národní
úložiště
šedé
literatury

Numerické optimalizační metody. Nepodmíněná minimalizace

Lukšan, Ladislav
2011

Dostupný z http://www.nusl.cz/ntk/nusl-81271

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 23.04.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .

Institute of Computer Science
Academy of Sciences of the Czech Republic

Numerické optimalizační metody
Nepodmíněná minimalizace

L.Lukšan

Technical report No. 1152

Prosinec 2015

Pod Vodárenskou věží 2, 182 07 Prague 8 phone: +420 2 688 42 44, fax: +420 2 858 57 89,
e-mail:e-mail:ics@cs.cas.cz

Institute of Computer Science
Academy of Sciences of the Czech Republic

Numerické optimalizační metody
Nepodmíněná minimalizace

L.Lukšan [1]

Technical report No. 1152

Prosinec 2015

Abstract:

Tato zpráva popisuje teoretické i praktické vlastnosti numerických metod pro nepodmíněnou optimalizaci. Studují se metody pro obecné i speciální optimalizační úlohy mezi které patří minimalizace součtu čtverců, součtu absolutních hodnot, maximní hodnoty a dalších nehladkých funkcí. Kromě metod pro standardní úlohy středních rozměrů jsou studovány i metody pro rozsáhlé řídké a strukturované úlohy. Velká pozornost je věnována soustavám nelineárních rovnic.

Keywords:
Numerická optimalizace, nelineární aproximace, systémy nelineárních rovnic, algoritmy.

[1]This work was supported by the Grant Agency of the Czech Republic, project No. 201/09/1957, and the institutional research plan No. AV0Z10300504

Obsah

1 Úvod do problematiky numerické optimalizace **4**
1.1 Základní pojmy . . . 4
1.2 Podmínky optimality . . . 9
1.3 Základní pojmy z teorie konvergence . . . 11
1.4 Základní optimalizační metody . . . 16
1.5 Testování optimalizačních metod . . . 18

2 Metody spádových směrů **20**
2.1 Základní vlastnosti metod spádových směrů . . . 20
2.2 Globální konvergence . . . 24
2.3 Asymptotická rychlost konvergence . . . 34
2.4 Výběr délky kroku . . . 44
2.5 Nemonotonní metody spádových směrů . . . 47
2.6 Využití směrů se zápornou křivostí . . . 50
2.7 Maticové rozklady pro symetrické indefinitní matice . . . 54
2.8 Metody sdružených směrů . . . 61

3 Metody sdružených gradientů **70**
3.1 Základní metody sdružených gradientů . . . 70
3.2 Globální konvergence základních metod sdružených gradientů . . . 73
3.3 Asymptotická rychlost konvergence . . . 79
3.4 Spádové metody sdružených gradientů . . . 86
3.5 Globální konvergence spádových metod sdružených gradientů . . . 96
3.6 Implementace metod sdružených gradientů . . . 107
3.7 Numerické porovnání metod sdružených gradientů . . . 109
3.8 Předpodmíněná metoda sdružených gradientů pro řešení soustav lineárních rovnic . . . 111

4 Metody s proměnnou metrikou **121**
4.1 Základní vlastnosti metod s proměnnou metrikou . . . 121
4.2 Součinový tvar metod s proměnnou metrikou . . . 135
4.3 Variační odvození metod s proměnnou metrikou . . . 148
4.4 Výběr parametrů (škálování a korekce) . . . 157
4.5 Globální konvergence . . . 169
4.6 Superlineární konvergence . . . 176
4.7 Aktualizace trojúhelníkového rozkladu . . . 183
4.8 Modifikace a implementace metod s proměnnou metrikou . . . 185
4.9 Davidonova třída metod s proměnnou metrikou . . . 192
4.10 Numerické porovnání metod s proměnnou metrikou . . . 203

5 Metody s lokálně omezeným krokem **206**
5.1 Základní vlastnosti metod s lokálně omezeným krokem . . . 206
5.2 Metody s optimálním lokálně omezeným krokem . . . 217
5.3 Newtonova metoda s lokálně omezeným krokem . . . 219
5.4 Nemonotonní metody s lokálně omezeným krokem . . . 223
5.5 Kombinované metody s lokálně omezeným krokem . . . 225
5.6 Metody kvadratické regularizace . . . 228

6 Výpočet lokálně omezeného kroku **232**
6.1 Výpočet optimálního lokálně omezeného kroku . . . 232
6.2 Využití směru největšího spádu (metody psí nohy) . . . 236
6.3 Nepřesné metody s lokálně omezeným krokem . . . 240
6.4 Použití symetrické Lanczosovy metody . . . 242
6.5 Posunuté nepřesné metody s lokálně omezeným krokem . . . 247
6.6 Numerické porovnání jednotlivých algoritmů . . . 249
6.7 Iterační metody pro řešení lineárních soustav se symetrickou indefinitní maticí . . . 251

7 Metody kubické regularizace **264**
7.1 Základní vlastnosti metod kubické regularizace . . . 264
7.2 Optimální metody kubické regularizace . . . 269
7.3 Výpočet optimálního směrového vektoru . . . 272

8 Metody pro minimalizaci součtu čtverců **277**
8.1 Gaussova–Newtonova metoda . . . 281
8.2 Použití kvazinewtonovských aktualizací . . . 283
8.3 Numerické porovnání metod pro minimalizaci součtu čtverců . . . 288
8.4 Řešení lineární úlohy nejmenších čtverců . . . 289

9 Metody pro rozsáhlé husté úlohy **294**
9.1 Vektorové metody s proměnnou metrikou s omezenou pamětí . . . 294
9.2 Modifikované vektorové metody s proměnnou metrikou s omezenou pamětí . . . 304
9.3 Maticové metody s proměnnou metrikou s omezenou pamětí . . . 311
9.4 Modifikované maticové metody s proměnnou metrikou s omezenou pamětí . . . 320
9.5 Metody redukovaných Hessiánů s omezenou pamětí . . . 329
9.6 Posunuté metody s proměnnou metrikou s omezenou pamětí . . . 334
9.7 Vektorové diferenční verze Newtonovy metody . . . 343
9.8 Numerické porovnání . . . 362

10 Metody pro rozsáhlé řídké a separovatelné úlohy **365**
10.1 Řídké matice a grafy . . . 365
10.2 Diferenční verze Newtonovy metody pro řídké úlohy . . . 367
10.3 Metody s proměnnou metrikou pro řídké úlohy . . . 372
10.4 Diferenční verze Newtonovy metody pro separovatelné úlohy . . . 382
10.5 Metody s proměnnou metrikou pro separovatelné úlohy . . . 386
10.6 Modifikace Gaussovy–Newtonovy metody pro řídké a separovatelné úlohy . . . 392
10.7 Iterační řešení rozsáhlých lineárních úloh nejmenších čtverců . . . 395
10.8 Numerické porovnání . . . 400

11 Metody pro řešení soustav nelineárních rovnic **404**
11.1 Základní vlastnosti metod pro řešení soustav nelineárních rovnic . . . 404
11.2 Metody spádových směrů . . . 407
11.3 Metody s lokálně omezeným krokem . . . 413
11.4 Newtonova metoda . . . 418
11.5 Kvazinewtonovské metody . . . 419
11.6 Nemonotonní kvazinewtonovské metody . . . 428
11.7 Sdružené kvazinewtonovské metody . . . 431
11.8 Tenzorové metody . . . 435
11.9 Aktualizace ortogonálního rozkladu . . . 439
11.10 Numerické porovnání . . . 440

12 Metody pro rozsáhlé soustavy nelineárních rovnic **442**
12.1 Kvazinewtonovské metody s omezenou pamětí . . . 442
12.2 Diferenční verze Newtonovy metody pro husté úlohy . . . 445
12.3 Diferenční verze Newtonovy metody pro řídké úlohy . . . 446
12.4 Kvazinewtonovské metody pro řídké úlohy . . . 447
12.5 Sdružené kvazinewtonovské metody pro řídké úlohy . . . 451
12.6 Metody založené na aktualizaci nesymetrického trojúhelníkového rozkladu . . . 454
12.7 Nedokonalé diferenční verze Newtonovy metody . . . 455
12.8 Iterační řešení systémů lineárních rovnic s nesymetrickou maticí . . . 456
12.9 Metody s lokálně omezeným krokem . . . 466
12.10 Numerické porovnání . . . 468

13 Optimalizace dynamických systémů **471**
13.1 Výpočet gradientu . . . 472
13.2 Výpočet Hessovy matice . . . 473
13.3 Aproximace Hessovy matice pro kritérium nejmenších čtverců . . . 476
13.4 Metody pro optimalizaci dynamických systémů . . . 477

14 Automatické a numerické derivování **480**
14.1 Automatický výpočet prvních derivací . . . 481
14.2 Automatický výpočet druhých derivací . . . 486
14.3 Numerický výpočet gradientu . . . 488

15 Základy nehladké analýzy **492**
15.1 Konvexní množiny . . . 493
15.2 Konvexní kužely . . . 502
15.3 Konvexní funkce . . . 509
15.4 Lipschitzovské funkce . . . 520
15.5 Lipschitzovská zobrazení . . . 530
15.6 Polohladká zobrazení . . . 537

16 Metody pro řešení soustav nehladkých rovnic **543**
16.1 Newtonova metoda . . . 543
16.2 Aplikace nehladkých rovnic . . . 548

17 Metody pro nehladkou optimalizaci **553**
17.1 Svazkové metody . . . 553

18 Úvod do problematiky nelineárního programování **564**
18.1 Základní pojmy . . . 564
18.2 Podmínky optimality pro úlohy s konvexními omezeními . . . 565

19 Minimalizace s lineárními omezeními **573**
19.1 Minimalizace na lineární varietě . . . 573
19.2 Změna lineární variety při přidání nebo ubrání aktivního omezení . . . 576
19.3 Metody aktivních omezení . . . 581

Učební texty **583**

Literatura **583**

1 Úvod do problematiky numerické optimalizace

V tomto textu jsou studovány základní metody pro nepodmíněnou minimalizaci včetně jejich konvergenčních vlastností. Po stručném úvodu do problematiky jsou v kapitole 2 uvedeny metody spádových směrů a jejich nejtypičtější realizace (metody sdružených gradientů a metody s proměnnou metrikou). Kapitola 3 je věnována metodám s lokálně omezeným krokem vhodným zejména ke globálně konvergentní realizaci Newtonovy metody a Gaussovy-Newtonovy metody pro minimalizaci součtu čtverců. V kapitole 4 jsou popsány speciální metody pro rozsáhlé a strukturované optimalizační úlohy. Kapitola 5 je věnována metodám pro řešení soustav nelineárních rovnic. V kapitole 6 jsou popsány speciální metody pro rozsáhlé a strukturované soustavy nelineárních rovnic. Věty a lemata jsou v této práci téměř vždy dokazovány. Tvrzení z příbuzných oborů, která lze nalézt v běžných učebních textech (například v [T1] – [T11], jsou uváděny bez důkazu.

1.1 Základní pojmy

Budeme používat označení $x \in R^n$ pro vektor dimenze n, $F(x)$ pro funkci $F : \mathcal{D}_F \to R$ a

$$g(x) = \begin{bmatrix} \dfrac{\partial F(x)}{\partial x_1} \\ \vdots \\ \dfrac{\partial F(x)}{\partial x_n} \end{bmatrix}, \qquad G(x) = \begin{bmatrix} \dfrac{\partial^2 F(x)}{\partial x_1^2}, & \ldots, & \dfrac{\partial^2 F(x)}{\partial x_1 \partial x_n} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial^2 F(x)}{\partial x_n \partial x_1}, & \ldots, & \dfrac{\partial^2 F(x)}{\partial x_n^2} \end{bmatrix}.$$

Zde $F(x)$ je účelová funkce definovaná na množině $\mathcal{D}_F \subset R^n$, $g(x)$ je její gradient a $G(x)$ je její Hessova matice (matice druhých parciálních derivací). Symboly $\underline{\lambda}(G(x))$ a $\overline{\lambda}(G(x))$ budeme označovat nejmenší a největší vlastní číslo matice $G(x)$. Většinou budeme předpokládat, že funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná na nějaké otevřené množině $\mathcal{D} \subset \mathcal{D}_F$. V tomto případě budeme psát $F \in \mathcal{C}^2$ nebo $F \in \mathcal{C}^2 : \mathcal{D} \to R$. Spojitost druhých parciálních derivací implikuje symetrii matice $G(x)$. Poznamenejme, že v mnoha případech stačí místo spojitosti druhých parciálních derivací předpokládat lipschitzovskost prvních parciálních derivací. Lipschitzovskost a konvexita jsou pojmy, které se probírají v základních kurzech matematické analýzy a podrobně se jimi budeme zabývat v kapitole 15. Protože tyto pojmy mají v teorii optimalizačních metod klíčový význam uvedeme zde základní definice, které budeme často používat.

Definice 1. *Řekněme, že množina $\mathcal{C} \in R^n$ je konvexní, jestliže*

$$\lambda_1 x_1 + \lambda_2 x_2 \in \mathcal{C},$$

pokud $x_1 \in \mathcal{C}$, $x_2 \in \mathcal{C}$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, a $\lambda_1 + \lambda_2 = 1$.

Definice 2. *Řekneme, že funkce $F : \mathcal{D}_F \to R$ je konvexní na konvexní množině $\mathcal{C} \subset \mathcal{D}_F \subset R^n$, jestliže*

$$F(\lambda_1 x_1 + \lambda_2 x_2) \leq \lambda_1 F(x_1) + \lambda_2 F(x_2),$$

pokud $x_1 \in \mathcal{C}$, $x_2 \in \mathcal{C}$, $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, a $\lambda_1 + \lambda_2 = 1$. Řekneme, že funkce $F : \mathcal{D}_F \to R$ je ryze konvexní na konvexní množině $\mathcal{C} \subset \mathcal{D}_F \subset R^n$, jestliže

$$F(\lambda_1 x_1 + \lambda_2 x_2) < \lambda_1 F(x_1) + \lambda_2 F(x_2),$$

pokud $x_1 \neq x_2$, $\lambda_1 > 0$, $\lambda_2 > 0$, a $\lambda_1 + \lambda_2 = 1$. Řekneme, že funkce $F : \mathcal{D}_F \to R$ je konkávní (ryze konkávní) na konvexní množině $\mathcal{C} \subset \mathcal{D}_F \subset R^n$, je-li funkce $-F$ konvexní (ryze konvexní) na $\mathcal{C}$.

Definice 3. *Řekneme, že funkce $F : \mathcal{D}_F \to R$ je lipschitzovská na konvexní množině $\mathcal{C} \subset \mathcal{D}_F \subset R^n$, jestliže existuje konstanta $L > 0$ taková, že*

$$|F(x_2) - F(x_1)| \leq L\|x_2 - x_1\|,$$

pokud $x_1 \in \mathcal{C}$ a $x_2 \in \mathcal{C}$ (v případě zobrazení používáme místo absolutní hodnoty normu).

Vlastnosti konvexních a lipschitzovských funkcí jsou studovány v kapitole 15. Zde pouze připomeneme, že dvakrát spojitě diferencovatelná funkce $F \in \mathcal{C}^2 : \mathcal{D}_F \to R$ je konvexní na konvexní množině $\mathcal{C} \subset \mathcal{D}_\mathcal{F}$, je-li její Hessova matice pozitivně semidefinitní na $\mathcal{C}$, a ryze konvexní na $\mathcal{C}$, je-li její Hessova matice pozitivně definitní na $\mathcal{C}$.

Při vyšetřování konvergence optimalizačních metod budeme používat tyto předpoklady.

Předpoklad F1. *Funkce $F : \mathcal{D}_F \to R$ je zdola omezená, takže existuje konstanta $\underline{F}$ taková, že*

$$F(x) \geq \underline{F} \quad \forall x \in \mathcal{D}_F. \tag{1}$$

Poznámka 1. Předpoklad F1 je vcelku logický. Hledáme-li lokální minimum, je žádoucí, aby funkce F byla zdola omezená. Přesto je někdy tento předpoklad omezující. Uvažujme funkci $F : R \to R$ definovanou vztahem $F(x) = x^3 - 3x$. Tato funkce má lokální minimum v bodě $x = 1$, ale $F(x) \to -\infty$, pokud $x \to -\infty$. Nicméně optimalizační metoda může nálézt lokální minimum $x = 1$, odstartujeme-li ji z vhodného bodu $x_1 > -1$. Podobné vlastnosti má celá řada pokutových funkcí používaných k hledání vázaných extrémů.

Předpoklad F2. *Množina*

$$\mathcal{D}_F(\overline{F}) = \{x \in \mathcal{D}_F : F(x) \leq \overline{F}\} \tag{2}$$

je kompaktní pro vhodnou hodnotu $\overline{F} \in R$.

Poznámka 2. Předpoklad F2 je opět logický. Je-li funkce $F : \mathcal{D}_F \subset R \to R$ zdola omezená a klesající, jako například funkce $F(x) = \exp(-x)$, může optimalizační metoda generovat divergentní posloupnost $x_i \to \infty$ takovou, že $g(x_i) \to 0$, takže i přes neschopnost nalézt lokální minimum je tato metoda globálně konvergentní (podle definice 14). Abychom vysvětlili, co rozumíme vhodnou hodnotou $\overline{F} \in R$, uvažujme funkci $F : R \to R$ definovanou vztahem $F(x) = -\cos(x)/(1+x^2)$. Tato funkce nabývá minima $F(x^*) = -1$ v bodě $x^* = 0$ a $F(x) \to 0$, pokud $|x| \to \infty$. Zvolíme-li $\overline{F} = -1/2$, je množina $\mathcal{D}_F(\overline{F})$ kompaktní (je obsažená v množině $\{x \in R^n : |x| \leq \pi/2\}$). Zvolíme-li $\overline{F} = 1/2$, platí $\mathcal{D}_F = R^n$, takže $\mathcal{D}_F$ není kompaktní. Jak uvidíme později, pokládáme obvykle $\overline{F} = F(x_1)$, kde x_1 je počáteční bod posloupnosti x_i, $i \in N$, generované optimalizační metodou.

V dalším výkladu budeme většinou předpokládat, že funkce F je spojitě diferencovatelná na otevřené konvexní množině $\mathcal{D} \subset \mathcal{D}_F$, která obsahuje všechny body posloupnosti x_i, $i \in N$. Pokud nebude řečeno jinak, budeme předpokládat, že $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$. Někdy, zejména při vyšetřování lokální konvergence, budeme předpokládat, že $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$, kde $\varepsilon > 0$ a bod x^* je hromadným bodem posloupnosti x_i, $i \in N$. Konvexita množiny $\mathcal{D}$ je důležitá proto, že při vyšetřování optimalizační metody generující posloupnost x_i, $i \in N$, potřebujeme, aby zvolené předpoklady platily na úsečce spojující dva po sobě jdoucí body x_i a x_{i+1}. Uvažujme funkci $F : R \setminus \{0\} \to R$ definovanou vztahem $F(x) = x^2 + x^{-2}$. Tato funkce má lokální minima v bodech $x = \pm 1$ a není definovaná v bodě $x = 0$ (platí $F(x) \to \infty$ pro $\|x\| \to 0$). Odstartujeme-li optimalizační metodu v bodě $x_1 < -1$, může se stát že $x_2 > 0$ a funkce F není definovaná na úsečce spojující body x_1 a x_2.

Předpoklad F3. *Funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená konvexní množina, má lipschitzovské první derivace na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$\|g(x_2) - g(x_1)\| \leq \overline{G}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{3}$$

Předpoklad F4. *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená konvexní množina, má omezené druhé derivace na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$|d^T G(x) d| \leq \overline{G}\|d\|^2 \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{4}$$

Podmínka (4) je ekvivalentní podmínce $\|G(x)\| \leq \overline{G}$ $\forall x \in \mathcal{D}$.

Poznámka 3. Předpoklad F4 je silnější než F3 (z F4 plyne F3). Jelikož se s nerovností (4) pracuje pohodlněji než s nerovností (3), budeme často předpokládat F4 místo F3, zejména v případech kdy používáme předpoklad F5.

Předpoklad F5. *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená konvexní množina, je stejnoměrně silně konvexní na $\mathcal{D}$, takže existuje konstanta $\underline{G} > 0$ taková, že*

$$d^T G(x) d \geq \underline{G} \|d\|^2 \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{5}$$

Předpoklad F6. *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená konvexní množina, má lipschitzovské druhé derivace na $\mathcal{D}$, takže existuje konstanta $\overline{L} > 0$ taková, že*

$$\|G(x_2) - G(x_1)\| \leq \overline{L} \|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{6}$$

Poznámka 4. Jsou-li splněny předpoklady F5 a F6, platí

$$\|G^{-1}(x_2) - G^{-1}(x_1)\| \leq \frac{\overline{L}}{\underline{G}^2} \|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{7}$$

Označíme-li $G_1 = G(x_1)$ a $G_2 = G(x_2)$, můžeme podle (5) psát

$$\begin{aligned} \frac{1}{\underline{G}} \|G_2 - G_1\| &\geq \|G_2^{-1/2}(G_2 - G_1)G_2^{-1/2}\| = \|I - G_2^{-1/2} G_1 G_2^{-1/2}\|, \\ \|G_2^{-1} - G_1^{-1}\| &= \|G_1^{-1/2}(G_1^{1/2} G_2^{-1} G_1^{1/2} - I) G_1^{-1/2}\| \leq \frac{1}{\underline{G}} \|G_1^{1/2} G_2^{-1} G_1^{1/2} - I\|. \end{aligned}$$

Matice $G_2^{-1/2} G_1 G_2^{-1/2}$ a $G_1^{1/2} G_2^{-1} G_1^{1/2}$ mají stejná vlastní čísla, neboť z $G_2^{-1/2} G_1 G_2^{-1/2} x = \lambda x$, kde $x \neq 0$, plyne $G_1^{1/2} G_2^{-1} G_1^{1/2} y = \lambda y$, kde $y = G_1^{1/2} G_2^{-1/2} x \neq 0$. Odtud plyne, že normy na pravých stranách uvedených nerovností jsou stejné, takže

$$\|G_2^{-1} - G_1^{-1}\| \leq \frac{1}{\underline{G}^2} \|G_2 - G_1\|,$$

což spolu s (6) dává (7).

Poznámka 5. Předpoklady F3–F6 lze též aplikovat na vektorová zobrazení tvaru $f = [f_1, \ldots, f_m]^T$, kde $f_k : \mathcal{D}_{f_k} \subset R^n \to R$, $1 \leq k \leq m$, jsou spojitě diferencovatelné funkce. Pak řekneme, že zobrazení $f : \mathcal{D}_f \to R^m$, kde $\mathcal{D}_f = \mathcal{D}_{f_1} \cap \cdots \cap \mathcal{D}_{f_m}$, splňuje některý z předpokladů F3–F6, splňuje-li tento předpoklad každá z funkcí f_k, $1 \leq k \leq m$.

Studujeme-li chování iteračního procesu v okolí lokálního minima $x^* \in R^n$, používáme v (3)–(6) množinu $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$, kde $\varepsilon > 0$. V tomto případě můžeme předpoklady F3–F6 nahradit slabšími předpoklady F3*–F6* (hvězdička značí, že se omezujeme na okolí bodu x^*). V předpokladech F3* a F6* nevyžadujeme lipschitzovskost gradientu a Hessovy matice v $\mathcal{B}(x^*, \varepsilon)$. Stačí předpokládat klidnost gradientu a Hessovy matice v okolí bodu x^*.

Definice 4. *Řekneme, že funkce $F : \mathcal{D}_F \to R$ je klidná v okolí bodu x^*, jestliže existují čísla $L > 0$ a $\varepsilon > 0$ taková, že*

$$|F(x) - F(x^*)| \leq L \|x - x^*\|,$$

pokud $x \in \mathcal{B}(x^*, \varepsilon)$ (v případě zobrazení používáme místo absolutní hodnoty normu).

Předpoklad F3*. *Funkce $F : \mathcal{D}_F \to R$ je spojitě diferencovatelná v okolí bodu $x^* \in R^n$ a její gradient je klidný v okolí bodu x^*, takže existují čísla $\overline{G} > 0$ a $\varepsilon > 0$ taková, že*

$$\|g(x) - g(x^*)\| \leq \overline{G}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{8}$$

Předpoklad F4*. *Funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná v okolí bodu $x^* \in R^n$. Pak pro libovolnou konstantu $\overline{G} > \|G(x^*)\|$ existuje číslo $\varepsilon > 0$ takové, že*

$$|d^T G(x) d| \leq \overline{G}\|d\|^2 \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{9}$$

Předpoklad F5*. *Funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná v okolí bodu $x^* \in R^n$ a matice $G(x^*)$ je pozitivně definitní. Pak pro libovolnou konstantu $0 < \underline{G} < \underline{\lambda}(G(x^*))$ existuje číslo $\varepsilon > 0$ takové, že*

$$d^T G(x) d \geq \underline{G}\|d\|^2 \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{10}$$

Předpoklad F6*. *Funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná v okolí bodu $x^* \in R^n$ a její Hessova marice je klidná v okolí bodu x^*, takže existují čísla $\overline{L}$ a $\varepsilon > 0$ taková, že*

$$\|G(x) - G(x^*)\| \leq \overline{L}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{11}$$

V konvergenčních důkazech budeme často používat věty o střední hodnotě známé z úvodních kurzů matematické analýzy. Symbolem $[x, x+d]$ označíme úsečku spojující body $x \in R^n$ a $x+d \in R^n$ a symbolem $(x, x+d)$ její vnitřek.

Tvrzení 1. *Nechť $F \in \mathcal{C}^1 : \mathcal{D} \to R$ a $[x, x+d] \subset \mathcal{D}$. Pak platí*

$$F(x+d) = F(x) + d^T g(\tilde{x}),$$

kde $\tilde{x} \in (x, x+d)$ (takže $\tilde{x} = x + \tilde{\lambda} d$, kde $0 < \tilde{\lambda} < 1$).

Použijeme-li tvrzení 1 a předpoklad F3, dostaneme

$$F(x+d) - F(x) \leq d^T g(x) + \overline{G}\|d\|^2. \tag{12}$$

Tvrzení 2. *Nechť $F \in \mathcal{C}^2 : \mathcal{D} \to R$, $x \in \mathcal{D}$ a $[x, x+d] \subset \mathcal{D}$. Pak platí*

$$F(x+d) = F(x) + d^T g(x) + \frac{1}{2} d^T G(\tilde{x}) d,$$

kde $\tilde{x} \in (x, x+d)$ (takže $\tilde{x} = x + \tilde{\lambda} d$, kde $0 < \tilde{\lambda} < 1$).

Použijeme-li tvrzení 2 a předpoklad F4, dostaneme

$$F(x+d) - F(x) \leq d^T g(x) + \frac{1}{2}\overline{G}\|d\|^2. \tag{13}$$

Použijeme-li tvrzení 2 a předpoklad F5, dostaneme

$$F(x+d) - F(x) \geq d^T g(x) + \frac{1}{2}\underline{G}\|d\|^2. \tag{14}$$

Tvrzení 3. *Nechť $F \in \mathcal{C}^2 : \mathcal{D} \to R$, $x \in \mathcal{D}$ a $[x, x+d] \subset \mathcal{D}$. Pak platí*

$$g(x+d) = g(x) + \int_0^1 G(x+\lambda d)d\mathrm{d}\lambda.$$

Použijeme-li předpoklad F3 nebo tvrzení 3 a předpoklad F4, dostaneme

$$\|g(x+d) - g(x)\| \leq \overline{G}\|d\|, \tag{15}$$

$$d^T(g(x+d) - g(x)) \leq \overline{G}\|d\|^2. \tag{16}$$

Použijeme-li tvrzení 3 a předpoklad F5, dostaneme

$$\|g(x+d) - g(x)\| \geq \underline{G}\|d\|, \tag{17}$$

$$d^T(g(x+d) - g(x)) \geq \underline{G}\|d\|^2. \tag{18}$$

Důkaz posledních dvou nerovností:

$$d^T(g(x+d) - g(x)) = \int_0^1 d^T G(x+\lambda d)d\mathrm{d}\lambda \geq \int_0^1 \underline{G}\|d\|^2\mathrm{d}\lambda = \underline{G}\|d\|^2,$$

$$\underline{G}\|d\|^2 \leq d^T(g(x+d) - g(x)) \leq \|d\|\|g(x+d) - g(x)\|.$$

Poznámka 6. Tvrzení 3, které je větou o střední hodnotě integrálního počtu, vychází z toho, že Hessova matice je primitivním zobrazením ke gradientu. Analogie tvrzení 1 pro obecná zobrazení neplatí. Označíme-li však $G_i(x)$, $1 \leq i \leq n$, řádky Hessovy matice $G(x)$, můžeme psát

$$g(x+d) = g(x) + \begin{bmatrix} G_1(x+\tilde{\lambda}_1 d) \\ \vdots \\ G_n(x+\tilde{\lambda}_n d) \end{bmatrix} d,$$

kde $0 \leq \tilde{\lambda}_i \leq 1$, $1 \leq i \leq n$ (tato čísla jsou obecně různá). Uvedené vyjádření však není tak praktické (pro stanovení odhadů (15)–(18) jako věta o střední hodnotě integrálního počtu. Poznamenejme, že existuje ještě jiné vyjádření pro přírůstek gradientu, uvedené v oddílu 15.5 (věta 344), které má tvar

$$g(x+d) - g(x) \in \left(\operatorname{conv} \bigcup_{0 \leq \lambda \leq 1} G(x+\lambda d)\right) d.$$

Uvedeme nyní dva důležité důsledky vět o střední hodnotě, které budeme často používat při odvozování podmínek optimality a při vyšetřování vlastností metod spádových směrů.

Věta 1. *Nechť $F \in \mathcal{C}^1 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená množina. Nechť $x \in \mathcal{D}$ a $s \in R^n$ je vektor takový, že $s^T g(x) < 0$. Pak existuje číslo $\overline{\alpha} > 0$ takové, že $x + \alpha s \in \mathcal{D}$ a $F(x+\alpha s) < F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$.*

Důkaz Jelikož $F \in \mathcal{C}^1$ na otevřené množině $\mathcal{D}$, existuje číslo $\overline{\alpha} > 0$, takové, že $x + \alpha s \in \mathcal{D}$ a $s^T g(x+\alpha s) \leq s^T g(x)/2$, pokud $0 \leq \alpha \leq \overline{\alpha}$ (plyne to ze spojitosti gradientu $g(x+\alpha s)$ a ze spojitosti skalárního součinu). Nechť $0 < \alpha \leq \overline{\alpha}$. Podle věty o střední hodnotě (tvrzení 1) existuje číslo $0 \leq \tilde{\alpha} \leq \alpha \leq \overline{\alpha}$ takové, že

$$F(x+\alpha s) = F(x) + \alpha s^T g(x+\tilde{\alpha} s) \leq F(x) + \frac{1}{2}\alpha s^T g(x) < F(x).$$

□

Věta 2. *Nechť $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D} \subset \mathcal{D}_F$ je otevřená množina. Nechť $x \in \mathcal{D}$ a $s \in R^n$ je vektor takový, že $s^T g(x) = 0$. Jestliže $s^T G(x) s > 0$, existuje číslo $\overline{\alpha} > 0$ takové, že $x + \alpha s \in \mathcal{D}$ a $F(x + \alpha s) > F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$. Jestliže $s^T G(x) s < 0$, existuje číslo $\overline{\alpha} > 0$ takové, že $x + \alpha s \in \mathcal{D}$ a $F(x + \alpha s) < F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$.*

Důkaz Nechť $s^T g(x) = 0$ a $s^T G(x) s > 0$. Jelikož $F \in \mathcal{C}^2$ na otevřené množině $\mathcal{D}$, existuje číslo $\overline{\alpha} > 0$, takové, že $x + \alpha s \in \mathcal{D}$ a $s^T G(x + \alpha s) s \geq s^T G(x) s / 2$, pokud $0 \leq \alpha \leq \overline{\alpha}$ (plyne to ze spojitosti Hessovy matice $G(x + \alpha s)$ a ze spojitosti skalárního součinu). Nechť $0 < \alpha \leq \overline{\alpha}$. Podle věty o střední hodnotě (tvrzení 2) existuje číslo $0 \leq \tilde{\alpha} \leq \alpha \leq \overline{\alpha}$ takové, že

$$F(x + \alpha s) = F(x) + \alpha s^T g(x) + \frac{1}{2} \alpha^2 s^T G(x + \tilde{\alpha} s) s \geq F(x) + \frac{1}{4} \alpha^2 s^T G(x) s > F(x)$$

(neboť $s^T g(x) = 0$). Druhá část tvrzení (pro $s^T G(x) s < 0$) se dokazuje analogicky (nebo se použije první část tvrzení na funkci $-F$). □

1.2 Podmínky optimality

Definice 5. *Řekneme, že bod $x^* \in R^n$ je lokálním minimem funkce $F : \mathcal{D}_F \to R$, existuje-li číslo $\varepsilon > 0$ takové, že*

$$F(x^*) \leq F(x) \quad \forall x \in \mathcal{B}(x^*, \varepsilon).$$

Jestliže navíc $F(x^*) < F(x)$ pokud $x^* \neq x$, řekneme, že bod $x^* \in R^n$ je ostrým lokálním minimem funkce F. Jestliže lze $\varepsilon > 0$ zvolit tak, že $\mathcal{B}(x^*, \varepsilon)$ již neobsahuje žádné jiné lokalní minimum funkce F, řekneme, že bod $x^* \in R^n$ je izolovaným lokálním minimem funkce F.

Poznámka 7. Pojem ostrého lokálního minima není totožný s pojmem izolovaného lokálního minima. Uvažujme funkci $F : R \to R$ zadanou předpisem

$$\begin{aligned} F(x) &= 0, & x = 0, \\ F(x) &= x^4(2 + \cos(1/x)), & x \neq 0. \end{aligned}$$

Tato funkce je spojitě diferencovatelná v R, má ostré lokální minimum v bodě $x = 0$ a platí

$$\begin{aligned} F'(x) &= 0, & x = 0, \\ F'(x) &= 4x^3(2 + \cos(1/x)) + x^2 \sin(1/x), & x \neq 0. \end{aligned}$$

Ostatní extrémy tedy vyhovují rovnici $4x(2 + \cos(1/x)) + \sin(1/x) = 0$ (věta 3). Zvolme libovolně číslo $0 < \varepsilon < 1/(4\pi)$. Pak funkce $\sin(1/x)$ nabývá v intervalu $[\varepsilon/2, \varepsilon]$ alespoň dvakrát všech hodnot z intervalu $(-1, 1)$ a jelikož pro $0 < x < 1/(4\pi)$ platí $0 < 4x \leq 4x(2 + \cos(1/x)) \leq 12x \leq 3/\pi < 1$, má funkce $F'(x)$ ma intervalu $[\varepsilon/2, \varepsilon]$ alespoň dva kořeny (odpovídající minimu a maximu funkce $F(x)$). Jelikož číslo $0 < \varepsilon < 1/(4\pi)$ můžeme volit libovolě malé, nemá funkce F v bodě $x = 0$ izolované lokální minimum. Je zřejmé, že každé izolované lokální minimum je také ostrým lokálním minimem.

Věta 3. *(Nutné podmínky) Nechť bod $x^* \in R^n$ je lokálním minimem funkce $F : \mathcal{D}_F \to R$ a nechť $F \in \mathcal{C}^1$ (spojitě diferencovatelná) na $\mathcal{B}(x^*, \varepsilon)$. Pak platí*

$$g(x^*) = 0.$$

Jestliže navíc $F \in \mathcal{C}^2$ (dvakrát spojitě diferencovatelná) na $\mathcal{B}(x^*, \varepsilon)$, pak platí

$$G(x^*) \succeq 0$$

(matice $G(x^*)$ je pozitivně semidefinitní).

Důkaz (a) Nechť $F \in \mathcal{C}^1$ na $\mathcal{B}(x^*, \varepsilon)$. Předpokládejme, že $g^* = g(x^*) \neq 0$ a položme $s = -g^*$ Jelikož $F \in \mathcal{C}^1$ na $\mathcal{B}(x^*, \varepsilon)$ a $s^T g^* < 0$, existuje podle věty 1 číslo $\overline{\alpha} > 0$ takové, že $x^* + \alpha s \in \mathcal{B}(x^*, \varepsilon)$ a $F(x + \alpha s) < F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$, což je ve sporu s definicí 5.

(b) Nechť navíc $F \in \mathcal{C}^2$ na $\mathcal{B}(x^*, \varepsilon)$. Předpokládejme, že $g(x^*) = 0$, ale matice $G(x^*)$ není pozitivně semidefinitní. Pak existuje vektor $s \in R^n$ takový, že $s^T g^* = 0$ a $s^T G(x^*) s < 0$ a podle věty 2 existuje číslo $\overline{\alpha} > 0$ takové, že $x^* + \alpha s \in \mathcal{B}(x^*, \varepsilon)$ a $F(x + \alpha s) < F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$, což je opět ve sporu s definicí 5. □

Poznámka 8. Nutná podmínka prvního řádu $g(x^*) = 0$ udává, že lokální minimum diferencovatelné funkce je jejím stacionárním bodem. Tuto podmínku splňují i lokální maxima a řada dalších bodů. Patří mezi ně sedlové body, které se vyznačují tím, že funkce $\varphi(\alpha) = F(x^* + \alpha d)$ má v bodě $\alpha = 0$ pro některé směrové vektory lokální minimum a pro jiné lokální maximum. Jako příklad lze uvést funkci $F(x_1, x_2) = x_1^2 - x_2^2$, pro kterou je bod $x^* = 0$ stacionárním bodem a pro kterou platí $\varphi(\alpha) = \alpha^2(d_1^2 - d_2^2)$. Funkce $\varphi(\alpha)$ má v bodě $\alpha = 0$ lokální minimum, pokud $d_1^2 - d_2^2 > 0$, a lokální maximum, pokud $d_1^2 - d_2^2 < 0$. Hessova matice funkce $F(x)$ má v bodě $x^* = 0$ jedno kladné a jedno záporné vlastní číslo. Dalším případem jsou inflexní body. Uvažujme funkci $F(x_1, x_2) = x_1^3 + x_2^3$, pro kterou je bod $x^* = 0$ stacionárním bodem a pro kterou platí $\varphi(\alpha) = \alpha^3(d_1^3 + d_2^3)$. Funkce $\varphi(\alpha)$ je pro $d_1^3 + d_2^3 > 0$ rostoucí a pro $d_1^3 + d_2^3 < 0$ klesající a bod $\alpha = 0$ je jejím inflexním bodem. Hessova matice funkce $F(x)$ má v bodě $x^* = 0$ dvě nulová vlastní čísla.

Věta 4. *(Postačující podmínky) Nechť $F \in \mathcal{C}^2$ na $\mathcal{B}(x^*, \varepsilon)$ a nechť platí*

$$g(x^*) = 0$$

a

$$G(x^*) \succ 0$$

(matice $G(x^*)$ je pozitivně definitní). Pak bod $x^* \in R^n$ je izolovaným lokálním minimem funkce $F : \mathcal{D}_F \to R$.

Důkaz Jelikož matice $G(x^*)$ je pozitivně definitní, platí $\lambda^* > 0$, kde λ^* je nejmenší vlastní číslo matice $G(x^*)$. Nechť $s \in R^n$ a $s \neq 0$. Z extremálních vlastností vlastních čísel plyne, že $s^T G(x^*) s \geq \lambda^* s^T s > 0$. Jelikož $F \in \mathcal{C}^2$ na $\mathcal{B}(x^*, \varepsilon)$, existuje podle věty 2 číslo $\overline{\alpha} > 0$ takové, že $x^* + \alpha s \in \mathcal{B}(x^*, \varepsilon)$ a $F(x + \alpha s) > F(x)$, pokud $0 < \alpha \leq \overline{\alpha}$, takže bod x^* je ostrým lokálním minimem funkce F. Ze spojitosti Hessovy matice plyne existence čísla $\overline{\alpha} > 0$ takového, že $s^T G(x^* + \alpha s) s \geq \lambda^* s^T s / 2$ pro $0 < \alpha \leq \overline{\alpha}$. Pak pro $0 < \alpha \leq \overline{\alpha}$ platí

$$s^T g(x^* + \alpha s) = s^T \int_0^1 G(x^* + \lambda \alpha s) \alpha s \mathrm{d}\lambda \geq \frac{\alpha}{2} \lambda^* s^T s > 0,$$

takže $g(x^* + \alpha s) \neq 0$ a bod x^* je izolovaným lokálním minimem funkce F. □

Poznámka 9. Jsou-li splněny předpoklady věty 4 (postačující podmínky druhého řádu) je funkce $F \in \mathcal{C}^2 : \mathcal{D}_F \to R$ ryze konvexní v okolí bodu x^* (platí předpoklad F4*).

Příklad 1. Uvažujme funkci $F : R^2 \to R$ tvaru

$$F(x) = \frac{1}{4}(x_1^2 + x_2^2 - 1)^2.$$

Derivováním této funkce dostaneme

$$g(x) = \begin{bmatrix} x_1(x_1^2 + x_2^2 - 1) \\ x_2(x_1^2 + x_2^2 - 1) \end{bmatrix}, \qquad G(x) = \begin{bmatrix} x_1^2 + x_2^2 - 1 + 2x_1^2, & 2x_1 x_2 \\ 2x_1 x_2, & x_1^2 + x_2^2 - 1 + 2x_2^2 \end{bmatrix}.$$

Z rovnice $g(x) = 0$ zjistíme, že stacionárními body jsou počátek souřadnic $x = [0, 0]^T$ a body ležící na jednotkové kružnici $x_1^2 + x_2^2 = 1$. V počátku souřadnic platí $G(0) = \mathrm{diag}(-1, -1)$, takže Hessova matice je negativně definitní a daný bod je izolovaným lokálním maximem. V bodech na jednotkové kružnici platí

$$G(x) = 2 \begin{bmatrix} x_1^2, & x_1 x_2 \\ x_1 x_2, & x_2^2 \end{bmatrix}.$$

Tato matice má nezáporné diagonální prvky a nulový determinant, takže je pozitivně semidefinitní, ale není pozitivně definitní (je singulární). Stacionární body na jednotkové kružnici tedy mohou být neostrá lokální minima. Abychom se o tom přesvědčili, všimneme si, že funkce F je rotačně symetrická a můžeme ji vyjádřit v polárních souřadnicích ve tvaru $F(x) = \varphi(r) = (1/4)(r^2-1)^2$, kde $r^2 = x_1^2 + x_2^2$. Pro funkci $\varphi(r)$ platí $\varphi'(r) = r(r^2-1)$ a $\varphi''(r) = 3r^2 - 1$, takže má ostré lokální maximum v bodě $r = 0$ a ostrá lokální minima v bodech $r = \pm 1$, což ukazuje, že body na jednotkové kružnici jsou neostrá lokální minima funkce F.

Při vyšetřování metod pro nepodmíněnou minimalizaci budeme někdy potřebovat nutné podmínky prvního řádu pro vázané extrémy. Uvedeme proto bez důkazu jednoduchou variantu těchto podmínek.

Tvrzení 4. *Nechť funkce $F : R^n \to R$ a $c_i : R^n \to R$, $1 \le i \le m$, jsou spojitě diferencovatelné. Nechť vektory $\{\nabla c_i(x) : c_i(x) = 0,\ 1 \le i \le m\}$ jsou lineárně nezávislé v bodě $x \in R^n$, který je lokálním minimem funkce F na množině zadané omezeními $c_i(x) \le 0$, $i \in I$, $c_i(x) = 0$, $i \in E$ (kde $I \cup E = \{1, \dots, m\}$ a $I \cap E = \emptyset$). Pak existuje vektor Lagrangeových multiplikátorů $\lambda \in R^m$ takový, že platí*

$$\nabla F(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0,$$

$$c_i(x) = 0, \quad i \in E,$$

$$c_i(x) \le 0, \quad \lambda_i \ge 0, \quad \lambda_i c_i(x) = 0, \quad i \in I.$$

Jsou-li funkce F, c_i, $i \in I$, konvexní, funkce c_i, $i \in E$, lineární a jsou-li splněny uvedené nutné podmínky prvního řádu, je bod $x \in R^n$ globálním minimem funkce F na množině zadané omezeními $c_i(x) \le 0$, $i \in I$, $c_i(x) = 0$, $i \in E$.

Podmínky pro vázané extrémy jsou podrobně studovány v druhé části práce.

1.3 Základní pojmy z teorie konvergence

Nyní se budeme zabývat asymptotickými vlastnostmi konvergentních posloupností, tedy jejich chováním v okolí limitního bodu a jejich asymptotickou rychlostí konvergence.

Definice 6. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů. Jestliže pro libovolné $\varepsilon > 0$ existuje index $k \in N$ takový, že $x_i \in \mathcal{B}(x^*, \varepsilon)$ $\forall i \ge k$, řekneme, že posloupnost $x_i \in R^n$, $i \in N$ konverguje k bodu $x^* \in R^n$ a píšeme $x_i \to x^*$. Používáme značení $F_i = F(x_i)$, $g_i = g(x_i)$, $G_i = G(x_i)$.*

Poznámka 10. Při studiu asymptotického chování konvergentních posloupností budeme často používat symboly $o(\xi_i)$ a $O(\xi_i)$, kde ξ_i, $i \in N$, je nějaká omezená posloupnost kladných čísel. Nechť u_i, v_i, $i \in N$, jsou dvě posloupnosti (čísel, vektorů nebo matic) a $k \ge 0$. Jestliže $\|u_i\|/\|v_i\|^k \to 0$, budeme psát $u_i = o(\|v_i\|^k)$. Jestliže existuje konstanta $C > 0$ taková, že $\|u_i\| \le C\|v_i\|^k$ $\forall i \in N$, budeme psát $u_i = O(\|v_i\|^k)$. Místo $o(\|v_i\|^0)$ a $O(\|v_i\|^0)$ budeme psát $o(1)$ a $O(1)$. Pokud současně platí $u_i = O(\|v_i\|)$ a $v_i = O(\|u_i\|)$, čili pokud existují konstanty $0 < \underline{c} \le \overline{c} < \infty$ takové, že

$$\underline{c}\|v_i\| \le \|u_i\| \le \overline{c}\|v_i\| \quad \forall i \in N,$$

budeme psát $u_i \sim v_i$ nebo $\|u_i\| \sim \|v_i\|$. Pro práci se symboly $o(\xi_i)$ a $O(\xi_i)$ platí jednoduchá pravidla. Nejčastěji použijeme toho, že pro libovolný exponent $r \in R$ platí $(1 + o(\xi_i))^r = 1 + o(\xi_i)$ a $(1 + O(\xi_i))^r = 1 + O(\xi_i)$, pokud $o(\xi_i) \to 0$ a $O(\xi_i) \to 0$ (k důkazu těchto vztahů lze použít binomickou větu nebo rozvoj v mocninnou řadu). Poznamenejme ještě, že jednotlivé veličiny $o(\xi_i)$ a $O(\xi_i)$ nemusíme rozlišovat, takže lze například psát $u_i v_i = o(\xi_i)o(\xi_i) = o(\xi_i)^2 = o(\xi_i^2)$, pokud $u_i = o(\xi_i)$ a $v_i = o(\xi_i)$, nebo $u_i v_i = (1 + O(\xi_i))(1 + O(\xi_i)) = (1 + O(\xi_i))^2 = (1 + O(\xi_i))$, pokud $u_i = (1 + O(\xi_i))$ a $v_i = (1 + O(\xi_i))$.

Věta 5. *Nechť $x_i \in R^n$, $d_i \in R^n$, $i \in N$, jsou dvě posloupnosti takové, že $x_i \to x^*$ a $d_i \to 0$, kde $x^* \in R^n$ je stacionární bod funkce $F \in \mathcal{C}^2 : R^n \to R$. Označme $e_i = x_i - x^*$, $i \in N$, a $G^* = G(x^*)$. Pak platí*

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + o(\|d_i\|^2),$$

$$g(x_i + d_i) - g(x_i) = G_i d_i + o(\|d_i\|)$$

a

$$F(x_i) - F(x^*) = \frac{1}{2} e_i^T G^* e_i + o(\|e_i\|^2),$$

$$g(x_i) = G^* e_i + o(\|e_i\|)$$

a

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G^* d_i + o(\|d_i\|^2),$$

$$g(x_i + d_i) - g(x_i) = G^* d_i + o(\|d_i\|).$$

Důkaz Použijeme-li tvrzení 2 o střední hodnotě, dostaneme

$$\begin{aligned} F(x_i + d_i) - F(x_i) &= d_i^T g_i + \frac{1}{2} d_i^T G(x_i + \tilde{\lambda} d_i) d_i \\ &= d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{1}{2} d_i^T (G(x_i + \tilde{\lambda} d_i) - G_i) d_i, \end{aligned}$$

kde $0 \le \tilde{\lambda} \le 1$ a

$$|d_i^T (G(x_i + \tilde{\lambda} d_i) - G_i) d_i| \le \|G(x_i + \tilde{\lambda} d_i) - G_i\| \|d_i\|^2.$$

Ze spojitosti druhých derivací plyne $\|G(x_i + \tilde{\lambda} d_i) - G(x_i)\| \le \|G(x_i + \tilde{\lambda} d_i) - G^*\| + \|G(x_i) - G^*\| \to 0$, neboť $x_i \to x^*$ a $d_i \to 0$ (takže $x_i + \tilde{\lambda} d_i \to x^*$). Použijeme-li tvrzení 3 o střední hodnotě, dostaneme

$$\begin{aligned} g(x_i + d_i) - g(x_i) &= \int_0^1 G(x_i + \lambda d_i) d_i \mathrm{d}\lambda \\ &= G_i d_i + \int_0^1 (G(x_i + \lambda d_i) - G_i) d_i \mathrm{d}\lambda, \end{aligned}$$

kde

$$\begin{aligned} \left\| \int_0^1 (G(x_i + \lambda d_i) - G_i) d_i \mathrm{d}\lambda \right\| &\le \int_0^1 \|G(x_i + \lambda d_i) - G_i\| \|d_i\| \mathrm{d}\lambda \\ &\le \max_{0 \le \lambda \le 1} \|G(x_i + \lambda d_i) - G_i\| \|d_i\|. \end{aligned}$$

Ze spojitosti druhých derivací plyne opět $\max_{0 \le \lambda \le 1} \|G(x_i + \lambda d_i) - G(x_i)\| \to 0$. Tím jsme dokázali první dva vztahy. Druhé dva vztahy se dokazují úplně stejně. Provede se záměna x_i místo $x_i + d_i$, x^* místo x_i, $e_i = x_i - x^*$ místo $d_i = x_i + d_i - x_i$ a přihlédne se k tomu, že $g(x^*) = 0$. Poslední dva vztahy plynou z toho, že $G_i d_i = G^* d_i + (G_i - G^*) d_i$, kde $\|(G_i - G^*)\| \to 0$ pokud $x_i \to x^*$. □

Je-li navíc splněn předpoklad F6, kde $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$ a $\varepsilon > 0$, dostaneme silnější odhady.

Věta 6. *Nechť $x_i \in R^n$, $d_i \in R^n$, $i \in N$, jsou dvě posloupnosti takové, že $x_i \to x^*$ a $d_i \to 0$, kde $x^* \in R^n$ je stacionární bod funkce $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje předpokladu F6. Označme $e_i = x_i - x^*$, $i \in N$, a $G^* = G(x^*)$. Pak platí*

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + O(\|d_i\|^3),$$

$$g(x_i + d_i) - g(x_i) = G_i d_i + O(\|d_i\|^2)$$

a

$$F(x_i) - F(x^*) = \frac{1}{2} e_i^T G^* e_i + O(\|e_i\|^3),$$

$$g(x_i) = G^* e_i + O(\|e_i\|^2)$$

a

$$F(x_i + d_i) - F(x_i) = d_i^T g_i + \frac{1}{2} d_i^T G^* d_i + \|d_i\|^2 O(\|e_i\|),$$

$$g(x_i + d_i) - g(x_i) = G^* d_i + \|d_i\| O(\|e_i\|).$$

Důkaz Důkaz této věty je prakticky stejný jako důkaz věty 5. Vztahy typu $\|G(x_i + \lambda d_i) - G(x_i)\| \to 0$ se nahradí odhady typu $\|G(x_i + \lambda d_i) - G(x_i)\| \le \overline{L}\|\lambda d_i\|$. □

Definice 7. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) R-lineárně, jestliže*

$$\limsup_{i \to \infty} \|x_i - x^*\|^{1/i} < 1.$$

Věta 7. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) R-lineárně právě tehdy, když existují index $k \in N$ a čísla $M_k > 0$ a $0 < q < 1$, tak že*

$$\|x_i - x^*\| \le M_k q^{i-k} \|x_k - x^*\|$$

$\forall i \ge k$.

Důkaz (a) Nechť $\|x_i - x^*\| \le M_k q^{i-k} \|x_k - x^*\|$ $\forall i \ge k$, kde $q < 1$. Pak platí

$$\limsup_{i \to \infty} \|x_i - x^*\|^{1/i} \le \lim_{i \to \infty} (M_k \|x_k - x^*\|)^{1/i} \lim_{i \to \infty} (q^{i-k})^{1/i} = \lim_{i \to \infty} q^{1-k/i} = q < 1.$$

(b) Označme $\tilde{q} = \limsup_{i \to \infty} \|x_i - x^*\|^{1/i} < 1$. Pak pro libovolné číslo q takové že $\tilde{q} < q < 1$ existuje index $k \in N$ takový, že platí

$$\|x_i - x^*\|^{1/i} \le q$$

$\forall i \ge k$, neboli

$$\|x_i - x^*\| \le q^i$$

$\forall i \ge k$. Zvolme

$$M_k = \frac{q^k}{\|x_k - x^*\|}.$$

Pak platí

$$\|x_i - x^*\| \le M_k q^{i-k} \|x_k - x^*\|.$$

□

Poznámka 11. Výraz použitý v definici 7 nezávisí na posunu indexů. Pro libovolné číslo $k \in N$ platí

$$\limsup_{i\to\infty} \|x_i - x^*\|^{1/i} = \limsup_{i\to\infty} \|x_{i+k} - x^*\|^{1/i}.$$

Definice 8. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* R-superlineárně, jestliže*

$$\lim_{i\to\infty} \|x_i - x^*\|^{1/i} = 0.$$

Definice 9. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) Q-lineárně, jestliže*

$$\limsup_{i\to\infty} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} < 1.$$

Poznámka 12. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) Q-lineárně právě tehdy, když existuje index $k \in N$ a konstanta $0 < q < 1$ tak, že*

$$\frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \leq q \qquad \forall i \geq k.$$

Definice 10. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* Q-superlineárně, jestliže*

$$\lim_{i\to\infty} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} = 0.$$

Věta 8. *Nechť $x_i \to x^*$ Q-lineárně (Q-superlineárně). Pak $x_i \to x^*$ R-lineárně (R-superlineárně).*

Důkaz R-lineární konvergence plyne z Q-lineární konvergence bezprostředně (stačí položit $M_k = 1$ ve větě 7). Nechť $0 < \varepsilon < 1$ je libovolné (malé) číslo. Z Q-superlineární konvergence plyne existence indexu $k \in N$ takového, že

$$\frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \leq \varepsilon \qquad \forall i \geq k,$$

takže

$$\|x_i - x^*\| \leq \varepsilon^{i-k} \|x_k - x^*\| \qquad \forall i \geq k,$$

neboli

$$\limsup_{i\to\infty} \|x_i - x^*\|^{1/i} \leq \lim_{i\to\infty} (\|x_k - x^*\|)^{1/i} \lim_{i\to\infty} (\varepsilon^{1-k/i}) = \varepsilon.$$

Protože číslo ε je libovolné, musí platit

$$\limsup_{i\to\infty} \|x_i - x^*\|^{1/i} = 0.$$

□

Poznámka 13. Q-lineární (Q-superlineární) konvergence implikuje monotonnost posloupnosti $\|x_i - x^*\|$, $i \in N$ (počínaje vhodným indexem $k \in N$).

Definice 11. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* m-krokově Q-superlineárně, jestliže existuje číslo $m \in N$ takové, že*

$$\lim_{i\to\infty} \frac{\|x_{i+m} - x^*\|}{\|x_i - x^*\|} = 0.$$

Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* cyklicky m-krokově Q-superlineárně, jestliže existují čísla $\underline{l} \in N$, $m \in N$ taková, že

$$\lim_{k\to\infty} \frac{\|x_{(k+1)m+\underline{l}} - x^*\|}{\|x_{km+\underline{l}} - x^*\|} = 0.$$

Poznámka 14. Je zřejmé, že m-kroková Q-superlineární konvergence implikuje cyklickou m-krokovou Q-superlineární konvergenci. V dalších úvahách budeme často předpokládat, že $\underline{l} = 1$ (jestliže $\underline{l} > 1$ můžeme členy posloupnosti přečíslovat tak, že i nahradíme $i - \underline{l} + 1$).

Ukážeme nyní, že cyklická m-kroková Q-superlineární konvergence implikuje R-superlineární konvergenci.

Lemma 1. *Nechť ξ_i, $i \in N$, je posloupnost nezáporných čísel. Pak jestliže $\xi_i \to 0$, platí*

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \xi_i = 0.$$

Důkaz Jelikož $\xi_i \to 0$, existuje pro libovolné číslo $\varepsilon > 0$ index $l(\varepsilon) \in N$ takový, že $\xi_i < \varepsilon \ \forall i \geq l(\varepsilon)$. Nechť $k > l(\varepsilon)$. Pak

$$\frac{1}{k} \sum_{i=1}^{k} \xi_i = \frac{1}{k} \sum_{i=1}^{l(\varepsilon)} \xi_i + \frac{1}{k} \sum_{i=l(\varepsilon)+1}^{k} \xi_i < \frac{1}{k} S(\varepsilon) + \frac{k - l(\varepsilon)}{k} \varepsilon,$$

kde $S(\varepsilon) = \sum_{i=1}^{l(\varepsilon)} \xi_i$ a $l(\varepsilon)$ jsou (konečná) čísla, která závisí pouze na zvoleném $\varepsilon > 0$. Platí tedy

$$\lim_{k\to\infty} \frac{1}{k} \sum_{i=1}^{k} \xi_i \leq \lim_{k\to\infty} \frac{1}{k} S(\varepsilon) + \lim_{k\to\infty} \frac{k - l(\varepsilon)}{k} \varepsilon = \varepsilon.$$

Jelikož číslo ε bylo zvoleno libovolně, je tím lemma dokázáno. □

Lemma 2. *Nechť ξ_i, $1 \leq i \leq m$, jsou nezáporná čísla. Pak platí*

$$\left(\prod_{i=1}^{m} \xi_i \right)^{\frac{1}{m}} \leq \frac{1}{m} \sum_{i=1}^{m} \xi_i. \tag{19}$$

Důkaz Jelikož logaritmická funkce je konkávní (definice 2), platí

$$\log \left(\prod_{i=1}^{m} \xi_i \right)^{\frac{1}{m}} = \frac{1}{m} \sum_{i=1}^{m} \log \xi_i = \sum_{i=1}^{m} \frac{1}{m} \log \xi_i \leq \log \sum_{i=1}^{m} \frac{1}{m} \xi_i = \log \left(\frac{1}{m} \sum_{i=1}^{m} \xi_i \right)$$

a jelikož logaritmická funkce je rostoucí, dostaneme tvrzení lemmatu. □

Věta 9. *Nechť $x_i \to x^*$ cyklicky m-krokově Q-superlineárně a nechť existuje konstanta $C > 0$ taková, že $\|e_{i+1}\| \leq C\|e_i\| \ \forall i \in N$. Pak $x_i \to x^*$ R-superlineárně.*

Důkaz Předpokládejme pro jednoduchost, že $\underline{l} = 1$ (poznámka 14) a označme $i = km + l$, kde $1 \leq l \leq m$. Abychom dokázali, že $\lim_{i\to\infty} \|e_i\|^{1/i} = 0$, stačí dokázat, že pro libovolné celé číslo $1 \leq l \leq m$ platí $\lim_{k\to\infty} \|e_{km+l}\|^{1/(km+l)} = 0$. Označme $\overline{C} = \|e_1\| \max_{1\leq l\leq m} C^{l-1}$. Pak

$$\|e_{km+l}\| = \|e_1\| \left(\prod_{j=1}^{k} \frac{\|e_{jm+1}\|}{\|e_{(j-1)m+1}\|} \right) \frac{\|e_{km+l}\|}{\|e_{km+1}\|} \leq \overline{C} \left(\frac{1}{k} \sum_{j=1}^{k} \frac{\|e_{jm+1}\|}{\|e_{(j-1)m+1}\|} \right)^k = \overline{C}(o(1))^k$$

(používáme nerovnost (19) a tvrzení lemmatu 1, podle kterého platí $\frac{1}{k}\sum_{j=1}^{k} o(1) = o(1)$). Můžeme tedy psát

$$\lim_{k\to\infty} \|e_{km+l}\|^{\frac{1}{km+l}} \leq \lim_{k\to\infty} \overline{C}^{\frac{1}{km+l}} (o(1))^{\frac{k}{km+l}} = \lim_{k\to\infty} (o(1))^{\frac{1}{m+l/k}} = 0.$$

□

Poznámka 15. Předpoklady věty 9 jsou splněny pro cyklicky přerušovanou metodu sdružených gradientů s asymptoticky přesným výběrem délky kroku (věta 32).

Definice 12. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) kvadraticky, jestliže existuje index $k \in N$ a konstanta $0 < M_k < \infty$ tak, že*

$$\|x_{i+1} - x^*\| \leq M_k \|x_i - x^*\|^2 \quad \forall i \geq k.$$

Definice 13. *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) m-krokově kvadraticky, jestliže existuje index $k \in N$, číslo $m \in N$ a konstanta $0 < M_k < \infty$ tak, že*

$$\|x_{i+m} - x^*\| \leq M_k \|x_i - x^*\|^2 \quad \forall i \geq k.$$

V důkazech globální konvergence budeme často používat následující nerovnost.

Lemma 3. *Nechť $u_i \in R^n$, $1 \leq i \leq m$. Pak platí*

$$\left\| \sum_{i=1}^{m} u_i \right\|^2 \leq m \sum_{i=1}^{m} \|u_i\|^2. \tag{20}$$

Důkaz Použijeme-li Schwarzovu nerovnost a nerovnost (19), můžeme psát

$$\begin{aligned} \left\| \sum_{i=1}^{m} u_i \right\|^2 &= \sum_{i=1}^{m}\sum_{j=1}^{m} u_i^T u_j \leq \sum_{i=1}^{m}\sum_{j=1}^{m} \|u_i\|\|u_j\| = \sum_{i=1}^{m}\sum_{j=1}^{m} \left(\|u_i\|^2\|u_j\|^2\right)^{\frac{1}{2}} \\ &\leq \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}(\|u_i\|^2 + \|u_j\|^2) = \frac{1}{2}\left(m\sum_{i=1}^{m}\|u_i\|^2 + m\sum_{j=1}^{m}\|u_j\|^2 \right) = m\sum_{i=1}^{m}\|u_i\|^2. \end{aligned}$$

□

1.4 Základní optimalizační metody

Základní optimalizační metoda je iterační proces, jehož výsledkem je posloupnost $x_i \in R^n$, $i \in N$, taková, že

$$x_{i+1} = x_i + \alpha_i s_i, \tag{21}$$

kde směrový vektor $s_i \in R^n$ se určuje pomocí hodnot x_j, F_j, g_j, G_j, $1 \leq j \leq i$, a délka kroku $\alpha_i > 0$ se určuje na základě chování funkce $F : R^n \to R$ v okolí bodu $x_i \in R^n$.

Definice 14. *Řekneme, že základní optimalizační metoda je globálně konvergentní, jestliže pro libovolný počáteční vektor $x_1 \in R^n$ buď existuje index $i \in N$ takový, že $g(x_i) = 0$, nebo platí*

$$\liminf_{i\to\infty} \|g(x_i)\| = 0. \tag{22}$$

Poznámka 16. Není-li základní optimalizační metoda globálně konvergentní, existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g(x_i)\| \geq \underline{\varepsilon}\ \forall i \in N$ (stačí zvolit $\underline{\varepsilon} < \liminf_{i\to\infty} \|g(x_i)\|$). Proto budeme tvrzní (22) dokazovat tak, že přivedeme do sporu předpoklad, že $\|g(x_i)\| \geq \underline{\varepsilon}\ \forall i \in N$.

Mezi nejjednodušší a nejznámější optimalizační metody patří metoda největšího spádu a Newtonova metoda. Metoda největšího spádu je definována vztahy

$$s_i = -g(x_i),$$

$$\alpha_i = \arg\min_{\alpha\geq 0} F(x_i + \alpha s_i).$$

Výhody:

(1) Metoda největšího spádu je globálně konvergentní.
(2) Metoda největšího spádu používá pouze vektory dimenze n. Vyžaduje tedy $O(n)$ paměťových míst a $O(n)$ operací na iteraci.

Nevýhody:

(3) Metoda největšího spádu vyžaduje přesný výběr délky kroku.
(4) Metoda největšího spádu je pouze R-lineárně konvergentní s asymptotickou rychlostí

$$\limsup_{i\to\infty} \|x_i - x^*\|^{1/i} \leq \frac{\kappa(G(x^*)) - 1}{\kappa(G(x^*)) + 1}.$$

Odhad asymptotické rychlosti je obvykle realistický (není nadhodnocený). Jestliže $\kappa(G(x^*)) = 10^3$, potřebujeme ke snížení chyby $\|x - x^*\|$ o 4 řády zhruba 4600 iterací a jestliže $\kappa(G(x^*)) = 10^6$, potřebujeme ke snížení chyby $\|x - x^*\|$ o 8 řádů zhruba 9200000 iterací.

Newtonova metoda je definována vztahy

$$s_i = -G^{-1}(x_i)g(x_i),$$

$$\alpha_i = 1.$$

Výhody:

(1) Newtonova metoda je Q – kvadraticky konvergentní. Pokud tato metoda konverguje, stačí k nalezení lokálního minima pouze několik iterací.
(2) Newtonova metoda používá jednoduchý výběr délky kroku.

Nevýhody:

(3) Newtonova metoda není globálně konvergentní. Pokud x_1 je daleko od x^*, nemusí tato metoda konvergovat.
(4) Newtonova metoda používá matici řádu n a je třeba řešit soustavu n lineárních rovnic. Vyžaduje tedy $O(n^2)$ paměťových míst a $O(n^3)$ operací na iteraci.
(5) Je třeba počítat druhé derivace.

Aby se odstranily nevýhody těchto jednoduchých metod, byly vyvinuty důmyslnější a tudíž i složitější metody. Můžeme je zhruba rozdělit na metody spádových směrů a metody s lokálně omezeným krokem. Metody spádových směrů byly vyvinuty z metody největšího spádu. Předně byl odstraněn požadavek přesného výběru délky kroku, který byl nahražen slabšími (Wolfeho) podmínkami. Dále byla použitím principu sdružených směrů podstatně urychlena konvergence. Výsledkem tohoto vývoje jsou metody sdružených gradientů a metody s proměnnou metrikou.

Metody s lokálně omezeným krokem byly vyvinuty z Newtonovy metody tak, aby byla zaručena jejich globální konvergence i v případě, že Hessova matice není pozitivně definitní. Dále byl snížen počet operací, tím že není třeba hledat optimální lokálně omezený krok, stačí pouze nepřesné iterační přiblížení. Výsledkem jsou modifikace nepřesné Newtonovy metody s lokálně omezeným krokem a hybridní metody pro minimalizaci součtu čtverců.

1.5 Testování optimalizačních metod

Numerické testování a porovnávání optimalizačních metod je nezbytné k potvrzení jejich účinnosti a vhodnosti jejich použití pro zvolenou třídu optimalizačních úloh. Nejznámější sbírkou testovacích úloh je testovací prostředí CUTE [10], které obsahuje stovky úloh různých typů. Každá úloha je popsána posloupností příkazů speciálního vstupního jazyka v souboru typu SIF (standard input format) v adresáři SIF souborů. Prostředí CUTE je opatřeno preprocesorem, který z každého SIF souboru vygeneruje procedury realizující funkce popisující optimalizační úlohu. Některé optimalizační systémy, například systém UFO [110], obsahují rozhraní, které dovoluje přímo využívat prostředí CUTE. Nevýhodou prostředí CUTE je skutečnost, že nelze generovat procedury Používající více testovacích funkcí v jednom běhu výpočtu (každou testovací úlohu je třeba použít zvlášť).

Kromě prostředí CUTE existuje celá řada specializovaných sbírek testovacích úloh. V této práci jsou uváděny výsledky získané pomocí systému UFO [110], který lze nalézt na `www.cs.cas.cz/luksan/ufo`. V následující tabulce jsou uvedeny některé sbírky testovacích úloh pro minimalizaci bez omezení, které jsou součástí systému UFO. Sbírky TEST11, TEST14, TEST15, TEST18, TEST25, TEST28 jsou dostupné na ŵww.cs.cas.cz/luksan/test.

Sbírka	Typ úlohy	počet úloh	citace
TEST01	Obecná hustá	15	–
TEST11	Obecná hustá	58	[107]
TEST12	Obecná hustá	73	[3]
TEST14	Obecná řídká	22	[112]
TEST15	Řídký součet čtverců	24	[112]
TEST18	Řídká soustava rovnic	44	[112]
TEST24	Hustý součet čtverců	102	–
TEST25	Obecná řídká	82	[106]
TEST26	Řídký součet čtverců	60	–
TEST28	Obecná hustá	92	[114]
TEST37	Hustá soustava rovnic	64	–

úlohu považujeme za řídkou, je-li zadána řídká struktura Hessovy nebo Jacobiovy matice. Husté úlohy ve sbírkách uvedených v tabulce jso většinou také řídké, ale není zadána řídká struktura Hessovy nebo Jacobiovy matice.

Pro porovnání optimalizačních metod je výhodné používat výkonnostní profily, které berou v úvahu výsledky odpovídající jednotlivým úlohám. Výkonnostní profil $\rho_m(\tau)$ je definován vztahem

$$\rho_m(\tau) = \frac{\text{počet úloh kde } \log_2(\tau_{p,m}) \leq \tau}{\text{celkový počet úloh}},$$

kde $0 \leq \tau \leq \overline{\tau}$ a $\tau_{p,m}$ je podíl zvolené hodnoty (počtu vyčíslení hodnoty funkce nebo času výpočtu) potřebné k řešení úlohy p metodou m k nejmenší hodnotě potřebné k řešení úlohy p. Podíl $\tau_{p,m}$ je nekonečný (nebo velmi veliký) pokud metoda m nedokáže vyřešit úlohu p. Hodnota $\rho_m(\tau)$ pro $\tau = 0$ udává relativní počet testovacích úloh, pro které je metoda m nejúspěšnější (hodnota zvolené veličiny je minimální) a hodnota $\rho_m(\tau)$ pro $\tau = \infty$ udává relativní počet testovacích úloh které metoda m dokáže vyřešit. výkonnostních profilů lze určit účinnost a robustnost dané metody: čím výše leží odpovídající křivka, tím je daná metoda efektivnější.

2 Metody spádových směrů

Metody spádových směrů jsou nejjednodušší třídou optimalizačních metod, která vznikla zobecněním metody největšího spádu. Směrový vektor s_i vystupující v (21) se vybírá tak, aby platilo $F(x_i + \alpha s_i) < F(x_i)$ pro $0 < \alpha < \tilde{\alpha}_i$, kde $\tilde{\alpha}_i > 0$. Jestliže $F \in C^1$, je tato podmínka splněna pokud $s_i^T g(x_i) < 0$.

2.1 Základní vlastnosti metod spádových směrů

V tomto oddílu budeme předpokládat, že $s_i \neq 0$ a $g_i \neq 0$ $\forall i \in N$ a označíme

$$\cos\theta_i = -\frac{s_i^T g_i}{\|s_i\|\|g_i\|} \tag{23}$$

směrové kosíny úhlů, které svírají směrové vektory s_i, $i \in N$, se záporně vzatými gradienty. Klíčový význam pro konstrukci metod spádových směrů má pojem spádovosti směrových vektorů.

Definice 15. *Řekneme, že směrové vektory $s_i \in R^n$, $i \in N$, jsou spádové, jestliže platí*

$$\cos\theta_i > 0 \quad \forall i \in N. \tag{S1a}$$

Řekneme, že směrové vektory $s_i \in R^n$, $i \in N$, jsou stejnoměrně spádové, jestliže existuje konstanta $0 < \varepsilon_0 \leq 1$ taková, že platí

$$\cos\theta_i \geq \varepsilon_0 \quad \forall i \in N. \tag{S1b}$$

Řekneme, že směrové vektory $s_i \in R^n$, $i \in N$, jsou dostatečně spádové, jestliže platí

$$\cos\theta_i \geq 1/C_i \quad \forall i \in N \tag{S1c}$$

a čísla C_i, $i \in N$, vyhovují rekurentním nerovnostem

$$C_{i+1} \leq C_i + \overline{C}\|d_i\|,$$

kde $d_i = x_{i+1} - x_i = \alpha_i s_i$ a kde $C_1 \geq 1$ a $\overline{C} \geq 0$ jsou vhodné konstanty.

Poznámka 17. Metoda (S1a) největšího spádu je metodou stejnoměně spádových směrů, neboť $s_i = -g_i$, takže $s_i^T g_i = -\|g_i\|^2 = -\|s_i\|\|g_i\|$ a (S1b) platí pro $\varepsilon_0 = 1$.

Poznámka 18. Definice dostatečné spádovosti se může zdát dosti umělá. Nicméně je tato vlastnost často velmi užitečná (věta 222). Podmínka dostatečné spádovosti je jedním z principů omezeného znehodnocení, které se používají v konvergenčních důkazech. Poznamenejme, že z rekurentních nerovností pro čísla C_i, $i \in N$, plyne vztah

$$C_i \leq C_1 + \sum_{j=1}^{i-1} \overline{C}\|d_j\| \leq C_1 + \sum_{j=1}^{i} \overline{C}\|d_j\|. \tag{24}$$

Jsou-li směrové vektory stejnoměrně spádové, jsou též dostatečně spádové (stačí položit $C_1 = 1/\varepsilon_0$ a $\overline{C} = 0$). Za určitých předpokladů platí i obrácená implikace (věta 16).

Směrové vektory $s_i \in R^n$, $i \in N$, se často určují řešením soustav lineárních rovnic $B_i s_i = -g_i$.

Věta 10. *Nechť $B_i s_i = -g_i$, kde B_i, $i \in N$, je posloupnost symetrických pozitivně definitních matic. Pak platí*

$$\cos^2\theta_i \geq \frac{1}{\kappa_i} \qquad \forall i \in N, \tag{25}$$

kde κ_i je spektrální číslo podmíněnosti matice B_i.

Důkaz Podle předpokladu platí

$$-g_i = B_i s_i$$

a

$$-s_i = B_i^{-1} g_i,$$

takže

$$-s_i^T g_i = s_i^T B_i s_i \geq \underline{\lambda}_i \|s_i\|^2$$

a

$$-s_i^T g_i = g_i^T B_i^{-1} g_i \geq \frac{1}{\overline{\lambda}_i} \|g_i\|^2,$$

kde $\underline{\lambda}_i$ a $\overline{\lambda}_i$ je nejmenší a největší vlastní číslo matice B_i. Vynásobíme-li obě tyto nerovnosti, dostaneme

$$(-s_i^T g_i)^2 \geq \frac{\underline{\lambda}_i}{\overline{\lambda}_i} \|s_i\|^2 \|g_i\|^2 = \frac{1}{\kappa_i} \|s_i\|^2 \|g_i\|^2,$$

takže $\cos^2 \theta_i \geq 1/\kappa_i$. □

Poznámka 19. Podle věty 70 (vztah (253)) platí odhad (25), určuje-li se směrový vektor s_i metodou sdružených gradientů aplikovanou na soustavu lineárních rovnic $B_i s_i = -g_i$. Přitom soustavu lineárních rovnic není nutné řešit přesně, odhad platí v každém iteračním kroku metody sdružených gradientů.

Poznámka 20. Platí-li odhad (25), lze podmínky kladené na $\cos^2 \theta_i$ nahradit stejnými podmínkami, ve kterých místo $\cos^2 \theta_i$ vystupuje $1/\kappa_i$. Například podmínku (S1c) lze nahradit podmínkou $\kappa_i \leq C_i^2$, $i \in N$, kde $C_1 \geq 1$ a $C_{i+1} \leq C_i + \overline{C}\|d_i\|$, $i \in N$.

Další významnou součástí metod spádových směrů je výběr délky kroku, na který je třeba klást řadu omezení. Ukážeme na příkladech jaké problémy mohou nastat, pokud neklademe na délky kroku žádná omezení.

Příklad 2. Uvažujme kvadratickou funkci $F : R \to R$ danou vzorcem $F(x) = x^2$ a iterační proces (21), kde $x_1 = 2$ a $s_i = (-1)^i$, $\alpha_i = 2 + 3/2^i$, $i \in N$. Indukčním postupem se můžeme přesvědčit, že

$$\begin{aligned}
x_i &= (-1)^{i-1} \frac{2^{i-1}+1}{2^{i-1}}, \\
F(x_{i+1}) - F(x_i) &= \left(\frac{2^i+1}{2^i}\right)^2 - \left(\frac{2^{i-1}+1}{2^{i-1}}\right)^2 \\
&\leq 2\left(\frac{2^i+1}{2^i} - \frac{2^{i-1}+1}{2^{i-1}}\right) = -\frac{1}{2^{i-1}} \to 0, \\
\alpha_i s_i^T g(x_i) &= -2\frac{2^{i-1}+1}{2^{i-1}}\left(2 + \frac{3}{2^i}\right) \to -4.
\end{aligned}$$

Posloupnost $F(x_i)$, $i \in N$, je klesající, směrové vektory s_i, $i \in N$, jsou spádové (dokonce stejnoměrně), ale posloupnost x_i, $i \in N$, má dva hromadné body 1, -1, které nejsou stacionárními body funkce F (tím je bod $x^* = 0$, který je zároveň globálním minimem). Potíž je v tom, že kroky jsou příliš dlouhé a skutečný pokles $F(x_{i+1}) - F(x_i)$ je malý ve srovnání s předpovězeným poklesem $\alpha_i s_i^T g(x_i)$.

Příklad 3. Uvažujme kvadratickou funkci $F : R \to R$ danou vzorcem $F(x) = x^2$ a iterační proces (21), kde $x_1 = 2$ a $s_i = -1$, $\alpha_i = 1/2^i$, $i \in N$. Pak

$$x_i = 2 - \sum_{j=1}^{i-1} \frac{1}{2^j} = 2 - \left(1 - \frac{1}{2^{i-1}}\right) = \frac{2^{i-1}+1}{2^{i-1}} \to 1,$$

$$\frac{s_i^T g(x_{i+1})}{s_i^T g(x_i)} = \frac{2^i+1}{2^i+2} \to 1.$$

Posloupnost $F(x_i)$, $i \in N$, je klesající, směrové vektory s_i, $i \in N$, jsou spádové (dokonce stejnoměrně), ale posloupnost x_i, $i \in N$, má hromadný bod 1, který není stacionárním bodem funkce F. Potíž je v tom, že kroky jsou příliš krátké a směrová derivace se dostatečně nezvětší.

Abychom vyloučili tyto případy, je třeba klást na délky kroku $\alpha_i > 0$, $i \in N$, další dodatečné podmínky.

Definice 16. *Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje Armijovu podmínku, jestliže existuje číslo $0 < \varepsilon_1 < 1$ (nezávislé na indexu $i \in N$) takové, že*

$$F_{i+1} - F_i \le \varepsilon_1 \alpha_i s_i^T g_i. \tag{S2a}$$

Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje Wolfeho podmínku, jestliže existují čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ a $\varepsilon_3 \ge 0$ (nezávislá na indexu $i \in N$) taková, že platí (S2a) a

$$\varepsilon_2 s_i^T g_i \le s_i^T g_{i+1} \le \varepsilon_3 |s_i^T g_i|. \tag{S3a}$$

Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje Goldsteinovu podmínku, existují-li čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) taková, že platí (S2a) a

$$F_{i+1} - F_i \ge \varepsilon_2 \alpha_i s_i^T g_i. \tag{S3b}$$

Poznámka 21. Při vyšetřování globální konvergence vystačíme s nerovnostmi $0 < \varepsilon_1 < \varepsilon_2 < 1$. Pro zaručení superlineární konvergence (věta 20) je třeba, aby platilo $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$ (v případě podmínky (S3b) navíc $1/2 < \varepsilon_2 < 1$).

Poznámka 22. Existují různé varianty Wolfeho podmínky:

(a) Slabá Wolfeho podmínka, kdy $\varepsilon_3 = \infty$, takže druhá nerovnost v (S3a) odpadne.
(b) Zobecměná Wolfeho podmínka, kdy $0 \le \varepsilon_3 < \infty$.
(c) Silná Wolfeho podmínka, kdy $\varepsilon_3 = \varepsilon_2$.
(d) Přesný výběr délky kroku, kdy $\varepsilon_3 = \varepsilon_2 = 0$, takže $s_i^T g_{i+1} = 0$. Obvykle se přesným výběrem délky kroku rozumí nalezení lokálního minima funkce $F(x_i + \alpha s_i)$ s nejmenší hodnotou parametru α.

Poznámka 23. Armijova podmínka (S2a) je součástí zbylých dvou podmínek. Samotatně ji lze použít v Armijově výběru délky kroku. V tomto případě je $\alpha_i > 0$ prvním členem vyhovující podmínce (S2a) v posloupnosti α_i^j, $j \in N$, takové, že $\underline{\alpha}\|g_i\|/\|s_i\| \le \alpha_i^1 \le \overline{\alpha}\|g_i\|/\|s_i\|$, a

$$\underline{\beta}\alpha_i^j \le \alpha_i^{j+1} \le \overline{\beta}\alpha_i^j \quad \forall j \in N,$$

kde $0 < \underline{\alpha} \le \overline{\alpha}$ a $0 < \underline{\beta} \le \overline{\beta} < 1$ jsou konstanty nezávislé na indexu $i \in N$.

Podmínky (S1)–(S3) tvoří základ definice metod spádových směrů.

Definice 17. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou spádových směrů, jestliže směrové vektory $s_i \in R^n$, $i \in N$, jsou spádové (podmínka (S1a)) a délky kroku $\alpha_i > 0$, $i \in N$, se vybírají tak, aby byla splněna zobecněná Wolfeho podmínka, nebo Goldsteinova podmínka, nebo se použije Armijův výběr délky kroku. Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou stejnoměrně spádových směrů, je-li metodou spádových směrů a platí-li (S1b). Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou dostatečně spádových směrů, je-li metodou spádových směrů a platí-li (S1c).*

Poznámka 24. Při realizaci metod sdružených gradientů odvozených z metody největšího spádu se používá silná Wolfeho podmínka s $0 < \varepsilon_1 < \varepsilon_2 < 1/2$ a $\varepsilon_3 = \varepsilon_2$. Při realizaci metod s proměnnou metrikou odvozených z Newtonovy metody (kde $\alpha_i \to 1$ pro $i \to \infty$) se používá slabá Wolfeho podmínka s $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$ a $\varepsilon_3 = \infty$. Při realizaci metod založených na numerickém výpočtu gradientů se používá Goldsteinova podmínka s $0 < \varepsilon_1 < 1/2 < \varepsilon_2 < 1$ (obvykle $\varepsilon_2 = 1 - \varepsilon_1$). Při realizaci metod pro nehladké úlohy se používá Armijův výběr délky kroku s $0 < \varepsilon_1 < 1/2$.

Lemma 4. *(Konzistence) Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3 a směrový vektor $s_i \in R^n$ vyhovuje podmínce (S1a). Pak zobecněná Wolfeho podmínka, Goldsteinova podmínka i Armijův výběr délky kroku jsou konzistentní v tom smyslu, že existuje délka kroku $\alpha_i > 0$, která daný požadavek splňuje.*

Důkaz (a) Nechť $0 < \varepsilon_1 < \varepsilon_2 < 1$ a nechť $\tilde{\alpha}_i \geq 0$ je největší číslo takové, že

$$F(x_i + \alpha s_i) - F_i \leq \varepsilon_1 \alpha s_i^T g_i, \qquad 0 \leq \alpha \leq \tilde{\alpha}. \tag{26}$$

Jelikož $s_i^T g_i < 0$, platí $\tilde{\alpha}_i > 0$ (věta 1). Podle předpokladu F1 platí $F(x_i + \tilde{\alpha}_i s_i) \geq \underline{F}$, což spolu s (26) dává $\tilde{\alpha}_i \leq (\underline{F} - F_i)/(\varepsilon_1 s_i^T g_i)$, takže číslo $\tilde{\alpha}_i$ je konečné. Ukážeme nejprve, že

$$F(x_i + \tilde{\alpha}_i s_i) - F_i = \varepsilon_1 \tilde{\alpha}_i s_i^T g_i > \varepsilon_2 \tilde{\alpha}_i s_i^T g_i, \tag{27}$$

$$s_i^T g(x_i + \tilde{\alpha}_i s_i) \geq \varepsilon_1 s_i^T g_i > \varepsilon_2 s_i^T g_i \tag{28}$$

(neboť $s_i^T g_i < 0$), takže délka kroku $\alpha_i = \tilde{\alpha}_i$ splňuje slabou Wolfeho podmínku i Goldsteinovu podmínku Slabá Wolfeho podmínka i Goldsteinova podmínka jsou tedy konzistentní. Rovnost (27) plyne ze spojitosti funkce $F : \mathcal{D} \to R$, nerovnost (28) dokážeme sporem. Předpokládejme, že

$$s_i^T g(x_i + \tilde{\alpha}_i s_i) = \varepsilon s_i^T g_i < \varepsilon_1 s_i^T g_i \tag{29}$$

pro nějaké číslo $\varepsilon > \varepsilon_1$. Jelikož množina $\mathcal{D}$ je otevřená, existuje číslo $\delta > 0$ takové, že $x_i + \alpha s_i \in \mathcal{D}$, pokud $(\alpha - \tilde{\alpha}_i)\|s_i\| < \delta$. Pro takové $\alpha > \tilde{\alpha}_i$ podle (12), (27) a (29) platí

$$\begin{aligned} F(x_i + \alpha s_i) - F_i &\leq F(x_i + \tilde{\alpha}_i s_i) - F_i + s_i^T g(x_i + \tilde{\alpha}_i s_i)(\alpha - \tilde{\alpha}_i) + \overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2 \\ &= \varepsilon_1 \tilde{\alpha}_i s_i^T g_i + \varepsilon(\alpha - \tilde{\alpha}_i) s_i^T g_i + \overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2 \\ &= \varepsilon_1 \alpha s_i^T g_i - (\varepsilon_1 - \varepsilon)(\alpha - \tilde{\alpha}_i) s_i^T g_i + \overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2. \end{aligned}$$

Nechť $0 < \lambda < 1$ je libovolné číslo takové, že $\lambda(\varepsilon_1 - \varepsilon) s_i^T g_i / (\overline{G}\|s_i\|) < \delta$. Pak pro

$$\alpha = \tilde{\alpha}_i + \lambda \frac{(\varepsilon_1 - \varepsilon) s_i^T g_i}{\overline{G}\|s_i\|^2} > \tilde{\alpha}_i$$

platí $x_i + \alpha s_i \in \mathcal{D}$ a

$$F(x_i + \alpha s_i) - F_i \leq \varepsilon_1 \alpha s_i^T g_i - \lambda(1 - \lambda) \frac{(\varepsilon - \varepsilon_1)^2 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2} < \varepsilon_1 \alpha s_i^T g_i,$$

což je spor, neboť $\tilde{\alpha}_i$ je největší číslo splňující podmínku (26).

(b) Nechť $\varepsilon_3 \geq 0$. Jestliže $s_i^T g(x_i + \tilde{\alpha}_i s_i) \leq \varepsilon_3 |s_i^T g_i|$, splňuje délka kroku $\alpha = \tilde{\alpha}_i$ zobecněnou Wolfeho podmínku. Jestliže $s_i^T g(x_i + \tilde{\alpha}_i s_i) > \varepsilon_3 |s_i^T g_i|$, pak z nerovnosti $s_i^T g_i < 0$ a ze spojité diferencovatelnosti funkce F plyne existence čísla $0 < \alpha_i < \tilde{\alpha}_i$ takového, že $s_i^T g(x_i + \alpha_i s_i) = \varepsilon_3 |s_i^T g_i|$. Tato délka kroku splňuje podmínku (S3a) a podle (26) i podmínku (S2a), takže zobecněná Wolfeho podmínka je konzistentní.

(c) Jelikož $\tilde{\alpha}_i > 0$, $\overline{\alpha}\|g_i\|/\|s_i\| < \infty$ a $0 < \underline{\beta} < \overline{\beta} < 1$, existuje číslo $j \in N$ takové, že pro $\alpha_i = \alpha_i^j$ platí

$$0 < \underline{\beta}^{j-1}\overline{\alpha}\|g_i\|/\|s_i\| \leq \alpha_i \leq \overline{\beta}^{j-1}\overline{\alpha}\|g_i\|/\|s_i\| \leq \tilde{\alpha}_i,$$

což dokazuje konzistenci Armijova výběru délky kroku. □

V definici 15 záleží pouze na velikosti směrového kosínu $cos\,\theta_i$, nikoli na velikosti normy $\|s_i\|$. Z teoretických důvodů je však vhodné, aby tato norma nebyla příliš malá nebo příliš velká ve srovnání s normou gradientu.

Definice 18. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou gradientního typu, existují-li čísla $0 < \underline{s} \leq \overline{s}$ taková, že*

$$\underline{s}\|g_i\| \leq \|s_i\| \leq \overline{s}\|g_i\|, \qquad \forall i \in N \tag{30}$$

Je-li metoda spádových směrů metodou gradientního typu, můžeme podíl $\|g_i\|/\|s_i\|$, vystupující v definici Armijova výběru délky kroku (a také ve vzorcích použitých v dalších oddílech) vynechat, neboť v tomto případě platí $s_i \sim g_i$.

Poznámka 25. Jsou-li splněny předpoklady věty 10 (platí-li $B_i s_i = -g_i$) a existují-li čísla $0 < \underline{B} \leq \overline{B}$ taková, že $\underline{B} \leq \underline{\lambda}(B_i) \leq \overline{\lambda}(B_i) \leq \overline{B}$, $i \in N$, je výsledná optimalizační metoda metodou gradientního typu a platí $\underline{s} = 1/\overline{B}$, $\overline{s} = 1/\underline{B}$.

2.2 Globální konvergence

Nyní budeme studovat globální konvergenci metod spádových směrů. Nejprve dokážeme pomocnou větu, která zdůvodňuje použití podmínek (S2)–(S3).

Lemma 5. *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3, směrový vektor $s_i \in R^n$ vyhovuje podmínce (S1a) a délka kroku $\alpha_i > 0$ je určena Armijovým výběrem nebo tak, že splňuje podmínku (S2a) a některou z podmínek (S3a), (S3b). Pak existuje konstanta $\varepsilon_4 > 0$ taková, že pro libovolný index $i \in N$ platí*

$$\alpha_i \geq -\frac{\varepsilon_4 s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{\varepsilon_4 \cos\theta_i \|g_i\|}{\overline{G}\|s_i\|} \tag{31}$$

a

$$F_{i+1} - F_i \leq -\frac{\varepsilon_1\varepsilon_4 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_4}{\overline{G}} \cos^2\theta_i \|g_i\|^2. \tag{32}$$

Důkaz Nerovnost (32) plyne bezprostředně z nerovnosti (31), neboť podle (S2a) platí

$$F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \leq -\frac{\varepsilon_1\varepsilon_4 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_4 \cos^2\theta_i}{\overline{G}} \|g_i\|^2.$$

Zbývá tedy dokázat nerovnost (31).

(a) Předpokládejme nejprve, že $0 < \alpha_i \leq \tilde{\alpha}_i$, kde $\tilde{\alpha}_i > 0$ je hodnota použitá v důkazu lemmatu 4. Pak $x + \alpha s_i \in \mathcal{D}$ pro $0 < \alpha \leq \alpha_i$, takže lze použít předpoklad F3. Platí-li (S3a), můžeme podle (16) psát

$$\varepsilon_2 s_i^T g_i \le s_i^T g(x_i + \alpha_i s_i) \le s_i^T g_i + \alpha_i \overline{G} \|s_i\|^2,$$

neboli

$$\alpha_i \ge \frac{(\varepsilon_2 - 1) s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{(1-\varepsilon_2)\cos\theta_i \|g_i\|}{\overline{G}\|s_i\|},$$

takže platí (31) s $\varepsilon_4 = 1 - \varepsilon_2 > 0$. Platí-li (S3b), můžeme s použitím odhadu (12) psát

$$\varepsilon_2 \alpha_i s_i^T g_i \le F_{i+1} - F_i \le \alpha_i s_i^T g_i + \alpha_i^2 \overline{G} \|s_i\|^2,$$

neboli

$$\alpha_i \ge \frac{(\varepsilon_2 - 1) s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{(1-\varepsilon_2)\cos\theta_i \|g_i\|}{\overline{G}\|s_i\|},$$

takže platí (31) s $\varepsilon_4 = 1 - \varepsilon_2 > 0$.

(b) Nechť $\alpha_i \ge \tilde{\alpha}_i$. Protože hodnota $\tilde{\alpha}_i > 0$ splňuje podmínku (S3b) s $\varepsilon_2 = \varepsilon_1$ (důkaz lemmatu 4), platí

$$\alpha_i \ge \tilde{\alpha}_i \ge \frac{(\varepsilon_1 - 1) s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{(1-\varepsilon_1)\cos\theta_i \|g_i\|}{\overline{G}\|s_i\|}.$$

takže platí (31) s $\varepsilon_4 = 1 - \varepsilon_1 > 0$.

(c) Používáme-li Armijův výběr délky kroku, pak buď $\alpha_i = \alpha_i^1$, takže platí (31) s $\varepsilon_4 = \underline{\alpha}\overline{G} > 0$, nebo $\alpha_i \ge \tilde{\alpha}_i$, takže platí (31) s $\varepsilon_4 = 1 - \varepsilon_1 > 0$, nebo $\alpha_i \ge \underline{\beta}\tilde{\alpha}_i$ takže platí (31) s $\varepsilon_4 = \underline{\beta}(1 - \varepsilon_1) > 0$. □

Poznámka 26. Jsou-li splněny předpoklady lemmatu 5, platí

$$\sum_{i=1}^{\infty} \frac{(s_i^T g_i)^2}{\|s_i\|^2} < \infty. \tag{33}$$

To plyne bezprostředně z (32), neboť

$$F_1 - \underline{F} \ge \sum_{i=1}^{\infty} (F_i - F_{i+1}) \ge \sum_{i=1}^{\infty} \frac{\varepsilon_1 \varepsilon_4 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2}$$

a výraz na levé straně je konečný (podrobnější argumentaci lze nalézt v důkazu věty 11).

Poznámka 27. V některých případech, například u metod sdružených gradientů, lze místo nerovnosti (S1b) dokázat nerovnost

$$-s_i^T g_i \ge \underline{s}\|g_i\|^2, \qquad \forall i \in N, \tag{34}$$

kde $\underline{s} > 0$. Pak podle (33) platí

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} < \infty. \tag{35}$$

Jestliže navíc

$$\|s_i\| \le \overline{s}\|g_i\|, \qquad \forall i \in N, \tag{36}$$

kde $\overline{s} \ge \underline{s}$, je uvažovaná metoda metodou stejnoměrně spádových směrů gradientního typu (platí (30) a (S1b) s $\varepsilon_0 = \underline{s}/\overline{s}$).

Poznámka 28. Podmínka (34) sama o sobě nezaručuje globální konvergenci metody spádových směrů. Metoda je však globálně konvergentní, platí-li (34) a navíc (36) (poznámka 27) nebo

$$\sum_{i=1}^{\infty} \frac{1}{\|s_i\|^2} = \infty. \tag{37}$$

Jestliže totiž $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$ (poznámka 16), pak podle (37) platí

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} \geq \sum_{i=1}^{\infty} \frac{\underline{\varepsilon}^4}{\|s_i\|^2} = \infty,$$

což je ve sporu s (35). Rovnost (37) je splněna například tehdy, když $\|s_i\|^2 \leq \overline{c}\, i$, $i \in N$, pro nějakou konstantu $\overline{c} > 0$.

Věta 11. *(Globální konvergence) Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda spádových směrů, pro kterou platí*

$$\sum_{i=1}^{\infty} \cos^2 \theta_i = \infty \tag{38}$$

je globálně konvergentní.

Důkaz Použijeme-li (32), můžeme psát

$$F_{i+1} = F_1 + \sum_{j=1}^{i} (F_{j+1} - F_j) \leq F_1 - \frac{\varepsilon_1 \varepsilon_4}{\overline{G}} \sum_{j=1}^{i} \cos^2 \theta_j \|g_j\|^2.$$

Podle (32) je posloupnost F_i, $i \in N$ klesající a podle předpokladu F1 je zdola omezená. Existuje tedy limita

$$\underline{F} \leq \lim_{i \to \infty} F_i \leq F_1 - (\varepsilon_1 \varepsilon_4 / \overline{G}) \sum_{i=1}^{\infty} \cos^2 \theta_i \|g_i\|^2,$$

takže

$$\sum_{i=1}^{\infty} \cos^2 \theta_i \|g_i\|^2 \leq \frac{(F_1 - \underline{F})\overline{G}}{\varepsilon_1 \varepsilon_4} < \infty.$$

Předpokládejme, že neplatí $\liminf_{i \to \infty} \|g_i\| = 0$. Pak nutně existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$. Platí tedy

$$\underline{\varepsilon}^2 \sum_{i=1}^{\infty} \cos^2 \theta_i \leq \sum_{i=1}^{\infty} \cos^2 \theta_i \|g_i\|^2 < \infty,$$

což je ve sporu s předpokladem věty. □

Poznámka 29. Pro metodu stejnoměrně spádových směrů platí (S1b), takže

$$\varepsilon_0^2 \sum_{i=1}^{\infty} \|g_i\|^2 \leq \sum_{i=1}^{\infty} \cos^2 \theta_i \|g_i\|^2 < \infty,$$

což dává $\|g_i\| \to 0$ pro $i \to \infty$. Speciálně metoda největšího spádu, která generuje stejnoměrně spádové směry, je globálně konvergentní a platí $\|g_i\| \to 0$ pro $i \to \infty$.

Poznámka 30. Podle věty 10 a věty 11 je metoda spádových směrů používající směrové vektory $s_i \in R^n$, $i \in N$, určené řešením soustav lineárních rovnic $B_i s_i = -g_i$ globálně konvergentní, platí-li

$$\sum_{i=1}^{\infty} \frac{1}{\kappa_i} = \infty, \tag{39}$$

kde κ_i jsou spektrální čísla podmíněnosti matic B_i. Jestliže existuje číslo $\overline{\kappa} > 0$ takové že $\kappa_i \leq \overline{\kappa}\ \forall i \in N$, je tato metoda metodou stejnoměně spádových směrů (s $\varepsilon_0^2 = 1/\overline{\kappa}$) a platí $\|g_i\| \to 0$ pro $i \to \infty$.

Poznámka 31. Podmínka 38 je splněna například tehdy, existuje-li konstanta $0 < c \leq 1$ taková, že platí buď

$$\sum_{j=1}^{i} \cos^2 \theta_j \geq c\, i, \quad i \in N,$$

nebo

$$\cos^2 \theta_i \geq \frac{c}{i}, \qquad i \in N.$$

Označíme-li $\kappa_i = 1/cos^2\theta_i$, je poslední nerovnost splněna například tehdy, existuje-li konstanta $\kappa > 0$ taková, že $\kappa_i \leq \kappa\, i$, $i \in N$. Zvolíme-li $\kappa \geq \kappa_1$, je tato podmínka splněna, pokud $\kappa_{i+1} - \kappa_i \leq \kappa$, $i \in N$.

Poznámka 32. Větu 11 lze použít ke globalizaci metod spádových směrů pomocí restartování. Restartováním rozumíme přerušení a nové spuštění iteračního procesu tak, aby v iteračím kroku po restartu byla splněna podmínka (S1b) (pokládáme například $s_i = -g_i$). Restartování se provádí buď tehdy, je-li porušena podmínka (S1b), pak dostaneme stejnoměrnou metodu spádových směrů, nebo cyklicky v krocích s indexy $i = mk + 1$, kde $m \geq n$ a $k \in N$. Při cyklickém restartování platí

$$\sum_{i=1}^{\infty} \cos^2 \theta_i \geq \sum_{k=1}^{\infty} \cos^2 \theta_{mk+1} \geq \sum_{k=1}^{\infty} \varepsilon_0^2 = \infty,$$

takže jsou splňeny předpoklady věty 11 a metoda spádových směrů je globálně konvergentní.

Ukážeme, že metoda dostatečně spádových směrů je globálně konvergentní

Lemma 6. *Nechť $0 \leq z_i < 1$, $i \in N$, přičemž $\sum_{i=1}^{\infty} z_i < \infty$. Pak platí*

$$0 < \prod_{i=1}^{\infty}(1 - z_i) \leq \prod_{i=1}^{\infty}(1 + z_i) < \infty \tag{40}$$

Důkaz Z Taylorova rozvoje funkce $\exp(z_i)$ podle mocnin proměnné z_i plyne nerovnost $1 + z_i \leq \exp(z_i)$. Platí tedy

$$\prod_{i=1}^{\infty}(1 + z_i) \leq \exp\left(\sum_{i=1}^{\infty} z_i\right) < \infty, \tag{41}$$

což dokazuje pravou nerovnost v (40). Jelikož z $\sum_{i=1}^{\infty} z_i < \infty$ plyne $z_i \to 0$, existuje index $k \in N$ takový, že $z_i < 1/2$ a tedy $z_i/(1 - z_i) \leq 2z_i < 1\ \forall i \geq k$. Můžeme tedy psát

$$\begin{aligned} \frac{1}{\prod_{i=1}^{\infty}(1 - z_i)} &= \prod_{i=1}^{\infty}\left(1 + \frac{z_i}{1 - z_i}\right) \leq \prod_{i=1}^{k-1}\left(1 + \frac{z_i}{1 - z_i}\right) + \exp\left(\sum_{i=k}^{\infty} \frac{z_i}{1 - z_i}\right) \\ &\leq \prod_{i=1}^{k-1}\left(1 + \frac{z_i}{1 - z_i}\right) + 2\exp\left(\sum_{i=k}^{\infty} z_i\right) < \infty, \end{aligned}$$

což dokazuje levou nerovnost v (40). □

Věta 12. *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda dostatečně spádových směrů je globálně konvergentní.*

Důkaz Předpokládejme, že neplatí $\liminf_{i\to\infty} \|g_i\| = 0$. Pak nutně existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$. Použijeme-li předpoklad F1, nerovnost (S2a) a vztahy (23), (24), dostaneme (podobně jako v důkazu věty 11)

$$\begin{aligned} F_1 - \underline{F} &\geq \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq -\varepsilon_1 \sum_{i=1}^{\infty} d_i^T g_i = \varepsilon_1 \sum_{i=1}^{\infty} \cos\theta_i \|d_i\| \|g_i\| \\ &\geq \underline{\varepsilon}\varepsilon_1 \sum_{i=1}^{\infty} \cos\theta_i \|d_i\| \geq \frac{\underline{\varepsilon}\varepsilon_1}{\overline{C}} \sum_{i=1}^{\infty} \frac{\overline{C}\|d_i\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|d_j\|} \triangleq \frac{\underline{\varepsilon}\varepsilon_1}{\overline{C}} \sum_{i=1}^{\infty} z_i, \end{aligned}$$

takže součet na pravé straně je konečný. Můžeme tedy použít levou nerovnost v (40), s jejíž pomocí dostaneme

$$\prod_{i=1}^{\infty}\left(1 - \frac{\overline{C}\|d_i\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|d_j\|}\right) = \prod_{i=1}^{\infty}(1 - z_i) > 0.$$

Existuje tedy číslo $0 < \underline{C} < 1$ takové že

$$\underline{C} \leq \prod_{i=1}^{k}\left(1 - \frac{\overline{C}\|d_i\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|d_j\|}\right) = \prod_{i=1}^{k} \frac{C_1 + \sum_{j=1}^{i-1} \overline{C}\|d_j\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|d_j\|} = \frac{C_1}{C_1 + \sum_{j=1}^{k} \overline{C}\|d_j\|}$$

$\forall k \in N$, neboli

$$C_k \leq C_1 + \sum_{j=1}^{k} \overline{C}\|d_j\| \leq \frac{C_1}{\underline{C}},$$

což spolu s předpoklady věty dává $\cos\theta_k \geq 1/C_k \geq \varepsilon_0\ \forall k \in N$, kde $\varepsilon_0 = \underline{C}/C_1$. To je však spor, neboť stejnoměrná metoda spádových směrů je podle poznámky 29 globálně konvergentní. □

Ukážeme ještě jeden způsob, jak lze konstruovat globálně konvergentní metody pomocí korekcí směrových vektorů, který je obvykle šetrnější než způsob uvedený v poznámce 32.

Věta 13. *Uvažujme metodu spádových směrů, která používá směrové vektory*

$$s_i = -H_i g_i - \sigma\gamma_i \|H_i g_i\|\, g_i,$$

kde H_i, $i \in N$, jsou pozitivně semidefinitní matice takové, že $H_i g_i \neq 0$, kde $\gamma_i = \min(1, 1/\|g_i\|)$ a kde $\sigma > 0$ je číslo, které nezávisí na indexu $i \in N$. Splňuje-li funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ předpoklady F1 a F3, je tato metoda globálně konvergentní.

Důkaz Nechť $s_i = -H_i g_i - \sigma\gamma_i \|H_i g_i\|\, g_i$, kde $H_i g_i \neq 0$ a $g_i^T H_i g_i \geq 0$. Pak platí

$$\begin{aligned} s_i^T s_i &= \|H_i g_i\|^2 + 2\sigma\gamma_i g_i^T H_i g_i \|H_i g_i\| + \sigma^2\gamma_i^2 \|H_i g_i\|^2 \|g_i\|^2 \\ &\leq (1 + 2\sigma\gamma_i\|g_i\| + \sigma^2\gamma_i^2\|g_i\|^2)\|H_i g_i\|^2 = (1 + \sigma\gamma_i\|g_i\|)^2 \|H_i g_i\|^2 \end{aligned}$$

a

$$-s_i^T g_i = g_i^T H_i g_i + \sigma\gamma_i \|H_i g_i\| \|g_i\|^2 \geq \sigma\gamma_i \|H_i g_i\| \|g_i\|^2.$$

Můžeme tedy psát

$$\frac{-s_i^T g_i}{\|s_i\|\|g_i\|} \geq \frac{\sigma\gamma_i\|H_i g_i\|\|g_i\|^2}{(1+\sigma\gamma_i\|g_i\|)\|H_i g_i\|\|g_i\|} = \frac{\sigma\gamma_i\|g_i\|}{1+\sigma\gamma_i\|g_i\|}.$$

Předpokládejme, že neplatí $\liminf_{i\to\infty}\|g_i\| = 0$. Pak nutně existuje konstanta $0 < \underline{\varepsilon} < 1$ taková, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$. Pro $\gamma_i = 1/\|g_i\|$ platí

$$\frac{-s_i^T g_i}{\|s_i\|\|g_i\|} \geq \frac{\sigma}{1+\sigma} \geq \frac{\sigma\underline{\varepsilon}}{1+\sigma\underline{\varepsilon}}$$

a pro $\gamma_i = 1$ platí

$$\frac{-s_i^T g_i}{\|s_i\|\|g_i\|} \geq \frac{\sigma\|g_i\|}{1+\sigma\|g_i\|} \geq \frac{\sigma\underline{\varepsilon}}{1+\sigma\underline{\varepsilon}},$$

neboť funkce $t/(1+t)$ je pro $t > 0$ rostoucí. Směrové vektory s_i, $i \in N$, jsou tedy stejnoměrně spádové, což podle poznámky 29 implikuje $\|g_i\| \to 0$. To je však ve sporu s předpokladem, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$. □

Poznámka 33. Metody s proměnnou metrikou používají směrové vektory $s_i = -H_i g_i$, $i \in N$, kde H_i, $i \in N$, jsou pozitivně definitní matice. Věta 101 tvrdí, že metody s proměnnou metrikou jsou globálné konvergentní, splňuje-li funkce $F \in \mathcal{C}^2 : R^n \to R$ předpoklady F1, F4, F5. Bez požadavku stejnoměrné konvexity (předpoklad F5) tato věta neplatí. Věta 13 dává návod, jak lze metody s proměnnou metrikou korigovat tak, aby byly globálně konvergentní i tehdy, jsou-li splněny pouze předpoklady F1 a F3 (jiný způsob korekce metod s proměnnou metrikou je uveden v poznámce 171). Poznamenejme, že číslo $\sigma > 0$ volíme obvykle velmi malé, například $\sigma = 10^{-12}$.

Je-li metoda spádových směru globálně konvergentní (definice 14), nemusí ještě platit $x_i \to x^*$. Splňuje-li funkce $F : \mathcal{D}_F \to R$ předpoklad F2 nemůže posloupnost $x_i \in R^n$, $i \in N$, divergovat, může však mít více hromadných bodů. Ukážeme nyní, že vyhovuje-li nějaký hromadný bod $x^* \in R^n$ posloupnosti, generované metodou stejnoměrně spádových směrů, postačujícím podmínkám pro lokální minimum (věta 4), pak platí $x_i \to x^*$.

Věta 14. *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje předpoklady F1 a F2 a nechť $x^* \in \mathcal{D}$ je hromadným bodem posloupnosti $x_i \in R^n$, $i \in N$, generované metodou stejnoměrně spádových směrů. Pak, vyhovuje-li bod $x^* \in R^n$ předpokladům věty 4, platí $x_i \to x^*$.*

Důkaz Protože bod $x^* \in \mathcal{D}$ vyhovuje předpokladům věty 4, platí $g(x^*) = 0$ a $0 < \underline{\lambda}(G(x^*)) \leq \overline{\lambda}(G(x^*))$, kde $\underline{\lambda}(G(x^*))$ a $\overline{\lambda}(G(x^*))$ je nejmenší a největší vlastní číslo matice $G(x^*)$. Nechť

$$0 < \underline{G} < \underline{\lambda}(G(x^*)) \leq \overline{\lambda}(G(x^*)) < \overline{G}.$$

Ze spojitosti Hessovy matice $G(x)$ v okolí bodu $x^* \in \mathcal{D}$ plyne existence čísla ε takového, že

$$\underline{G}\|d\|^2 \leq d^T G(x) d \leq \overline{G}\|d\|^2 \quad \forall d \in R^n,$$

pokud $x \in \mathcal{B}(x^*, \varepsilon)$, takže podle (13)–(15) a (14)–(17) platí

$$F - F^* \leq \frac{1}{2}\overline{G}\|x - x^*\|^2, \tag{42}$$

$$F - F^* \geq \frac{1}{2}\underline{G}\|x - x^*\|^2, \tag{43}$$

$$\|g\| \leq \overline{G}\|x - x^*\|, \tag{44}$$

$$\|g\| \geq \underline{G}\|x - x^*\|, \tag{45}$$

pokud $x \in \mathcal{B}(x^*, \varepsilon)$. Protože $F_i \to F^*$, existuje index $l \in N$ takový, že

$$F_i - F^* < \frac{\underline{G}}{2}\varepsilon^2 \left(1 + \frac{\overline{G}^2}{2\varepsilon_0\varepsilon_1\underline{G}^2}\right)^{-2} \tag{46}$$

$\forall i \geq l$. Protože bod $x^* \in R^n$ je hromadným bodem posloupnosti $x_i \in R^n$, $i \in N$, existuje index $k \geq l$ takový, že $x_k \in \mathcal{B}(x^*, \varepsilon)$, takže podle (42) a (45) platí

$$F_k - F^* \leq \frac{1}{2}\overline{G}\|x_k - x^*\|^2 \leq \frac{\overline{G}}{2\underline{G}^2}\|g_k\|^2,$$

což spolu s $F_{k+1} \geq F^*$ a (S2b) dává

$$\alpha_k \leq \frac{F^* - F_k}{\varepsilon_1 s_k^T g_k} \leq \frac{F_k - F^*}{\varepsilon_0\varepsilon_1\|s_k\|\|g_k\|} \leq \frac{\overline{G}}{2\varepsilon_0\varepsilon_1\underline{G}^2}\frac{\|g_k\|}{\|s_k\|} \tag{47}$$

Použijeme-li tuto nerovnost spolu s (44), dostaneme

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \alpha_k\|s_k\| \leq \|x_k - x^*\| + \frac{\overline{G}}{2\varepsilon_0\varepsilon_1\underline{G}^2}\|g_k\| \leq \left(1 + \frac{\overline{G}^2}{2\varepsilon_0\varepsilon_1\underline{G}^2}\right)\|x_k - x^*\| \tag{48}$$

a podle (43) a (46) platí

$$\|x_k - x^*\| \leq \sqrt{\frac{2}{\underline{G}}(F_k - F^*)} < \varepsilon\left(1 + \frac{\overline{G}^2}{2\varepsilon_0\varepsilon_1\underline{G}^2}\right)^{-1},$$

což po dosazení do (48) dává $x_{k+1} \in \mathcal{B}(x^*, \varepsilon)$. Postupujeme-li takto dále, dostaneme $x_i \in \mathcal{B}(x^*, \varepsilon)$ $\forall i \geq k$ a tudíž i

$$\|x_i - x^*\| \leq \sqrt{\frac{2}{\underline{G}}(F_i - F^*)}$$

$\forall i \geq k$, což spolu s $F_i \to F^*$ dává $x_i \to x^*$. □

Poznámka 34. Věta 14 vyžaduje stejnoměrnou spádovost směrových vektorů. Abychom dostali podobný výsledek v obecném případě (kdy neplatí (S1b), takže nemůžeme použít nerovnost (47)), je třeba, aby délky kroku α_i, $i \in N$, splňovaly dodatečnou podmínku $\alpha_i \leq \overline{\alpha}\|g_i\|/\|s_i\|$. Tato podmínka není příliš omezující. Splňuje ji Armijův výběr délky kroku a také ostatní pravidla lze upravit tak aby platila (stačí položit $\alpha_i = \overline{\alpha}\|g_i\|/\|s_i\|$, kdykoliv hodnota α_i vychází větší). Splňují-li délky kroku tuto dodatečnou podmínku, platí obecnější věta, která nevyžaduje, aby funkce F byla dvakrát spojitě diferencovatelná.

Věta 15. *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$, která splňuje předpoklady F1–F3, má na $\mathcal{D}$ konečný počet stacionárních bodů. Nechť bod $x^* \in \mathcal{D}$ je hromadným bodem posloupnosti $x_i \in R^n$, $i \in N$, generované metodou stejnoměrně spádových směrů takovou, že $\alpha_i \leq \overline{\alpha}\|g_i\|/\|s_i\|$ $\forall i \in N$. Pak platí $x_i \to x^*$.*

Důkaz Jelikož hromadné body posloupnosti generované metodou stejnoměrně spádových směrů jsou podle poznámky 29 stacionárními body funkce F, je jich konečný počet. Předpokládejme, že posloupnost $x_i \in R^n$, $i \in N$, má $m > 1$ hromadných bodů x_k^*, $1 \leq k \leq m$, a označme $\varepsilon = \min_{1\leq k<l\leq m}\|x_l^* - x_k^*\|$. Podle definice hromadných bodů existují podposloupnosti x_{i_k}, $i_k \in K_k$, $1 \leq k \leq m$, přičemž $K_1 \cup \cdots \cup K_m = N$, takové, že $x_{i_k} \to x_k^*$. Existuje tedy index $i_1 \in N$ takový, že pokud $i \geq i_1$, platí $\|x_i - x_k^*\| < \varepsilon/3$ pro nějaký index $1 \leq k \leq m$. Jelikož $g_i \to 0$, existuje index $i_2 \in N$ takový, že pokud $i \geq i_2$, platí $\|g_i\| < \varepsilon/(3\,\overline{\alpha})$. Vzhledem k tomu, že nejmenší vzdálenost mezi jednotlivými hromadnými body je ε, nemůže pro všechny indexy $i \geq \max(i_1, i_2)$ platit $\|x_i - x_k^*\| < \varepsilon/3$ pro tentýž hromadný bod x_k^*. Existuje tedy index $i \geq \max(i_1, i_2)$ takový, že $\|x_i - x_k^*\| < \varepsilon/3$ a $\|x_{i+1} - x_l^*\| < \varepsilon/3$, kde $k \neq l$ Pak ale platí

$$\|x_l^* - x_k^*\| \leq \|x_l^* - x_{i+1}\| + \|x_{i+1} - x_i\| + \|x_k^* - x_i\| = \|x_{i+1} - x_l^*\| + \alpha_i\|s_i\| + \|x_i - x_k^*\| < \frac{2}{3}\varepsilon + \overline{\alpha}\|g_i\| < \varepsilon,$$

což je ve sporu s předpokladem, že $\min_{1\le k<l\le m}\|x_l^*-x_k^*\|=\varepsilon$. Posloupnost $x_i\in R^n$, $i\in N$, tedy může mít pouze jeden hromadný bod $x^*\in R^n$ a platí $x_i\to x^*$. □

V další části tohoto oddílu budeme předpokládat, že $x_i\to x^*$ a že bod $x^*\in R^n$ vyhovuje postačujícím podmínkám pro extrém (věta 4), takže existuje číslo $\varepsilon>0$, pro které platí (4) a (5) s $\mathcal{D}=\mathcal{B}(x^*,\varepsilon)$. Abychom nemuseli stále ověřovat zda pro daný index $i\in N$ již platí $x_i\in\mathcal{B}(x^*,\varepsilon)$, nahradíme předpoklady věty 4 silnějšími předpoklady F4 a F5. Tím se formálně zjednoduší a zpřehlední většina důkazů aniž by došlo k újmě na obecnosti. Nejprve ukážeme, že metoda dostatečně spádových směrů je za těchto předpokladů metodou stejnoměrně spádových směrů.

Věta 16. *Nechť $x_i\in R^n$, $i\in N$, je posloupnost generovaná metodou spádových směrů. Nechť $x_i\to x^*$, kde $x^*\in R^n$ je stacionárním bodem funkce $F\in\mathcal{C}^2:\mathcal{D}\to R$, která vyhovuje předpokladům F4 a F5. Pak jsou-li směrové vektory dostatečně spádové, jsou též stejnoměrně spádové.*

Důkaz Použijeme-li (S2a), definici čísla $\cos\theta_i$ a definici dostatečné spádovosti (nerovnost z poznámky 18), můžeme (podobně jako v důkazu věty 12) psát

$$\frac{F_i-F_{i+1}}{\|g_i\|}\ge\varepsilon_1\cos\theta_i\|d_i\|\ge\frac{\varepsilon_1}{\overline{C}}\frac{\overline{C}\|d_i\|}{C_1+\sum_{j=1}^i\overline{C}\|d_j\|}$$

$\forall i\in N$. Z druhé strany nerovnosti (42) a (45) implikují, že

$$\frac{F_i-F_{i+1}}{\|g_i\|}\le\frac{1}{\underline{G}}\sqrt{\frac{\overline{G}}{2}}\frac{(F_i-F^*)-(F_{i+1}-F^*)}{\sqrt{F_i-F^*}}\le\frac{\sqrt{2\overline{G}}}{\underline{G}}\left(\sqrt{F_i-F^*}-\sqrt{F_{i+1}-F^*}\right)$$

(neboť pro libovolná čísla $a\ge b>0$ platí $(a-b)/\sqrt{a}=(\sqrt{a}-\sqrt{b})(\sqrt{a}+\sqrt{b})/\sqrt{a}\le 2(\sqrt{a}-\sqrt{b})$), což po dosazení do předchozí nerovnosti dává

$$\frac{\varepsilon_1}{\overline{C}}\sum_{i=1}^{\infty}\frac{\overline{C}\|d_i\|}{C_1+\sum_{j=1}^i\overline{C}\|d_j\|}\le\sum_{i=1}^{\infty}\frac{F_i-F_{i+1}}{\|g_i\|}\le\frac{\sqrt{2\overline{G}}}{\underline{G}}\sqrt{F_1-F^*},$$

takže součet na levé straně je konečný. Postupujeme-li stejným způsobem jako v důkazu věty 12, dokážeme že existuje číslo $0<\underline{C}<1$ takové, že $\cos\theta_k\ge 1/C_k\ge\varepsilon_0\ \forall k\in N$, kde $\varepsilon_0=\underline{C}/C_1>0$. □

Jsou li splněny předpoklady věty 16, je metoda stejnoměrně spádových směrů (a tudíž i metoda dostatečně spádových směrů) lineárně konvergentní. Vyplývá to z následující věty, která je poněkud obecnější, neboť používá slabší podmínku uvedenou v poznámce 31.

Věta 17. *(Lineární konvergence) Nechť $x_i\in R^n$, $i\in N$, je posloupnost generovaná metodou spádových směrů takovou, že*

$$\sum_{j=1}^{i}cos^2\theta_j\ge\underline{c}\,i\quad\forall i\in N,$$

kde $0<\underline{c}\le 1$. Nechť $x_i\to x^*$, kde $x^*\in R^n$ je stacionárním bodem funkce $F\in\mathcal{C}^2:R^n\to R$, která vyhovuje přpokladům F4 a F5. Pak platí

$$\frac{\|x_{i+1}-x^*\|}{\|x_1-x^*\|}\le\sqrt{\frac{\overline{G}}{\underline{G}}}\ q^i,$$

kde $q=\sqrt{1-\underline{c}\varepsilon_1\varepsilon_4\underline{G}/\overline{G}}$ (ε_4 je číslo z lemmatu 5).

Důkaz Podle (14) platí $F^*-F\ge g^T(x^*-x)$, což po úpravě dává $F-F^*\le g^T(x-x^*)\le\|g\|\|x-x^*\|$ a použijeme-li (45), dostaneme

$$\|g\|^2 \geq \underline{G}(F - F^*). \tag{49}$$

Podle (32) tedy platí

$$\begin{aligned} F_{i+1} - F^* &\leq F_i - F^* - \frac{\cos^2\theta_i \varepsilon_1 \varepsilon_4}{\overline{G}} \|g_i\|^2 \leq \left(1 - \frac{\cos^2\theta_i \varepsilon_1 \varepsilon_4 \underline{G}}{\overline{G}}\right)(F_i - F^*) \\ &= \left(1 - \frac{\cos^2\theta_i}{\overline{c}}\right)(F_i - F^*) \end{aligned} \tag{50}$$

$\forall i \in N$, kde $\overline{c} = \overline{G}/(\varepsilon_1 \varepsilon_4 \underline{G})$. Použijeme-li tuto nerovnost několikrát po sobě, dostaneme

$$\frac{F_{i+1} - F^*}{F_1 - F^*} \leq \prod_{j=1}^{i} \left(1 - \frac{\cos^2\theta_j}{\overline{c}}\right) \leq \left[1 - \frac{1}{i}\sum_{j=1}^{i} \frac{\cos^2\theta_j}{\overline{c}}\right]^i \leq \left(1 - \frac{\underline{c}}{\overline{c}}\right)^i$$

(používáme lemma 2), což s použitím (42) a (43) dává

$$\frac{\|x_{i+1} - x^*\|}{\|x_1 - x^*\|} \leq \sqrt{\frac{\overline{G}}{\underline{G}}} \sqrt{\frac{F_{i+1} - F^*}{F_1 - F^*}} \leq \sqrt{\frac{\overline{G}}{\underline{G}}} q^i,$$

kde $q = \sqrt{1 - \underline{c}/\overline{c}} = \sqrt{1 - \underline{c}\varepsilon_1\varepsilon_4\underline{G}/\overline{G}}$. □

Poznámka 35. Z monotonie posloupnosti F_i, $i \in N$, a z nerovností (42), (43) plynou vztahy

$$\|e_{i+1}\| \leq \sqrt{\frac{\overline{G}}{\underline{G}}} \|e_i\|,$$

$$\|d_i\| = \|e_{i+1} - e_i\| \leq \|e_i\| + \|e_{i+1}\| \leq \left(1 + \sqrt{\frac{\overline{G}}{\underline{G}}}\right) \|e_i\|,$$

neboli $\|e_{i+1}\| = O(\|e_i\|)$ a $\|d_i\| = O(\|e_i\|)$.

Poznámka 36. Nechť jsou splněny předpoklady věty 17. Pak platí

$$\sum_{i=1}^{\infty} \|e_i\| = \sum_{i=1}^{\infty} \|x_i - x^*\| \leq \sqrt{\frac{\overline{G}}{\underline{G}}} \|x_1 - x^*\| \sum_{i=1}^{\infty} q^{i-1} = \sqrt{\frac{\overline{G}}{\underline{G}}} \|x_1 - x^*\| \frac{1}{1-q} < \infty$$

a také

$$\sum_{i=1}^{\infty} \|x_{i+1} - x_i\| \leq \sum_{i=1}^{\infty} (\|e_{i+1}\| + \|e_i\|) \leq 2 \sum_{i=1}^{\infty} \|e_i\| < \infty.$$

V některých případech, například u metod s proměnnou metrikou pro separovatelné úlohy (oddíl 10.5), není možné nalézt rozumné předpoklady pro globální konvergenci. Lze však zaručit lokální konvergenci, což znamená, že zvolíme-li počáteční bod x_1 dostatečné blízko k bodu x^*, platí $x_i \to x^*$.

Věta 18. *Nechť bod $x^* \in R^n$ vyhovuje předpokladům věty 4 (postačujícím podmínkám pro lokální minimum). Uvažujme metodu spádových směrů (definice 17) takovou, že $cos^2\theta_i \geq 1/\kappa_i$, kde $\kappa_1 \leq \overline{\kappa}$ a $\kappa_{i+1} \leq \kappa_i(1 + O(\|e_i\|)$. Pak existuje číslo $\delta > 0$ takové, že pokud $\|e_1\| < \delta$, platí $x_i \to x^*$ a $\sum_{i=1}^{\infty} \|e_i\| < \infty$.*

Důkaz Protože bod $x^* \in R^n$ vyhovuje předpokladům věty 4, existuje číslo $\varepsilon > 0$ takové, že jsou splněny předpoklady F4 a F5 v $\mathcal{B}(x^*, \varepsilon)$ (poznámka 5 a poznámka **??**).

(a) Předpokládejme, že $x_i \in \mathcal{B}(x^*, \varepsilon)$, $i \in N$. Ze vztahu $\kappa_{i+1} \leq \kappa_i(1 + O(\|e_i\|))$ a z nerovnosti (43) plyne existence konstanty $C > 0$ takové, že

$$\kappa_{i+1} \leq \kappa_i(1 + C\|e_i\|) \leq \kappa_i \left(1 + C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_i - F^*}\right),$$

takže podle (41) platí

$$\kappa_i \leq \kappa_1 \prod_{j=1}^{i-1} \left(1 + C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_j - F^*}\right) \leq \overline{\kappa} \exp\left(C\sqrt{\frac{2}{\underline{G}}} \sum_{j=1}^{i-1} \sqrt{F_j - F^*}\right). \tag{51}$$

Použijeme-li (50), můžeme pro $i \in N$ psát

$$\sqrt{F_{i+1} - F^*} \leq \sqrt{1 - \frac{1}{\kappa_i\, \overline{c}}}\sqrt{F_i - F^*} \leq \left(1 - \frac{1}{2\kappa_i\, \overline{c}}\right)\sqrt{F_i - F^*}$$

(kde $\kappa_i\, \overline{c} > 1$), neboť pro libovolné číslo $a > 0$ platí $\sqrt{1-a} \leq 1 - a/2$. Ukážeme, že bod x_1 lze volit tak, aby platilo $\kappa_i \leq 2\overline{\kappa}$ $\forall i \in N$. Předpokládejme, že $\kappa_j \leq 2\overline{\kappa}$ pro $1 \leq j \leq i-1$ (podle předpokladu to platí pro $i = 2$). Pak lze psát

$$\sum_{j=1}^{i-1} \sqrt{F_j - F^*} \leq \sqrt{F_1 - F^*} \sum_{j=1}^{i-1} \left(1 - \frac{1}{4\,\overline{\kappa}\,\overline{c}}\right)^{j-1} \leq \sqrt{F_1 - F^*} \sum_{j=1}^{\infty} \left(1 - \frac{1}{4\,\overline{\kappa}\,\overline{c}}\right)^{j-1} = 4\,\overline{\kappa}\,\overline{c}\sqrt{F_1 - F^*}.$$

Dosadíme-li tento vztah do (51), dostaneme

$$\kappa_i \leq \overline{\kappa} \exp\left(C\sqrt{\frac{2}{\underline{G}}} \sum_{j=1}^{i-1} \sqrt{F_j - F^*}\right) \leq \overline{\kappa} \exp\left(C\sqrt{\frac{2}{\underline{G}}} 4\,\overline{\kappa}\,\overline{c}\sqrt{F_1 - F^*}\right),$$

takže

$$\sqrt{F_1 - F^*} \leq \frac{1}{8\,\overline{\kappa}\,\overline{c}\,C}\sqrt{\frac{\underline{G}}{2}} \quad \Rightarrow \quad 4\,\overline{\kappa}\,\overline{c}\,C\sqrt{\frac{2}{\underline{G}}}\sqrt{F_1 - F^*} \leq \frac{1}{2} \quad \Rightarrow \quad \kappa_i \leq 2\overline{\kappa} \tag{52}$$

(neboť $\exp(1/2) < 2$). Tím jsme provedli indukční krok a dokázali, že volba počátečního bodu, splňující podmínku (52), implikuje nerovnosti $\kappa_i \leq 2\overline{\kappa}$, $i \in N$.

(b) Potřebujeme ještě, aby platilo $x_i \in \mathcal{B}(x^*, \varepsilon)$ (neboli $\|e_i\| < \varepsilon$) $\forall i \in N$. Jelikož posloupnost $F_i - F^*$, $i \in N$ je nerostoucí, lze to podle (43) zajistit volbou $\sqrt{F_1 - F^*} < \sqrt{\underline{G}/2}\,\varepsilon$. Zvolíme-li

$$\delta = \sqrt{\frac{2}{\overline{\underline{G}}}} \min\left(\frac{1}{8\,\overline{\kappa}\,\overline{c}\,C}\sqrt{\frac{\underline{G}}{2}}, \sqrt{\frac{\underline{G}}{2}}\,\varepsilon\right) = \sqrt{\frac{\underline{G}}{\overline{\underline{G}}}} \min\left(\frac{1}{8\,\overline{\kappa}\,\overline{c}\,C}, \varepsilon\right) \tag{53}$$

a $x_1 \in \mathcal{B}(x^*, \delta)$, platí $x_i \in \mathcal{B}(x^*, \varepsilon)$ a podle (a) též $\kappa_i \leq 2\overline{\kappa}$ $\forall i \in N$, takže uvažovaná metoda je podle věty 10 metodou stejnoměrně spádových směrů, podle věty 14 platí $x_i \to x^*$ a podle věty 17 je rychlost konvergence alespoň lineární (platí $\sum_{i=1}^{\infty} \|e_i\| < \infty$). □

Poznámka 37. Podmínku $\kappa_{i+1} \leq \kappa_i(1 + O(\|e_i\|))$, $i \in N$, vystupující ve větě 18, lze nahradit existencí čísel $\overline{\kappa} > 0$ a $C > 0$ takových, že platí (51).

Poznámka 38. Význam věty 18 spočívá v tom že metodu, která je lokálně konvergentní ale není globálně konvergentní, lze upravit (například pomocí restartů) tak, aby byla zaručena globální konvergence. Věta 18 pak říká, že dostaneme-li se do blízkosti minima, další restarty již nejsou nutné.

2.3 Asymptotická rychlost konvergence

Nyní se budeme zabývat asymptotickým chováním metod spádových směrů. Budeme přitom používat symboly $o(\xi_i)$, $O(\xi_i)$, relaci $u_i \sim v_i$ a pravidla uvedená v poznámce 10.

Definice 19. *Řekneme, že výběr délky kroku je asymptoticky přesný, jestliže*

$$\lim_{i\to\infty} \frac{s_i^T g_{i+1}}{s_i^T g_i} = 0.$$

Lemma 7. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou stejnoměrně spádových směrů s asymptoticky přesným výběrem délky kroku taková, že $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje předpokladům F4 a F5. Pak platí*

$$\alpha_i = -\frac{s_i^T g_i}{s_i^T G^* s_i}(1 + o(1))$$

a

$$F_{i+1} - F_i = -\frac{1}{2}\frac{(s_i^T g_i)^2}{s_i^T G^* s_i}(1 + o(1)).$$

Důkaz Podle věty 5 platí

$$g_i = G^* e_i + o(\|e_i\|),$$

což s použitím předpokladů F4 a F5 dává $g_i \sim e_i$, takže podle poznámky 35 platí $\|d_i\| = O(\|e_i\|) = O(\|g_i\|)$. Dále z (S1b) plyne $d_i^T g_i \sim \|d_i\|\|g_i\|$. Použijeme-li tyto vztahy a větu 5, můžeme psát

$$\frac{s_i^T g_{i+1}}{s_i^T g_i} = \frac{d_i^T g_{i+1}}{d_i^T g_i} = 1 + \frac{d_i^T G^* d_i + o(\|d_i\|^2)}{d_i^T g_i} = 1 + \alpha_i \frac{s_i^T G^* s_i}{s_i^T g_i} + o(1),$$

(neboť $\|d_i\|^2/d_i^T g_i \sim \|d_i\|^2/\|d_i\|\|g_i\| = O(1)$, takže

$$\alpha_i = -\frac{s_i^T g_i}{s_i^T G^* s_i}(1 + o(1)).$$

Podle věty 5 platí

$$F_{i+1} - F_i = \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G^* s_i + o(\|d_i\|^2).$$

Dosadíme-li do tohoto vyjádření vztah pro asympticky přesný výběr délky kroku, dostaneme

$$F_{i+1} - F_i = -\frac{(s_i^T g_i)^2}{s_i^T G^* s_i}(1 + o(1)) + \frac{1}{2}\frac{(s_i^T g_i)^2}{s_i^T G^* s_i}(1 + o(1)) + o(\|d_i\|^2) = -\frac{1}{2}\frac{(s_i^T g_i)^2}{s_i^T G^* s_i}(1 + o(1)),$$

neboť podle předpokladů F4 a F5 platí $d_i^T G^* d_i \sim \|d_i\|^2$ a tudíž

$$\frac{(s_i^T g_i)^2}{s_i^T G^* s_i} = \frac{(d_i^T g_i)^2}{d_i^T G^* d_i} \sim \frac{\|d_i\|^2\|g_i\|^2}{\|d_i\|^2} \sim \|d_i\|^2$$

(připomeňme že $(1 + o(1))^2 = 1 + o(1)$). □

Poznámka 39. Asymptoticky přesný výběr délky kroku dostaneme, vybíráme-li délku kroku pomocí kvadratické nebo kubické interpolace (věta 24).

Lemma 8. *Nechť B je symetrická pozitivně definitní matice. Vyhovují-li vektory $u \in R^n$ a $v \in R^n$ podmínce*

$$\frac{(u^T v)^2}{u^T u v^T v} \leq \varepsilon^2, \tag{54}$$

kde $0 \leq \varepsilon \leq 1$, platí

$$\frac{(u^T B v)^2}{u^T B u v^T B v} \leq \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2, \tag{55}$$

kde $\kappa(B)$ je spektrální číslo podmíněnosti matice B.

Důkaz Nechť vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce (54). Bez újmy na obecnosti budeme předpokládat, že $\|u\| = 1$, $\|v\| = 1$ a budeme používat označení $V = [u, v]$. Nechť vektor w je lineární kombinací vektorů u a v, přičemž $\|w\| = 1$ a $u^T w = 0$. Pak existují čísla α a β taková, že

$$v = \alpha u + \beta w$$

a přihlédneme-li k tomu, že $\|u\| = 1$ a $\|w\| = 1$, platí $u^T v = \alpha$ a $v^T v = \alpha^2 + \beta^2$. Z nerovnosti (54) pak plyne

$$\alpha^2 \leq \varepsilon^2$$

a předpoklad $\|v\| = 1$ dává

$$\alpha^2 + \beta^2 = 1.$$

Položme $W = [u, w]$. Pak zřejmě platí $V = WM$, kde

$$M = \begin{bmatrix} 1, & \alpha \\ 0, & \beta \end{bmatrix}.$$

Jelikož $V^T B V = M^T W^T B W M$, můžeme psát

$$\kappa(V^T B V) \leq \kappa(M^T M)\kappa(W^T B W).$$

Jelikož vektor w byl zvolen tak, aby platilo $W^T W = I$, dostaneme

$$\frac{x^T W^T B W x}{x^T x} = \frac{x^T W^T B W x}{x^T W^T W x} = \frac{y^T B y}{y^T y},$$

kde $y = Wx$, takže nutně $\underline{\lambda}(W^T B W) = \underline{\lambda}(B)$, $\overline{\lambda}(W^T B W) = \overline{\lambda}(B)$ a

$$\kappa(W^T B W) = \frac{\overline{\lambda}(W^T B W)}{\underline{\lambda}(W^T B W)} = \frac{\overline{\lambda}(B)}{\underline{\lambda}(B)} = \kappa(B).$$

Jelikož $\alpha^2 + \beta^2 = 1$, platí

$$M^T M = \begin{bmatrix} 1, & \alpha \\ \alpha, & 1 \end{bmatrix},$$

takže $\underline{\lambda}(M^T M) = 1 - |\alpha|$ a $\overline{\lambda}(M^T M) = 1 + |\alpha|$, což spolu s $\alpha^2 \leq \varepsilon^2$ dává

$$\kappa(M^T M) = \frac{\overline{\lambda}(M^T M)}{\underline{\lambda}(M^T M)} = \frac{1 + |\alpha|}{1 - |\alpha|} \leq \frac{1 + \varepsilon}{1 - \varepsilon}.$$

Můžeme tedy psát

$$\kappa(V^T B V) \leq \kappa(M^T M)\kappa(W^T B W) \leq \kappa(B)\frac{1 + \varepsilon}{1 - \varepsilon}.$$

Nechť $\underline{\lambda}$ a $\overline{\lambda}$ jsou vlastní čísla matice $V^T B V$ seřazená podle velikosti. Pak platí

$$\det(V^T B V) = \underline{\lambda}\overline{\lambda} = \underline{\lambda}^2 \kappa(V^T B V).$$

Z nerovnosti $\left(\sqrt{u^TBu} - \sqrt{v^TBv}\right)^2 \geq 0$ plyne, že

$$\sqrt{u^TBuv^TBv} \leq \frac{1}{2}(u^TBu + v^TBv) = \frac{1}{2}Tr(V^TBV) = \frac{1}{2}(\underline{\lambda} + \overline{\lambda}) = \frac{1}{2}\underline{\lambda}(1 + \kappa(V^TBV)).$$

Můžeme tedy psát

$$\begin{aligned}
\frac{(u^TBv)^2}{u^TBuv^TBv} &= 1 - \frac{\det(V^TBV)}{u^TBuv^TBv} \leq 1 - \frac{4\kappa(V^TBV)}{(1+\kappa(V^TBV))^2} \\
&= \left(\frac{\kappa(V^TBV)-1}{\kappa(V^TBV)+1}\right)^2 \leq \left(\frac{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon} - 1}{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon} + 1}\right)^2 \\
&= \left(\frac{\kappa(B) - 1 + (\kappa(B)+1)\varepsilon}{\kappa(B) + 1 + (\kappa(B)-1)\varepsilon}\right)^2
\end{aligned}$$

(funkce $(t-1)/(t+1)$ je pro kladná t rostoucí). □

Důsledek 1. Jsou-li splněny předpoklady lemmatu 8 s $\varepsilon = 0$ (takže u a v jsou ortogonální), platí

$$\frac{(u^TBv)^2}{u^TBuv^TBv} \leq \left(\frac{\kappa(B)-1}{\kappa(B)+1}\right)^2,$$

Tato nerovnost se nazývá Wielandtovou nerovností.

Lemma 9. *Nechť B je symetrická pozitivně definitní matice. Vyhovují-li vektory $u \in R^n$ a $v \in R^n$ podmínce*

$$\frac{(u^Tv)^2}{u^Tuv^Tv} \geq 1 - \varepsilon^2, \tag{56}$$

kde $0 \leq \varepsilon \leq 1$, platí

$$\frac{(u^Tv)^2}{u^TBuv^TB^{-1}v} \geq \frac{4\kappa(B)(1-\varepsilon^2)}{(\kappa(B)+1+(\kappa(B)-1)\varepsilon)^2}, \tag{57}$$

kde $\kappa(B)$ je spektrální číslo podmíněnosti matice B.

Důkaz Nechť vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce (56). Položme $w = BHv$, kde

$$H = B^{-1} - u(u^TBu)^{-1}u^T.$$

Pak platí

$$u^Tw = u^TB(B^{-1} - u(u^TBu)^{-1}u^T)v = u^Tv - u^TBu(u^TBu)^{-1}u^Tv = 0,$$

takže vektory u a w jsou ortogonální. Zvolme v R^n ortonormální bázi v_i, $1 \leq i \leq n$, tak, aby platilo $v_1 = u/\|u\|$ a $v_2 = w/\|w\|$. Pak lze psát

$$v = \sum_{i=1}^{n}(v_i^Tv)v_i$$

a

$$v^Tv = \sum_{i=1}^{n}(v_i^Tv)^2 \geq (v_1^Tv)^2 + (v_2^Tv)^2 = \frac{(u^Tv)^2}{u^Tu} + \frac{(w^Tv)^2}{w^Tw},$$

neboli

$$\frac{(w^Tv)^2}{w^Tw} \leq v^Tv - \frac{(u^Tv)^2}{u^Tu} = \frac{u^Tuv^Tv - (u^Tv)^2}{u^Tu},$$

takže

$$\frac{(w^Tv)^2}{w^Twv^Tv} \leq 1 - \frac{(u^Tv)^2}{u^Tuv^Tv} \leq \varepsilon^2,$$

a použijeme-li lemma 8, dostaneme

$$\frac{(w^T B^{-1} v)^2}{w^T B^{-1} w v^T B^{-1} v} \leq \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2$$

(protože $\kappa(B^{-1}) = \kappa(B)$). Z druhé strany (vzhledem k definici matice H, vektoru w a ortogonalitě $u^T w = 0$) platí

$$\begin{aligned} w^T B^{-1} w &= w^T B^{-1} B H v = w^T H v = w^T B^{-1} v - \frac{w^T u u^T v}{u^T B u} \\ &= w^T B^{-1} v = v^T H B B^{-1} v = v^T H v \end{aligned}$$

a

$$v^T H v = v^T B^{-1} v - \frac{(u^T v)^2}{u^T B u},$$

takže

$$\begin{aligned} \frac{(u^T v)^2}{u^T B u v^T B^{-1} v} &= 1 - \frac{v^T H v}{v^T B^{-1} v} = 1 - \frac{(v^T H v)^2}{v^T H v v^T B^{-1} v} = 1 - \frac{(w^T B^{-1} v)^2}{w^T B^{-1} w v^T B^{-1} v} \\ &\geq 1 - \left(\frac{\kappa(B) - 1 + (\kappa(B) + 1)\varepsilon}{\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon} \right)^2 = \frac{4\kappa(B)(1 - \varepsilon^2)}{(\kappa(B) + 1 + (\kappa(B) - 1)\varepsilon)^2}. \end{aligned}$$

□

Důsledek 2. Jsou-li splněny předpoklady lemmatu 9 s $\varepsilon = 0$ (takže u a v jsou lineárně závislé), platí

$$\frac{(u^T v)^2}{u^T B u v^T B^{-1} v} \geq \frac{4\kappa(B)}{(\kappa(B) + 1)^2}.$$

Tato nerovnost se nazývá Kantorovičovou nerovností.

Věta 19. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou stejnoměrně spádových směrů s asymptoticky přesným výběrem délky kroku taková, že $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje předpokladům F4 a F5. Pak platí*

$$\limsup_{i \to \infty} \|x_i - x^*\|^{1/i} \leq \frac{\kappa(G^*) - 1 + (\kappa(G^*) + 1)\sqrt{1 - \varepsilon_0^2}}{\kappa(G^*) + 1 + (\kappa(G^*) - 1)\sqrt{1 - \varepsilon_0^2}}. \tag{58}$$

Důkaz Podle věty 5 platí

$$\begin{aligned} F_i - F^* &= \frac{1}{2} e_i^T G^* e_i + o(\|e_i\|^2), \\ g_i &= G^* e_i + o(\|e_i\|), \end{aligned}$$

takže s použitím předpoklad'ru F4, F5 a toho, že $\|g_i\| \sim \|e_i\|$ dostaneme

$$\begin{aligned} e_i &= (G^*)^{-1} g_i (1 + o(1)), \\ F_i - F^* &= \frac{1}{2} g_i^T (G^*)^{-1} g_i (1 + o(1)). \end{aligned}$$

Použijeme-li lemma 7 můžeme psát

$$\frac{F_{i+1} - F^*}{F_i - F^*} = 1 + \frac{F_{i+1} - F_i}{F_i - F^*} = 1 - \frac{(s_i^T g_i)^2}{s_i^T G^* s_i g_i^T (G^*)^{-1} g_i} (1 + o(1)).$$

Podle (S1b) platí $(s_i^T g_i)^2 \geq \varepsilon_0^2 \|s_i\|^2 \|g_i\|^2$ takže s použitím lemmatu 9 dostaneme

$$\frac{(s_i^T g_i)^2}{s_i^T G^* s_i g_i^T (G^*)^{-1} g_i} \geq \frac{4\kappa(G^*)\varepsilon_0^2}{(\kappa(G^*)+1+(\kappa(G^*)-1)\sqrt{1-\varepsilon_0^2})^2},$$

což po dosazení do předchozí rovnosti dává

$$\frac{F_{i+1}-F^*}{F_i-F^*} \leq \left(\frac{(\kappa(G^*)-1+(\kappa(G^*)+1)\sqrt{1-\varepsilon_0^2}}{(\kappa(G^*)+1+(\kappa(G^*)-1)\sqrt{1-\varepsilon_0^2}}\right)^2 (1+o(1)) \triangleq \hat{q}^2(1+o(1)).$$

K libovolnému číslu $q, \hat{q} < q < 1$, tedy existuje index $k \in N$ takový, že $(F_{i+1}-F^*)/(F_i-F^*) \leq q^2$ pro $i \geq k$. Můžeme tedy postupovat stejně jako v důkazu věty 16, takže

$$\frac{F_i-F^*}{F_k-F^*} \leq q^{2(i-k)} \quad \Rightarrow \quad \frac{\|x_i-x^*\|}{\|x_k-x^*\|} \leq \sqrt{\frac{\overline{G}}{\underline{G}}} q^{i-k}$$

a podle věty 7 platí

$$\limsup_{i\to\infty} \|x_i-x^*\|^{1/i} \leq q.$$

Jelikož to platí pro libovolné číslo q, $\hat{q} < q < 1$, dokázali jsme tvrzení věty. □

Poznámka 40. Pro metodu největšího spádu je $\varepsilon_0 = 1$, takže

$$\limsup_{i\to\infty} \|x_i-x^*\|^{1/i} \leq \frac{\kappa(G^*)-1}{\kappa(G^*)+1}. \tag{59}$$

Poznámka 41. Používáme-li směrové vektory $s_i = -H_i g_i$, platí

$$\limsup_{i\to\infty} \|x_i-x^*\|^{1/i} \leq \limsup_{i\to\infty} \frac{\kappa(R_i)-1}{\kappa(R_i)+1},$$

kde $R_i = (G^*)^{-1/2} B_i (G^*)^{-1/2}$, neboť matice R_i mají stejná vlastní čísla jako matice $\tilde{R}_i = B_i^{1/2}(G^*)^{-1}B_i^{1/2}$ a položíme-li $z_i = B_i^{1/2} s_i = -B_i^{-1/2} g_i$, můžeme stejně jako v důkazu věty 19 psát

$$\begin{aligned} \frac{F_{i+1}-F^*}{F_i-F^*} &= 1 - \frac{(s_i^T g_i)^2}{s_i^T G^* s_i g_i^T (G^*)^{-1} g_i}(1+o(1)) \\ &= 1 - \frac{(z_i^T z_i)^2}{z_i^T \tilde{R}_i^{-1} z_i z_i^T \tilde{R}_i z_i}(1+o(1)) \end{aligned}$$

a použitím lemmatu 9 dostaneme

$$\frac{F_{i+1}-F^*}{F_i-F^*} \leq \left(\frac{\kappa(\tilde{R}_i)-1}{\kappa(\tilde{R}_i)+1}\right)^2 (1+o(1)) = \left(\frac{\kappa(R_i)-1}{\kappa(R_i)+1}\right)^2 (1+o(1)).$$

Poznámka 42. Metoda největšího spádu používá směrový vektor $s_i = -g_i$, $i \in N$, takže platí (S1b), přičemž můžeme volit $\varepsilon_0 = 1$. Z toho je vidět, že metoda největšího spádu má výborné vlastnosti z hlediska teorie vyložené v oddílech 2.2 a 2.3. Je globálně konvergentní, přičemž asymptotická rychlost konvergence uvedená ve větě 19 je optimální (platí $\varepsilon_0 = 1$). Přesto je tato metoda neúčinná. Je to způsobeno tím, že nejsou splněny podmínky pro vícekrokovou kvadratickou konvergenci (metoda největšího spádu obvykle nenajde minimum kvadratické funkce po konečném počtu kroků), ani podmínky pro superlineární konvergenci (posloupnost g_i, $i \in N$, obvykle nekonverguje k vlastnímu vektoru matice G^*, takže neplatí (61)). Přitom odhad (59) je realistický, jak je ukázáno v následujícím příkladu.

Příklad 4. Uvažujeme kvadratickou funkci $F(x) = (1/2)x^T Gx$, kde $x \in R^2$ a G je pozitivně definitní matice. Nech $\underline{\lambda}$, $\overline{\lambda}$ jsou vlastní čísla matice G příslušná ortonormálním vlastním vektorům $\underline{v}$, $\overline{v}$ a nech $\kappa = \kappa(G) = \overline{\lambda}/\underline{\lambda}$. Zvolme počáteční bod $x_1 \in R^2$ tak, že $x_1 = c_1(\underline{v} + \nu_1\overline{v})$, kde $\nu_1^2\kappa^2 = 1$. Pak platí

$$g_1 = Gx_1 = c_1\underline{\lambda}(\underline{v} + \nu_1\overline{v}),$$

takže

$$\alpha_1 = \frac{g_1^T g_1}{g_1^T G g_1} = \frac{c_1^2\underline{\lambda}^2(1+\nu_1^2\kappa^2)}{c_1^2\underline{\lambda}^3(1+\nu_1^2\kappa^3)} = \frac{2}{\underline{\lambda}(\kappa+1)}\,.$$

Tato hodnota realizuje (pro kvadratickou funkci) přesný výběr délky kroku, takže

$$x_2 = x_1 - \alpha_1 g_1 = c_1(\underline{v} + \nu_1\overline{v}) - \frac{2c_1}{\kappa+1}\,(\underline{v} + \nu_1\kappa\overline{v}) = c_1\frac{\kappa-1}{\kappa+1}\,(\underline{v} - \nu_1\overline{v}) = c_2(\underline{v} + \nu_2\overline{v}),$$

kde $c_2 = c_1(\kappa-1)/(\kappa+1)$ a $\nu_2 = -\nu_1$. Platí tedy

$$\frac{\|e_2\|}{\|e_1\|} = \frac{\|x_2\|}{\|x_1\|} = \frac{c_2\sqrt{1+\nu_2^2}}{c_1\sqrt{1+\nu_1^2}} = \frac{\kappa-1}{\kappa+1}\,.$$

Jelikož $\nu_2^2\kappa^2 = \nu_1^2\kappa^2 = 1$, můžeme pokračovat dále, takže nakonec dostaneme

$$\frac{\|e_{i+1}\|}{\|e_i\|} = \frac{\kappa-1}{\kappa+1}, \quad \forall i \in N.$$

Pokud $\kappa \gg 1$, je rychlost konvergence velmi pomalá (například při $\kappa = 1000$ se chyba sníží o čtyři řády zhruba po 4600 iteračních krocích). V každém iteračním kroku platí

$$g_i = Gx_i = c_i\underline{\lambda}(\underline{v} + \nu_i\kappa\overline{v})$$

a

$$\alpha_i = \frac{2}{\underline{\lambda}(\kappa+1)}\,,$$

takže

$$\|d_i\| = \alpha_i\|g_i\| = \frac{2}{\underline{\lambda}(\kappa+1)}\,c_i\underline{\lambda}\sqrt{1+\nu_2^2\kappa^2} = \frac{2\sqrt{2}}{\kappa+1}\,c_i.$$

Z posledního výrazu je patrné, že při $\kappa \gg 1$ se délka kroku neúměrně zkracuje faktorem $\kappa+1$. Tento jev se projeví i při minimalizaci obecné nekvadratické funkce v oblastech, kde $\kappa(G(x)) \gg 1$, například v okolí dna strmého údolí.

Nyní se budeme zabývat superlineární konvergencí metod spádových směrů, které určují směrový vektor přibližným řešením soustavy rovnic $B_i s_i = -g_i$. Budeme používat označení

$$\omega_i = \frac{B_i s_i + g_i}{\|g_i\|}, \qquad \vartheta_i = \frac{(B_i - G_i)s_i}{\|s_i\|} \tag{60}$$

a budeme předpokládat, že konstanty ε_1 a ε_2 v podmínkách (S2)–(S3) vyhovují nerovnostem uvedeným v poznámce 21, tedy že platí $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$ a v případě podmínky (S3b) též $1/2 < \varepsilon_2 < 1$.

Věta 20. *(Superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů taková, že $x_i \to x^*$, kde $x^* \in \mathcal{D}$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňujícím postačující podmínky druhého řádu pro lokální minimum (matice $G^* = G(x^*)$ je pozitivně definitní). Nechť $\omega_i \to 0$, $\vartheta_i \to 0$, neboli*

$$\lim_{i\to\infty} \frac{\|B_i s_i + g_i\|}{\|g_i\|} = 0, \qquad \lim_{i\to\infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0, \tag{61}$$

a nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2)–(S3). Pak existuje index $k \in N$ takový, že $\alpha_i = 1$ $\forall i \geq k$, a posloupnost x_i, $i \in N$, konverguje superlineárně k bodu $x^* \in R^n$.

Důkaz (a) Nechť $0 < \underline{G} < \underline{\lambda}(G^*) \leq \overline{\lambda}(G^*) < \overline{G}$. Ukážeme, že existuje index $k_1 \in N$ takový, že

$$\frac{1}{\overline{G}}\|g_i\| \leq \|s_i\| \leq \frac{1}{\underline{G}}\|g_i\| \tag{62}$$

pro $i \geq k_1$. Použijeme-li označení (60), můžeme psát

$$G_i s_i = (B_i s_i + g_i) - (B_i - G_i)s_i - g_i = \omega_i\|g_i\| - \vartheta_i\|s_i\| - g_i, \tag{63}$$

takže

$$\begin{aligned}(\overline{\lambda}(G_i) + \|\vartheta_i\|)\|s_i\| &\geq (1 - \|\omega_i\|)\|g_i\|,\\ (\underline{\lambda}(G_i) - \|\vartheta_i\|)\|s_i\| &\leq (1 + \|\omega_i\|)\|g_i\|.\end{aligned}$$

neboli

$$\frac{1 - \|\omega_i\|}{\overline{\lambda}(G_i) + \|\vartheta_i\|}\|g_i\| \leq \|s_i\| \leq \frac{1 + \|\omega_i\|}{\underline{\lambda}(G_i) - \|\vartheta_i\|}\|g_i\|. \tag{64}$$

Jelikož $\|\vartheta_i\| \to 0$, $\|\omega_i\| \to 0$ (podle (61)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, $\overline{\lambda}(G_i) \to \overline{\lambda}(G^*) < \overline{G}$, existuje index $k_1 \in N$ takový, že pro $i \geq k_1$ platí

$$\frac{1}{\overline{G}} \leq \frac{1 - \|\omega_i\|}{\overline{\lambda}(G_i) + \|\vartheta_i\|}, \qquad \frac{1 + \|\omega_i\|}{\underline{\lambda}(G_i) - \|\vartheta_i\|} \leq \frac{1}{\underline{G}},$$

což spolu s (64) dává (62).

(b) Ukážeme, že existuje index $k_2 \geq k_1$ takový, že $-s_i^T g_i \geq (\underline{G}/\overline{G})\|s_i\|\|g_i\|$ pro $i \geq k_2$. Z definice vektorů ω_i, ϑ_i (vztah (60)) a z (62) plyne, že

$$\begin{aligned}-s_i^T g_i &= s_i^T(G_i s_i + (B_i - G_i)s_i - (B_i s_i + g_i)) \geq (\underline{\lambda}(G_i) - \|\vartheta_i\|)\|s_i\|^2 - \|\omega_i\|\|s_i\|\|g_i\|\\ &\geq \frac{1}{\overline{G}}\left(\underline{\lambda}(G_i) - \|\vartheta_i\| - \overline{G}\|\omega_i\|\right)\|s_i\|\|g_i\|\end{aligned}$$

a jelikož $\|\omega_i\| \to 0$, $\|\vartheta_i\| \to 0$ (podle (61)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, existuje index $k_2 \geq k_1$ takový, že $-s_i^T g_i \geq (\underline{G}/\overline{G})\|s_i\|\|g_i\|$ pro $i \geq k_2$.

(c) Ukážeme, že existuje index $k \geq k_2$ takový, že pro $i \geq k$ hodnota $\alpha_i = 1$ vyhovuje podmínkám (S2)–(S3). Označme

$$\eta_i = \frac{s_i^T g_i + s_i^T G_i s_i}{s_i^T g_i}. \tag{65}$$

Použijeme-li (b), dostaneme

$$|\eta_i| = \frac{|s_i^T g_i + s_i^T G_i s_i|}{|s_i^T g_i|} \leq \frac{\overline{G}\|s_i\|\|g_i + G_i s_i\|}{\underline{G}\|s_i\|\|g_i\|} \leq \frac{\overline{G}}{\underline{G}}\left(\frac{\|g_i + B_i s_i\|}{\|g_i\|} + \frac{\|(B_i - G_i)s_i\|}{\|g_i\|}\right)$$

pro $i \geq k_1$, takže podle (61) a (a) platí $|\eta_i| \to 0$. Nyní použijeme větu 5, podle které

$$F(x_i + s_i) - F(x_i) = s_i^T g_i + \frac{1}{2}s_i^T G_i s_i + o(\|s_i\|^2),$$

$$s_i^T g(x_i + s_i) = s_i^T g_i + s_i^T G_i s_i + o(\|s_i\|^2).$$

Můžeme tedy psát

$$\lim_{i\to\infty}\frac{F(x_i + s_i) - F(x_i)}{s_i^T g_i} = \frac{1}{2} + \lim_{i\to\infty}(\frac{1}{2}\eta_i + o(1)) = \frac{1}{2},$$

$$\lim_{i\to\infty}\frac{s_i^T g(x_i + s_i)}{s_i^T g_i} = \lim_{i\to\infty}(\eta_i + o(1)) = 0,$$

neboť $|s_i^T g_i| \sim \|s_i\|^2$ podle (a) a (b). Protože $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$ (v případě podmínky (S3b) též $1/2 < \varepsilon_2 < 1$), existuje index $k \geq k_2$ takový, že (S2) a (S3) (s $\alpha_i = 1$) platí pro $i \geq k$.

(d) Podle (c) pro $i \geq k$ platí $d_i = s_i$, neboli $g_{i+1} = g(x_i + s_i)$, a podle věty 5 lze psát

$$g(x_i + s_i) = g_i + G_i s_i + o(\|s_i\|).$$

Použijeme-li (44), (45) a (62), dostaneme

$$\begin{aligned}
\frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} &\leq \frac{\overline{G}}{\underline{G}} \frac{\|g_{i+1}\|}{\|g_i\|} \\
&\leq \frac{\overline{G}}{\underline{G}} \left(\frac{\|g(x_i + s_i) - g_i - G_i s_i\|}{\|g_i\|} + \frac{\|G_i s_i - B_i s_i\|}{\|g_i\|} + \frac{\|B_i s_i + g_i\|}{\|g_i\|} \right) \\
&\leq \frac{\overline{G}}{\underline{G}} \left(\frac{o(\|s_i\|)}{\underline{G}\|s_i\|} + \frac{\|(B_i - G_i)s_i\|}{\underline{G}\|s_i\|} + \frac{\|B_i s_i + g_i\|}{\|g_i\|} \right),
\end{aligned}$$

takže podle (61) platí

$$\lim_{i \to \infty} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} = 0.$$

□

Podmínky (61) splňuje přesná Newtonova metoda, kdy $\vartheta_i = 0$ a $\omega_i = 0$, $i \in N$. Pro nepřesnou Newtonovu metodu, kdy $\vartheta_i = 0$ a $\omega_i \neq 0$, je třeba zajistit, aby platilo $\omega_i \to 0$. V případě diferenční verze Newtonovy metody obvykle neplatí $\vartheta_i \to 0$, ale tato hodnota bývá velmi malá. Nejsou tedy splněny podmínky pro superlineární konvergenci, nicméně metoda konverguje velmi rychle lineárně. Abychom mohli studovat konvergenční vlastnosti metod, pro které neplatí (61), ale veličiny ϑ_i a ω_i jsou dostatečně malé, ukážeme, že tyto metody lze považovat za nepřesné Newtonovy metody, kdy $\|G_i s_i + g_i\|/\|g_i\| \leq \overline{\omega}'$, přičemž $\overline{\omega}' < 1$. Abychom zjednodušili argumentaci, budeme ve smyslu poznámky **??** předpokládat, že funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ vyhovuje předpokladům F4 a F5.

Věta 21. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů taková, že $x_i \to x^*$, kde $x^* \in \mathcal{D}$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje předpokladům F4 a F5. Nechť $\|\omega_i\| \leq \overline{\omega}$ a $\|\vartheta_i\| \leq \overline{\vartheta}$, $i \in N$, kde $\overline{\omega} < 1$ a $\overline{\vartheta} < (1 - \overline{\omega})\,\underline{G}/2$. Pak pro $i \in N$ platí*

$$\frac{\|G_i s_i + g_i\|}{\|g_i\|} \leq \overline{\omega}' < 1, \quad \text{kde} \quad \overline{\omega}' = \frac{\overline{\omega}\,\underline{G} + \overline{\vartheta}}{\underline{G} - \overline{\vartheta}}. \tag{66}$$

Důkaz Jelikož $B_i s_i = (B_i s_i + g_i) - g_i$, můžeme psát $\|B_i s_i\| \leq (1 + \|\omega_i\|)\|g_i\|$, takže dostaneme

$$(1 + \|\omega_i\|)\|g_i\| \geq \|B_i s_i\| \geq \|G_i s_i\| - \|(B_i - G_i)s_i\| \geq (\underline{G} - \|\vartheta_i\|)\|s_i\| > 0, \tag{67}$$

neboť podle předpokladu platí $\|\vartheta_i\| \leq \overline{\vartheta} < \underline{G}(1 - \overline{\omega})/2 < \underline{G}$. Označme $\omega_i' = (G_i s_i + g_i)/\|g_i\|$. Ze vztahu

$$\|G_i s_i + g_i\| \leq \|B_i s_i + g_i\| + \|(B_i - G_i)s_i\|$$

dostaneme $\|\omega_i'\| \leq \|\omega_i\| + \|\vartheta_i\|\|s_i\|/\|g_i\|$, což spolu s (67) dává

$$\|\omega_i'\| \leq \|\omega_i\| + \frac{1 + \|\omega_i\|}{\underline{G} - \|\vartheta_i\|}\|\vartheta_i\| = \frac{\|\omega_i\|\underline{G} + \|\vartheta_i\|}{\underline{G} - \|\vartheta_i\|} \leq \frac{\overline{\omega}\underline{G} + \overline{\vartheta}}{\underline{G} - \overline{\vartheta}} = \overline{\omega}',$$

neboť $\|\omega_i\| \leq \overline{\omega} < 1$ a $\|\vartheta_i\| \leq \overline{\vartheta}$. Jelikož $\overline{\vartheta} < (1 - \overline{\omega})\underline{G}/2$, platí

$$\overline{\omega}' = \frac{\overline{\omega}\underline{G} + \overline{\vartheta}}{\underline{G} - \overline{\vartheta}} < \frac{\overline{\omega} + (1 - \overline{\omega})/2}{1 - (1 - \overline{\omega})/2} = \frac{\overline{\omega} + 1}{1 + \overline{\omega}} = 1$$

. □

Věta 22. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů taková, že $x_i \to x^*$, kde $x^* \in \mathcal{D}$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňujícím postačující podmínky druhého řádu pro lokální minimum (matice $G^* = G(x^*)$ je pozitivně definitní). Nechť pro $i \in N$ platí*

$$\frac{\|B_i s_i + g_i\|}{\|g_i\|} \le \overline{\omega}, \qquad \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} \le \overline{\vartheta}, \tag{68}$$

kde

$$\overline{\omega} \le \frac{1}{1+\kappa+\kappa/\overline{\eta}} < \frac{1}{1+\kappa} < 1, \qquad \overline{\vartheta} \le \frac{1-(1+\kappa+\kappa/\overline{\eta})\,\overline{\omega}}{2+\kappa+\kappa/\overline{\eta}}\,\underline{G} < \frac{1-(1+\kappa)\,\overline{\omega}}{2+\kappa}\,\underline{G} < \frac{1-\overline{\omega}}{2}\,\underline{G}, \tag{69}$$

přičemž $0 < \overline{\eta} < \min(1-2\varepsilon_1, \varepsilon_2) < 1$ a $\kappa = \overline{G}/\underline{G} \ge 1$, a nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2)–(S3). Pak existuje index $k \in N$ takový, že $\alpha_i = 1$, $i \ge k$, a posloupnost x_i, $i \in N$, konverguje lineárně k bodu $x^* \in R^n$ s asymptotickou rychlostí alespoň $\overline{\omega}' = (\overline{\omega}\,\underline{G} + \overline{\vartheta})/(\underline{G} - \overline{\vartheta})$.

Důkaz Nechť $0 < \underline{G} < \underline{\lambda}(G^*) \le \overline{\lambda}(G^*) < \overline{G}$. Jelikož $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$ a $\overline{\lambda}(G_i) \to \overline{\lambda}(G^*) < \overline{G}$, existuje index $k_1 \in N$ takový, že $\underline{G} \le \underline{\lambda}(G_i) \le \overline{\lambda}(G_i) \le \overline{G}$ $\forall i \ge k_1$. Jelikož podle předpokladu (69) platí $\overline{\omega} < 1$ a $\overline{\vartheta} < (1-\overline{\omega})\,\underline{G}/2$, můžeme používat vztah (66).

(a) Ukážeme, že pro $i \ge k_1$ platí

$$\frac{1-\overline{\omega}'}{\overline{G}}\|g_i\| \le \|s_i\| \le \frac{1+\overline{\omega}'}{\underline{G}}\|g_i\|, \tag{70}$$

kde $0 \le \overline{\omega}' < 1$ je číslo definované v (66). Zřejmě $G_i s_i = (G_i s_i + g_i) - g_i$. Použijeme-li (66), můžeme psát

$$\overline{G}\|s_i\| \ge \|G_i s_i\| \ge \|g_i\| - \|G_i s_i + g_i\| \ge (1-\overline{\omega}')\|g_i\|,$$

$$\underline{G}\|s_i\| \le \|G_i s_i\| \le \|g_i\| + \|G_i s_i + g_i\| \le (1+\overline{\omega}')\|g_i\|.$$

Spojíme-li obě tyto nerovnosti, dostaneme po úpravě (70).

(b) Ukážeme, že pro $i \ge k_1$ platí $\cos\theta_i > 0$. Použijeme-li (70), dostaneme

$$\begin{aligned} -s_i^T g_i = s_i^T G_i s_i - s_i^T(G_i s_i + g_i) &\ge \underline{G}\|s_i\|^2 - \overline{\omega}'\|s_i\|\|g_i\| \ge \left(\underline{G}\frac{1-\overline{\omega}'}{\overline{G}} - \overline{\omega}'\right)\|s_i\|\|g_i\| \\ &\ge \frac{1}{\kappa}(1-(1+\kappa)\,\overline{\omega}')\|s_i\|\|g_i\|, \end{aligned} \tag{71}$$

takže $\cos\theta_i > 0$, pokud $\overline{\omega}' < 1/(1+\kappa)$. Použijeme-li definici čísla $\overline{\omega}'$ (vzorec (66)), můžeme nerovnost $\overline{\omega}' < 1/(1+\kappa)$ zapsat ve tvaru

$$\frac{\overline{\omega}\,\underline{G} + \overline{\vartheta}}{\underline{G} - \overline{\vartheta}} < \frac{1}{1+\kappa} \quad \Leftrightarrow \quad (2+\kappa)\,\overline{\vartheta} < \underline{G}(1-(1+\kappa)\,\overline{\omega}),$$

což platí právě tehdy, když

$$\overline{\omega} < \frac{1}{1+\kappa}, \qquad \overline{\vartheta} < \frac{1-(1+\kappa)\,\overline{\omega}}{2+\kappa}\,\underline{G}.$$

Tyto nerovnosti plynou z předpokladu (69).

(c) Nechť η_i, $i \in N$, jsou veličiny určené vzorcem (65). Budeme hledat postačující podmínku pro to, aby platilo $\|\eta_i\| \le \overline{\eta}$, $i \in N$. Podle (66) platí

$$\|\eta_i\| = \frac{|s_i^T g_i + s_i^T G_i s_i|}{|s_i^T g_i|} \le \frac{\kappa\,\overline{\omega}'\,\|s_i\|\|g_i\|}{(1-(1+\kappa)\,\overline{\omega}')\,\|s_i\|\|g_i\|} = \frac{\kappa\,\overline{\omega}'}{1-(1+\kappa)\,\overline{\omega}'},$$

takže nerovnost $\|\eta_i\| < \overline{\eta}$ je splněna tehdy, když

$$\frac{\kappa\,\overline{\omega}'}{1-(1+\kappa)\,\overline{\omega}'} \le \overline{\eta} \quad \Leftrightarrow \quad (\kappa+\overline{\eta}(1+\kappa))\,\overline{\omega}' \le \overline{\eta},$$

neboli

$$\overline{\omega}' \le \frac{\overline{\eta}}{\kappa+(1+\kappa)\overline{\eta}} = \frac{1}{(1+\kappa)+\kappa/\overline{\eta}}.$$

Použijeme-li (66), dostaneme

$$\frac{\overline{\omega}\,\underline{G}+\overline{\vartheta}}{\underline{G}-\overline{\vartheta}} \le \frac{\overline{\eta}}{\kappa+(1+\kappa)\overline{\eta}} \quad \Leftrightarrow \quad \overline{\vartheta}(\kappa+\overline{\eta}(2+\kappa)) \le \underline{G}(\overline{\eta}-\overline{\omega}(\kappa+\overline{\eta}(1+\kappa))).$$

Tato podmínka je splněna právě tehdy, když

$$\overline{\omega} \le \frac{1}{1+\kappa+\kappa/\overline{\eta}}, \qquad \overline{\vartheta} \le \frac{1-(1+\kappa+\kappa/\overline{\eta})\,\overline{\omega}}{2+\kappa+\kappa/\overline{\eta}}\;\underline{G},$$

což jsou nerovnosti použité v předpokladu (69). Podobně jako v části (c) důkazu věty 20, můžeme psát

$$\lim_{i\to\infty}\frac{F(x_i+s_i)-F(x_i)}{s_i^T g_i} = \frac{1}{2}+\lim_{i\to\infty}(\frac{1}{2}\eta_i+o(1)) \ge \frac{1}{2}(1-\overline{\eta}),$$

$$\lim_{i\to\infty}\frac{s_i^T g(x_i+s_i)}{s_i^T g_i} = \lim_{i\to\infty}(\eta_i+o(1)) \le \overline{\eta},$$

a jelikož $\overline{\eta} < \min(1-2\varepsilon_1, \varepsilon_2)$, existuje index $k_2 \ge k_1$ takový, že (S2) a (S3) (s $\alpha_i = 1$) platí pro $i \ge k_2$.

(d) Podle (c) pro $i \ge k_2$ platí $d_i = s_i$, neboli $g_{i+1} = g(x_i+s_i)$, a podle věty 5 lze psát $g(x_i+s_i) = g_i + G_i s_i + o(\|s_i\|)$. Použijeme-li (44), (45) a (70), dostaneme

$$\begin{aligned}\lim_{i\to\infty}\frac{\|g_{i+1}\|}{\|g_i\|} &\le \lim_{i\to\infty}\left(\frac{\|g(x_i+s_i)-g_i-G_i s_i\|}{\|g_i\|}+\frac{\|G_i s_i+g_i\|}{\|g_i\|}\right)\\ &\le \lim_{i\to\infty}\left(\frac{2\,o(\|s_i\|)}{\underline{G}\|s_i\|}+\frac{\|G_i s_i+g_i\|}{\|g_i\|}\right) = \lim_{i\to\infty}\frac{\|G_i s_i+g_i\|}{\|g_i\|} \le \overline{\omega}'\end{aligned}$$

(neboť $1+\overline{\omega}' < 2$). K libovolnému číslu q, $\overline{\omega}' < q < 1$, tedy existuje index $k \ge k_2$ takový, že $\|g_{i+1}\|/\|g_i\| \le q$ pro $i \ge k$. Můžeme tedy postupovat stejně jako v důkazu věty 16, takže

$$\frac{\|g_i\|}{\|g_k\|} \le q^{i-k} \quad \Rightarrow \quad \frac{\|x_i-x^*\|}{\|x_k-x^*\|} \le \frac{\overline{G}}{\underline{G}}q^{i-k}$$

a podle věty 7 platí

$$\limsup_{i\to\infty}\|x_i-x^*\|^{\frac{1}{i}} \le q.$$

Jelikož to platí pro libovolné číslo q, $\overline{\omega}' < q < 1$, konverguje posloupnost x_i, $i \in N$, lineárně k bodu $x^* \in R^n$ s asymptotickou rychlostí alespoň $\overline{\omega}'$. □

Poznámka 43. Teoretický význam věty 22 spočívá v tom, že dává odhad asymptotické rychlosti konvergence pro zadané horní meze $0 \le \overline{\omega} < 1$ a $0 \le \overline{\vartheta} < (1-\overline{\omega})\,\underline{G}/2$. Praktický význam nerovností (69), které zajišťují, aby platilo $\alpha_i = 1$ pro $i \ge k_2$, je však malý, neboť tyto nerovnosti jsou značně nadhodnocené (v jejich odvození bereme vždy ten nejhorší případ). To znamená, že odhad $\overline{\omega}'$ obvykle platí i pro mnohem větší hodnoty $\overline{\omega}$ a $\overline{\vartheta}$. O maticích B_i, $i \in N$, nepředpokládáme nic jiného, než že platí (68), takže věta 22 je vhodná zejména k vyšetřování diferenčních verzí Newtonovy metody, kdy umíme nalézt hodnotu $\overline{\vartheta}$ pro kterou platí $\|B_i - G_i\| \le \overline{\vartheta}$, $i \in N$ (věta **??**).

2.4 Výběr délky kroku

Nyní se budeme zabývat implementací metod spádových směrů, jejichž iterační krok má tvar $x_{i+1} = x_i + \alpha_i s_i$. Popíšeme nejprve algoritmus pro výběr délky kroku $\alpha_i > 0$, který používá slabou Wolfeho podmínku (pro ostatní podmínky je třeba tento algoritmus mírně modifikovat).

Algoritmus 1. Data $0 < \beta_1 < \beta_2 < 1 < \gamma_1 < \gamma_2$.

Krok 1 Zvolíme počáteční délku kroku $\alpha > 0$. Položíme $\overline{\alpha} := 0$.

Krok 2 Položíme $\underline{\alpha} := \overline{\alpha}$ a $\overline{\alpha} := \alpha$. Jsou-li splněny podmínky (S2a) a (S3a) s $\varepsilon_3 = \infty$, ukončíme výpočet s $\alpha_i = \alpha$. Není-li splněna podmínka (S2a), přejdeme na krok 4.

Krok 3 Určíme hodnotu α pomocí extrapolace (poznámka 45) tak, aby $\gamma_1\overline{\alpha} \le \alpha \le \gamma_2\overline{\alpha}$ a přejdeme na krok 2.

Krok 4 Určíme hodnotu α pomocí interpolace (poznámka 45) tak, aby $\beta_1(\overline{\alpha}-\underline{\alpha}) \le (\alpha-\underline{\alpha}) \le \beta_2(\overline{\alpha}-\underline{\alpha})$.

Krok 5 Jsou-li splněny podmínky (S2a) a (S3a) s $\varepsilon_3 = \infty$, ukončíme výpočet s $\alpha_i = \alpha$. Není-li splněna podmínka (S2a), položíme $\overline{\alpha} := \alpha$, v opačném případě položíme $\underline{\alpha} := \alpha$. Přejdeme na krok 4.

Poznámka 44. Algoritmus 1 je vnitřním cyklem iteračních metod spádových směrů, takže veličiny generované tímto algoritmem by měly mít dva indexy (vnější a vnitřní). Abychom zjednodušili symboliku, budeme vnější index vynechávat. Budeme tedy psát $\underline{\alpha}_j \le \alpha_j \le \overline{\alpha}_j$, $j \in N$, kde α_1 je počáteční délka kroku. Použijeme též označení $\varphi(\alpha) = F(x+\alpha s)$ a $\varphi'(\alpha) = s^T g(x+\alpha s)$.

Věta 23. *Jsou-li splněny předpoklady F1 a F3 najde algoritmus 1 délku kroku vyhovující podmínkám (S2a) a (S3a) s $\varepsilon_3 = \infty$ po konečném počtu kroků.*

Důkaz (a) V první fázi algoritmu platí $\varphi(\underline{\alpha}_j) - \varphi(0) \le \varepsilon_1\underline{\alpha}_j\varphi'(0)$, takže z předpokladu F1 (podobně jako v důkazu lemmatu 4) plyne

$$\underline{\alpha}_j \le \frac{\underline{F} - \varphi(0)}{\varepsilon_1\varphi'(0)}. \tag{72}$$

Jelikož pro $j > 1$ platí $\underline{\alpha}_j \ge \gamma_1^{j-2}\alpha_1$ a $\gamma_1 > 1$, dostaneme po konečném počtu extrapolací číslo které je větší než uvedená mez. První fáze algoritmu tedy obsahuje konečný počet kroků.

(b) Ve druhé fázi algoritmu platí $\varphi(\underline{\alpha}_j) - \varphi(0) \le \varepsilon_1\underline{\alpha}_j\varphi'(0)$ a $\varphi'(\underline{\alpha}_j) < \varepsilon_2\varphi'(0)$. Označme $\tilde{\alpha}_j > \underline{\alpha}_j$ největší číslo takové, že

$$\varphi(\alpha) - \varphi(0) \le \varepsilon_1\alpha\varphi'(0) \qquad \forall\, \underline{\alpha}_j \le \alpha \le \tilde{\alpha}_j.$$

Pak podobně jako v důkazu lemmatu 4 platí $\varphi(\tilde{\alpha}_j) - \varphi(0) = \varepsilon_1\tilde{\alpha}_j\varphi'(0)$ a $\varphi'(\tilde{\alpha}_j) \ge \varepsilon_1\varphi'(0)$. Použijeme-li tyto nerovnosti a předpoklad F3, dostaneme

$$\varepsilon_1\varphi'(0) \le \varphi'(\tilde{\alpha}_j) \le \varphi'(\underline{\alpha}_j) + (\tilde{\alpha}_j - \underline{\alpha}_j)\overline{G}\|s\|^2 < \varepsilon_2\varphi'(0) + (\tilde{\alpha}_j - \underline{\alpha}_j)\overline{G}\|s\|^2,$$

neboli

$$\tilde{\alpha}_j - \underline{\alpha}_j > \frac{\varepsilon_1 - \varepsilon_2}{\overline{G}\|s\|^2}\varphi'(0).$$

Jelikož $\varphi(\overline{\alpha}_j) - \varphi(0) > \varepsilon_1\underline{\alpha}_j\varphi'(0)$, musí platit $\overline{\alpha}_j > \tilde{\alpha}_j$, neboli

$$\overline{\alpha}_j - \underline{\alpha}_j > \frac{\varepsilon_1 - \varepsilon_2}{\overline{G}\|s\|^2}\varphi'(0). \tag{73}$$

Ve druhé fázi algoritmu, upravujeme interval tak, že $\overline{\alpha}_{j+1} - \underline{\alpha}_{j+1} \le \max(1-\beta_1, \beta_2)(\overline{\alpha}_j - \underline{\alpha}_j)$. Jelikož $\max(1-\beta_1, \beta_2) < 1$, dostaneme po konečném počtu kroků interval menší než $(\varepsilon_1 - \varepsilon_2)/(\overline{G}\|s\|^2)\varphi'(0)$. Druhá fáze algoritmu tedy obsahuje konečný počet kroků. □

Je-li splněna podmínka (S1b) (stejnoměrná spádovost) a vyhovuje-li funkce F předpokladům F4 a F5, můžeme předchozí tvrzení podstatně zesílit (budeme to potřebovat pro důkaz asymptotické přesnosti výběru délky kroku).

Lemma 10. *Uvažujme algoritmus 1 s počáteční délkou kroku $\delta_1\|g\|/\|s\| \le \alpha_1 \le \delta_2\|g\|/\|s\|$, kde konstanty δ_1 a δ_2 nezávisí na vnějším indexu. Nechť je splněna podmínka (S1b) a nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ vyhovuje předpokladům F4 a F5. Pak existují konstanty c_1 a c_2 nezávislé na vnějším indexu takové, že*

$$c_1\|g\|/\|s\| \le \overline{\alpha}_j - \underline{\alpha}_j \le \overline{\alpha}_j \le c_2\|g\|/\|s\|$$

$\forall j \in N$. V tomto případě existuje číslo $k \in N$ nezávislé na vnějším indexu takové, že počet kroků algoritmu 1 nepřekročí k.

Důkaz V prvním kroku algoritmu platí $\underline{\alpha}_j = 0$ a $\overline{\alpha}_j = \alpha_1$, takže lze položit $c_1 = \delta_1$ a $c_2 = \delta_2$. V dalších krocích algoritmu (pro $j > 1$) použijeme nerovnosti uvedené v důkazu věty 23.
(a) V první fázi algoritmu využijeme toho, že podle předpokladu F5 můžeme $\underline{F}$ nahradit F^* v (72), což s použitím nerovnosti (49) a nerovnosti (S1b) (zapsané ve tvaru $-\varphi'(0) \ge \varepsilon_0\|s\|\|g\|$) dává

$$\overline{\alpha}_j - \underline{\alpha}_j \le \overline{\alpha}_j \le \gamma_2\underline{\alpha}_j \le \gamma_2\frac{F^* - F}{\varepsilon_1\varphi'(0)} \le -\frac{\gamma_2\|g\|^2}{\varepsilon_1\underline{G}\varphi'(0)} \le \frac{\gamma_2}{\varepsilon_0\varepsilon_1\underline{G}}\frac{\|g\|}{\|s\|}.$$

S druhé strany víme že $\underline{\alpha}_j \ge \gamma_1^{j-2}\alpha_1$ a $\overline{\alpha}_j - \underline{\alpha}_j \ge (\gamma_1 - 1)\underline{\alpha}_j$. Platí tedy

$$\overline{\alpha}_j \ge \overline{\alpha}_j - \underline{\alpha}_j \ge (\gamma_1 - 1)\gamma_1^{j-2}\alpha_1 \ge (\gamma_1 - 1)\delta_1\|g\|/\|s\|.$$

Můžeme tedy položit $c_1 = \delta_1 \min(1, \gamma_1 - 1)$ a $c_2 = \gamma_2/(\varepsilon_0\varepsilon_1\underline{G})$. Jelikož $\overline{\alpha}_j \ge \gamma_1^{j-1}\alpha_1 \ge \gamma_1^{j-1}\delta_1\|g\|/\|s\|$ a $\gamma_1 > 1$, existuje index $k_1 \in N$ takový, že $\overline{\alpha}_j \ge c_2\|g\|/\|s\|$ $\forall j \ge k_1$, takže první fáze skončí po nejvýše k_1 krocích.
(b) Ve druhé fázi algoritmu se již $\overline{\alpha}_j$ nezvětšuje, takže s použitím (73) a (a) můžeme psát

$$\varepsilon_0\frac{\varepsilon_2 - \varepsilon_1}{\overline{G}}\frac{\|g\|}{\|s\|} \le \frac{\varepsilon_1 - \varepsilon_2}{\overline{G}\|s\|^2}\varphi'(0) \le \overline{\alpha}_j - \underline{\alpha}_j \le \overline{\alpha}_j \le \max\left(\frac{\gamma_2}{\varepsilon_0\varepsilon_1\underline{G}}, \delta_2\right)\frac{\|g\|}{\|s\|}.$$

Můžeme tedy položit $c_1 = \varepsilon_0(\varepsilon_2 - \varepsilon_1)/\overline{G}$ a $c_2 = \max(\gamma_2/(\varepsilon_0\varepsilon_1\underline{G}), \delta_2)$. Označme j_2 index kroku, ve kterém začíná druhá fáze algoritmu. Jelikož pro $j \ge j_2$ platí

$$\overline{\alpha}_j - \underline{\alpha}_j \le \max(1 - \beta_1, \beta_2)^{j-j_2}(\overline{\alpha}_{j_2} - \underline{\alpha}_{j_2}) \le \max(1 - \beta_1, \beta_2)^{j-j_2}c_2\|g\|/\|s\|$$

a $\max(1 - \beta_1, \beta_2) < 1$, existuje číslo $k_2 \in N$ takové, že $\overline{\alpha}_j - \underline{\alpha}_j \le c_1\|g\|/\|s\|$ $\forall j \ge j_2 + k_2$, takže druhá fáze skončí po nejvýše k_2 krocích.
(c) Podle (a) a (b) existuje číslo $k \le k_1 + k_2$ nezávislé na vnějším indexu takové, že počet kroků algoritmu 1 nepřekročí k. □

Poznámka 45. Hodnotu α použitou v algoritmu 1 můžeme určit pomocí kvadratické nebo kubické extrapolace či interpolace. Označme

$$A = \frac{\varphi(\overline{\alpha}) - \varphi(\underline{\alpha})}{(\overline{\alpha} - \underline{\alpha})\varphi'(\underline{\alpha})},$$

$$B = \frac{\varphi'(\overline{\alpha})}{\varphi'(\underline{\alpha})}.$$

Kvadratická interpolace (dvě hodnoty):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{2(1 - A)}. \tag{74}$$

Kvadratická interpolace (dvě derivace):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{1 - B}. \tag{75}$$

Kubická interpolace:

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{D + \sqrt{D^2 - 3C}}, \tag{76}$$

kde

$$C = (B-1) - 2(A-1),$$
$$D = (B-1) - 3(A-1).$$

Věta 24. *Nechť jsou splněny předpoklady lemmatu 10. Pak je-li délka kroku v algoritmu 1 spočtena podle (74) nebo (75) nebo (76), je výběr délky kroku asymptoticky přesný.*

Důkaz V důkazu budeme používat symboly $o(\xi_i)$, $O(\xi_i)$, relaci $u_i \sim v_i$ a pravidla uvedená v poznámce 10. Jelikož podle předpokladů F4 a F5 platí $s_i^T G^* s_i \sim \|s_i\|^2$ a z (S1b) plyne $s_i^T g_i \sim \|s_i\|\|g_i\|$, dostaneme $\alpha_i^* \sim \|g_i\|/\|s_i\|$, kde

$$\alpha_i^* = -\frac{s_i^T g_i}{s_i^T G^* s_i}.$$

Podle lemmatu 10 platí $\overline{\alpha}_i \sim \|g_i\|/\|s_i\|$ a $\overline{\alpha}_i - \underline{\alpha}_i \sim \|g_i\|/\|s_i\|$, takže také $\alpha_i^* - \underline{\alpha}_i = O(\|g_i\|/\|s_i\|)$ a $(\alpha_i^* - \underline{\alpha}_i)/(\overline{\alpha}_i - \underline{\alpha}_i) = O(1)$. Označme $\underline{d}_i = \underline{\alpha}_i s_i$, $\overline{d}_i = \overline{\alpha}_i s_i$ a $\underline{e}_{i+1} = x_i + \underline{d}_i - x^*$, $\overline{e}_{i+1} = x_i + \overline{d}_i - x^*$. Pak použitím nerovnosti $\underline{\alpha}_i < \overline{\alpha}_i = O(\|g_i\|/\|s_i\|)$, vztahu $\|g_i\| \sim \|e_i\|$, předpokladů (F4, F5 a věty 5) dostaneme $\underline{d}_i = O(\|g_i\|) = O(\|e_i\|)$, $\overline{d}_i = O(\|g_i\|) = O(\|e_i\|)$ a $\underline{e}_{i+1} = O(\|e_i\|)$, $\overline{e}_{i+1} = O(\|e_i\|)$. Použijeme-li větu 5 dostaneme úpravou výrazů A, B uvedených v poznámce 45

$$\begin{aligned} A &= \frac{F(x_i + \overline{\alpha}_i s_i) - F(x_i + \underline{\alpha}_i s_i)}{(\overline{\alpha}_i - \underline{\alpha}_i) s_i^T g(x_i + \underline{\alpha}_i s_i)} = \frac{(\overline{\alpha}_i - \underline{\alpha}_i) s_i^T g_i + \frac{1}{2}(\overline{\alpha}_i^2 - \underline{\alpha}_i^2) s_i^T G^* s_i + o(\|e_i\|^2)}{(\overline{\alpha}_i - \underline{\alpha}_i)(s_i^T g_i + \underline{\alpha}_i s_i^T G^* s_i + \|s_i\| o(\|e_i\|)} \\ &= \frac{s_i^T g_i + \frac{1}{2}(\overline{\alpha}_i + \underline{\alpha}_i) s_i^T G^* s_i + \|s_i\| o(\|e_i\|)}{s_i^T g_i + \underline{\alpha}_i s_i^T G^* s_i + \|s_i\| o(\|e_i\|)} = \frac{1 - (\overline{\alpha}_i + \underline{\alpha}_i)/(2\alpha_i^*) + o(1)}{1 - \underline{\alpha}_i/\alpha_i^* + o(1)}, \\ B &= \frac{s_i^T g(x_i + \overline{\alpha}_i s_i)}{s_i^T g(x_i + \underline{\alpha}_i s_i)} = \frac{s_i^T g_i + \overline{\alpha}_i s_i^T G^* s_i + \|s_i\| o(\|e_i\|)}{s_i^T g_i + \underline{\alpha}_i s_i^T G^* s_i + \|s_i\| o(\|e_i\|)} = \frac{1 - \overline{\alpha}_i/\alpha_i^* + o(1)}{1 - \underline{\alpha}_i/\alpha_i^* + o(1)}, \end{aligned}$$

takže

$$\begin{aligned} 1 - A &= 1 - \frac{1 - (\overline{\alpha}_i + \underline{\alpha}_i)/(2\alpha_i^*)}{1 - \underline{\alpha}_i/\alpha_i^*} + o(1) = \frac{1}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} + o(1), \\ 1 - B &= 1 - \frac{1 - \overline{\alpha}_i/\alpha_i^*}{1 - \underline{\alpha}_i/\alpha_i^*} + o(1) = \frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} + o(1), \end{aligned}$$

(předpokládáme, že $\alpha_i^* \neq \underline{\alpha}_i$, neboť pro $\underline{\alpha}_i$ neplatí (S3a), zatímco $s_i^T g(x_i + \alpha_i^* s_i)/s_i^T g_i \to 0$). Nyní se omezíme na vzorec (76) (důkaz pro (74) a (75) je mnohem jednodušší a přenecháme ho čtenáři). Použijeme-li právě získané vztahy, dostaneme

$$\begin{aligned} C &= 2(1-A) - (1-B) = \frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} - \frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} + o(1) = o(1) \\ D &= 3(1-A) - (1-B) = \frac{3}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} - \frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i} + o(1) = \frac{1}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}(1 + o(1)), \end{aligned}$$

neboť $(\alpha_i^* - \underline{\alpha}_i)/(\overline{\alpha}_i - \underline{\alpha}_i) = O(1)$. Platí tedy

$$\begin{aligned} D + \sqrt{D^2 - 3C} &= \frac{1}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}(1 + o(1)) + \sqrt{\frac{1}{4}\left(\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}\right)^2 (1 + o(1))^2 + o(1)} \\ &= \frac{1}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}(1 + o(1)) + \frac{1}{2}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}\sqrt{(1 + o(1))^2 + \left(\frac{\alpha_i^* - \underline{\alpha}_i}{\overline{\alpha}_i - \underline{\alpha}_i}\right)^2 o(1)} \\ &= \frac{\overline{\alpha}_i - \underline{\alpha}_i}{\alpha_i^* - \underline{\alpha}_i}(1 + o(1)), \end{aligned}$$

neboť $(\alpha_i^* - \underline{\alpha}_i)/(\overline{\alpha}_i - \underline{\alpha}_i) = O(1)$. Dosadíme-li tento výraz do (76), dostaneme

$$\begin{aligned} \alpha_i &= \underline{\alpha}_i + \frac{\overline{\alpha}_i - \underline{\alpha}_i}{D + \sqrt{D^2 - 3C}} = \underline{\alpha}_i + \frac{\alpha_i^* - \underline{\alpha}_i}{\overline{\alpha}_i - \underline{\alpha}_i}\frac{\overline{\alpha}_i - \underline{\alpha}_i}{1 + o(1)} \\ &= \underline{\alpha}_i + (\alpha_i^* - \underline{\alpha}_i)(1 + o(1)) = \alpha_i^*(1 + o(1)), \end{aligned}$$

neboť $\underline{\alpha}_i/\alpha_i^* = O(1)$. □

Poznámka 46. Je-li kromě předpokladů F4 a F5 splněn i předpoklad F6, můžeme místo věty 5 použít větu 6 a tudíž místo $o(1)$ psát $O(\|e_i\|)$. Dostaneme tak kvalitnější odhady

$$\frac{s_i^T g_{i+1}}{s_i^T g_i} = O(\|e_i\|)$$

a

$$\alpha_i = -\frac{s_i^T g_i}{s_i^T G^* s_i}(1 + O(\|e_i\|)).$$

Poznámka 47. Počáteční výběr délky kroku. Pokud $s_i \sim g_i$, což je případ většiny efektivních metod, je výhodné volit $\alpha_0 \sim 1$. Pro superlineárně konvergentní metody volíme $\alpha_0 = 1$. U metod sdružených gradientů volíme $\alpha_0 = \min(1, 2(F_i - F_{i-1})/s_i^T g_i, 2(\underline{F} - F_i)/s_i^T g_i)$ (v prvním iteračním kroku pokládáme $\alpha_0 = \min(1, 2(\underline{F} - F_i)/s_i^T g_i)$.

2.5 Nemonotonní metody spádových směrů

Zatím jsme se zabývali pouze metodami, kde posloupnost F_i, $i \in N$, byla nerostoucí. Někdy je výhodné (zejména ve spojení s Newtonovou metodou) používat nemonotonní metody spádových směrů, kdy posloupnost F_i, $i \in N$, není nerostoucí. V definici nemonotonních metod spádových směrů se místo hodnot F_i, $i \in N$, používají čísla $\overline{F}_i \geq F_i$, $i \in N$, jejichž výběr je určen konkrétní metodou. Poznamenejme, že pro tato čísla platí $\overline{F}_i \leq \overline{F}$, $\forall i \in N$ (kde $\overline{F} = F_1$), takže opět $x_i \in \mathcal{D}_F(\overline{F}) \subset \mathcal{D}$ $\forall i \in N$

Definice 20. *Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje nemonotonní Armijovu podmínku, jestliže existuje číslo $0 < \varepsilon_1 < 1$ (nezávislé na indexu $i \in N$) takové, že*

$$F_{i+1} - \overline{F}_i \leq \varepsilon_1 \alpha_i s_i^T g_i. \tag{S2b}$$

*Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje nemonotonní Wolfeho podmínku, jestliže existují čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ a $\varepsilon_3 \geq 0$ (nezávislá na indexu $i \in N$) taková, že platí (*S2b*) a*

$$\varepsilon_2 s_i^T g_i \leq s_i^T g_{i+1} \leq \varepsilon_3 |s_i^T g_i|. \tag{S3a}$$

*Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje nemonotonní Goldsteinovu podmínku, jestliže existují čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) taková, že platí (*S2b*) a*

$$F_{i+1} - F_i \geq \varepsilon_2 \alpha_i s_i^T g_i. \tag{S3b}$$

Poznámka 48. Nemonotonní Armijova podmínka (S2b) je součástí zbylých dvou nemonotonní podmínek. Samotatně ji lze použít v nemonotonním Armijově výběru délky kroku. V tomto případě je $\alpha_i > 0$ prvním členem vyhovující podmínce (S2b) v posloupnosti α_i^j, $j \in N$, takové, že $\underline{\alpha}\|g_i\|/\|s_i\| \leq \alpha_i^1 \leq \overline{\alpha}\|g_i\|/\|s_i\|$, a

$$\underline{\beta}\alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta}\alpha_i^j \quad \forall j \in N,$$

kde $0 < \underline{\alpha} \leq \overline{\alpha}$ a $0 < \underline{\beta} \leq \overline{\beta} < 1$ jsou konstanty nezávislé na indexu $i \in N$.

Definice 21. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je nemonotonní metodou spádových směrů, jestliže směrové vektory $s_i \in R^n$, $i \in N$, splňují podmínku (S1a) a délky kroku $\alpha_i > 0$, $i \in N$, splňují podmínku (S2b) a některou z podmínek (S3a), (S3b). Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je nemonotonní metodou stejnoměrně spádových směrů, je-li nemonotonní metodou spádových směrů a platí-li (S1b).*

Poznámka 49. *Pro každou nemonotonní metodou spádových směrů platí rovnice (31), neboť podmínky (S3a), (S3b) jsou stejné jako v případě monotonních metod spádových směrů. Rovnici (32) je třeba nahradit vztahem*

$$F_{i+1} - \overline{F}_i \leq -\frac{\varepsilon_1\varepsilon_4(s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_4}{\overline{G}}\cos^2\theta_i\|g_i\|^2. \tag{77}$$

Nejprve vyšetříme jednoduchou nemonotonní metodu spádových směrů, pro kterou platí

$$\overline{F}_i = \max\{F_j : i - \min(m,i) + 1 \leq j \leq i\}, \tag{78}$$

kde číslo m udává počet funkčních hodnot použitých k určení $\overline{F}_i$.

Věta 25. *(Globální konvergence metody (78)) Nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak nemonotonní metoda stejnoměrně spádových směrů definovaná vztahem (78) je globálně konvergentní.*

Důkaz (a) Z (S2b) a (78) vyplývá, že posloupnost $\overline{F}_i$, $i \in N$, je nerostoucí. Platí totiž

$$\overline{F}_{i+1} = \max\{F_{j+1} : i - \min(m,i) + 1 \leq j \leq i\} \leq \max(\overline{F}_i, F_{i+1}) \leq \max(\overline{F}_i, \overline{F}_i - \varepsilon_1\alpha_i s_i^T g_i) = \overline{F}_i.$$

Použijeme-li navíc (S1b) a (77) vidíme, že pro libovolné indexy $k \in N$ a $1 \leq j \leq m$ platí

$$F_{km+j} \leq \overline{F}_{km+j-1} - \frac{\varepsilon_1\varepsilon_4(s_{km+j-1}^T g_{km+j-1})^2}{\overline{G}\|s_{km+j-1}\|^2} \leq \overline{F}_{km} - \frac{\varepsilon_0\varepsilon_1\varepsilon_4}{\overline{G}}\|g_{km+j-1}\|^2$$

a s použitím (78) dostaneme

$$\overline{F}_{km+m} = \max_{1\leq j\leq m} F_{km+j} \leq \overline{F}_{km} - \frac{\varepsilon_0\varepsilon_1\varepsilon_4}{\overline{G}} \min_{1\leq j\leq m}\|g_{km+j-1}\|^2,$$

čili

$$\overline{F}_{(k+1)m} - \overline{F}_{km} \leq -\frac{\varepsilon_0\varepsilon_1\varepsilon_4}{\overline{G}}\|g_{j(k)}\|^2.$$

kde $\|g_{j(k)}\| = \min_{1\leq j\leq m}\|g_{km+j-1}\|$.

(b) Podle (a) platí

$$\frac{\varepsilon_0\varepsilon_1\varepsilon_4}{\overline{G}}\sum_{k=1}^{\infty}\|g_{j(k)}\|^2 \leq \sum_{k=1}^{\infty}(\overline{F}_{km} - \overline{F}_{(k+1)m}) = \overline{F}_m - \lim_{k\to\infty}\overline{F}_{(k+1)m} \leq \overline{F}_m - \underline{F} < \infty,$$

takže $\lim_{k\to\infty}\|g_{j(k)}\| = 0$ a tedy $\liminf_{i\to\infty}\|g_i\| = 0$. □

Nyní vyšetříme nemonotonní metodu spádových směrů, kde se čísla $\overline{F}_i$, $i \in N$, určují rekurentně tak, že $\overline{n}_1 = 1$, $\overline{F}_1 = F_1$ a

$$\overline{n}_{i+1} = \lambda \overline{n}_i + 1, \qquad \overline{F}_{i+1} = \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \tag{79}$$

pro $i \in N$, kde $0 \leq \lambda \leq 1$.

Poznámka 50. Pokud $\lambda = 0$, platí $\overline{n}_i = 1$ a $\overline{F}_i = F_i$ pro $i \in N$. Pokud $\lambda = 1$, platí $\overline{n}_i = i$ a

$$\overline{F}_i = \frac{1}{i} \sum_{j=1}^{i} F_j$$

pro $i \in N$. V obecném případě platí $1 \leq \overline{n}_i \leq i$ a

$$F_{i+1} \leq \overline{F}_{i+1} \leq \overline{F}_i \leq \frac{1}{i} \sum_{j=1}^{i} F_j$$

pro $i \in N$, neboť z (S2b) plyne $F_{i+1} \leq \overline{F}_i$, což spolu s (79) dává

$$\begin{aligned} F_{i+1} &= \frac{\lambda \overline{n}_i + 1}{\overline{n}_{i+1}} F_{i+1} \leq \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} = \overline{F}_{i+1}, \\ \overline{F}_{i+1} &= \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \leq \frac{\lambda \overline{n}_i + 1}{\overline{n}_{i+1}} \overline{F}_i = \overline{F}_i. \end{aligned}$$

Jelikož funkce

$$\begin{aligned} \overline{n}_{i+1}(\lambda) &= \lambda \overline{n}_i + 1, \\ \overline{F}_{i+1}(\lambda) &= \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\lambda \overline{n}_i + 1} \end{aligned}$$

(kde $F_{i+1} \leq \overline{F}_i$) jsou pro $i \in N$ neklesající, dostaneme $F_{i+1} \leq \overline{F}_{i+1}(1) = (1/i) \sum_{j=1}^{i} F_j$.

Věta 26. *(Globální konvergence metody (79)) Nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak nemonotonní metoda spádových směrů definovaná rekurentními vztahy (79) je globálně konvergentní, pokud*

$$\sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{i} = \infty.$$

Důkaz Podle (77) platí

$$F_{i+1} \leq \overline{F}_i - \frac{\varepsilon_1 \varepsilon_4 \cos^2 \theta_i}{\overline{G}} \|g_i\|^2,$$

což spolu s (79) dává

$$\overline{F}_{i+1} = \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \leq \frac{\lambda \overline{n}_i + 1}{\overline{n}_{i+1}} \overline{F}_i - \frac{\varepsilon_1 \varepsilon_4 \cos^2 \theta_i}{\overline{n}_{i+1} \overline{G}} \|g_i\|^2 = \overline{F}_i - \frac{\varepsilon_1 \varepsilon_4 \cos^2 \theta_i}{\overline{n}_{i+1} \overline{G}} \|g_i\|^2.$$

Jelikož podle předpokladu F1 a poznámky 50 platí $\overline{F}_{i+1} \geq F_{i+1} \geq \underline{F}$, můžeme psát

$$\frac{\varepsilon_1 \varepsilon_4}{\overline{G}} \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{\overline{n}_{i+1}} \|g_i\|^2 \leq \sum_{i=1}^{\infty} (\overline{F}_i - \overline{F}_{i+1}) \leq \overline{F}_1 - \underline{F}.$$

Dostaneme tedy

$$\frac{1}{2} \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{i} \|g_i\|^2 \leq \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{\overline{n}_{i+1}} \|g_i\|^2 \leq \frac{(\overline{F}_1 - \underline{F}) \overline{G}}{\varepsilon_1 \varepsilon_4} < \infty,$$

neboť podle poznámky 50 platí $\overline{n}_{i+1} \leq i + 1 \leq 2i$. Z poslední nerovnosti dostaneme dokazované tvrzení postupem uvedeným v důkazu věty 11. □

Poznámka 51. Podmínka použitá ve větě 26 je mnohem silnější než podmínka vystupující ve větě 11. Je však splněna pro nemonotonní metody stejnoměrně spádových směrů, kdy $\cos\theta_i \geq \varepsilon_0\ \forall i \in N$. Jestliže kromě $\cos\theta_i \geq \varepsilon_0\ \forall i \in N$ platí též $0 \leq \lambda < 1$, dá se dokázat, že $\lim_{i\to\infty} \|g_i\| = 0$.

Poznámka 52. Realizace nemonotonních metod spádových směrů se příliš neliší od realizace standardních metod spádových směrů. Stačí počítat hodnoty $\overline{F}_i$, $i \in N$, a v algoritmu 1 nahradit podmínku (S2a) podmínkou (S2b).

Poznámka 53. jsou-li splněny podmínky pro superlineární konvergenci (61) a pokládáme-li $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2b) a (S3), jsou nemonotonní metody spádových směrů superlineárně konvergentní. Plyne to z části (c) důkazu věty 20 a z toho, že slabší podmínky (S2b) a (S3) jsou splněny pokud platí (S2a) a (S3). Proto se nemonotonní metody spádových směrů používají zejména ve spojení s Newtonovou metodou.

2.6 Využití směrů se zápornou křivostí

Newtonovu metodu nemůžeme bez dalších úprav realizovat jako metodu spádových směrů, neboť matice G_i nemusí být pozitivně definitní, takže směrový vektor $s_i = -G_i^{-1}g_i$ nemusí být spádový. Metody spádových směrů hledají nižší hodnotu minimalizované funkce na polopřímce $x_i(\alpha) = x_i + \alpha s_i$. Nahradíme-li tuto polopřímku vhodnou křivkou $x_i(\alpha)$ začínající v bodě x_i, můžeme dosáhnout toho, že $F(x_i(\alpha_i)) < F(x_i)$, pro nějakou hodnotu $\alpha_i > 0$. Ke konstrukci této křivky se používají směry se zápornou křivostí.

Definice 22. *Dvojici vektorů $(s_i, z_i) \in R^n \times R^n$ nazveme spádovým párem pro funkci $F \in \mathcal{C}^2 : \mathcal{D} \to R$ v bodě x_i, jestliže*

$$\begin{aligned} s_i = 0, &\quad \text{pokud} \quad g_i = 0, \\ g_i^T s_i < 0, &\quad \text{pokud} \quad g_i \neq 0 \quad \text{a} \quad G_i \succeq 0, \\ g_i^T s_i \leq 0, &\quad \text{pokud} \quad g_i \neq 0 \quad \text{a} \quad G_i \not\succeq 0, \\ z_i = 0, &\quad \text{pokud} \quad G_i \succeq 0, \\ g_i^T z_i \leq 0 \quad \text{a} \quad z_i^T G_i z_i < 0, &\quad \text{pokud} \quad G_i \not\succeq 0 \end{aligned}$$

(spádový pár je nulový, je-li bod x_i lokálním minimem funkce F). Vektor s_i se nazývá spádovým směrem a vektor z_i je směrem se zápornou křivostí pro funkci F v bodě x_i.

Definice 23. *Spádový pár $(s_i, z_i) \in R^n \times R^n$ nazveme přijatelným spádovým párem, jestliže existují čísla $0 < \varepsilon_0 \leq 1$, $0 < \underline{s} \leq \overline{s}$ taková, že pro $i \in N$ platí*

$$-g_i^T s_i \geq \varepsilon_0 \|g_i\| \|s_i\|, \qquad \underline{s}\|g_i\| \leq \|s_i\| \leq \overline{s}\|g_i\|, \tag{80}$$

a čísla $0 < \delta_0 \leq 1$, $0 < \underline{z} \leq \overline{z}$ taková, že pro $G_i \not\succeq 0$ platí

$$z_i^T G_i z_i \leq \delta_0 \underline{\lambda}(G_i) \|z_i\|^2, \quad \underline{z} \leq \|z_i\| \leq \overline{z}. \tag{81}$$

Poznámka 54. Místo podmínky (80) se často používá podmínka

$$-g_i^T s_i \geq \underline{s}\|g_i\|^2, \qquad \|s_i\| \leq \overline{s}\|g_i\|. \tag{82}$$

Tato podmínka implikuje (80), neboť je-li splněna platí

$$-g_i^T s_i \geq \underline{s}\|g_i\|^2 \geq \frac{\underline{s}}{\overline{s}}\|g_i\|\|s_i\| \stackrel{\Delta}{=} \varepsilon_0 \|g_i\|\|s_i\|$$

a protože $\|g_i\|\|s_i\| \geq -g_i^T s_i \geq \underline{s}\|g_i\|^2$, můžeme psát $\|s_i\| \geq \underline{s}\|g_i\|$.

Poznámka 55. Z (80) plyne, že $g_i = 0$ a $s_i = 0$, pokud $g_i^T s_i = 0$ a také $g_i \to 0$ a $s_i \to 0$, pokud $g_i^T s_i \to 0$. Je li splněn předpoklad F2, jsou normy vektorů g_i a s_i shora omezené. Z (81) plyne, že $\min(0, \underline{\lambda}(G_i)) = 0$, pokud $z_i^T G_i z_i = 0$, a také $\min(0, \underline{\lambda}(G_i)) \to 0$, pokud $z_i^T G_i z_i \to 0$, přičemž normy vektorů z_i jsou shora omezené.

Máme li spádový pár (s_i, z_i), můžeme definovat spádovou křivku

$$x_i(\alpha) = x_i + d_i(\alpha), \qquad d_i(\alpha) = \varphi_1(\alpha)s_i + \varphi_2(\alpha)z_i, \tag{83}$$

kde $\varphi_1(\alpha)$, $\varphi_2(\alpha)$ jsou neklesající funkce definované pro $\alpha \geq 0$ takové, že $\varphi_1(0) = \varphi_2(0) = 0$. Délka kroku $\alpha_i > 0$ se volí tak, aby byla splněna zobecněná Goldsteinova podmínka

$$\varepsilon_2 \tilde{Q}_i(\alpha_i) \leq \tilde{F}_i(\alpha_i) - \tilde{F}_i(0) \leq \varepsilon_1 \tilde{Q}_i(\alpha_i), \tag{84}$$

kde $\tilde{F}_i(\alpha) = F(x_i(\alpha)) = F(x_i + \varphi_1(\alpha)s_i + \varphi_2(\alpha)z_i)$ a

$$\tilde{Q}_i(\alpha) = \tilde{F}_i'(0)\alpha + \frac{1}{2}\min(0, \tilde{F}_i''(0))\alpha^2. \tag{85}$$

Lemma 11. *Nechť dvojice vektorů (s_i, z_i) je spádovým párem (v bodě x_i) pro funkci $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňující předpoklad F1 a nechť $0 < \varepsilon_1 < \varepsilon_2 < 1$. Pak, platí-li buď $\tilde{F}_i'(0) < 0$ nebo $\tilde{F}_i'(0) = 0$ a $\tilde{F}_i''(0) < 0$, existuje číslo $\alpha_i > 0$ splňující zobecněnou Goldsteinovu podmínku (84).*

Důkaz Pokud $\tilde{F}_i'(0) < 0$, platí $\tilde{Q}_i(\alpha) = \tilde{F}_i'(0)\alpha + o(\alpha)$ a použijeme-li jeden člen Taylorova rozvoje, dostaneme $\tilde{F}_i(\alpha) - \tilde{F}_i(0) = \tilde{F}_i'(0)\alpha + o(\alpha)$, takže

$$\frac{\tilde{F}_i(\alpha) - \tilde{F}_i(0)}{\tilde{Q}_i(\alpha)} = \frac{\tilde{F}_i'(0)\alpha(1 + o(1))}{\tilde{F}_i'(0)\alpha(1 + o(1))} = 1 + o(1).$$

Pokud $\tilde{F}_i'(0) = 0$ a $\tilde{F}_i''(0) < 0$, platí $\tilde{Q}_i(\alpha) = (1/2)\tilde{F}_i''(0)\alpha^2$ a použijeme-li dva členy Taylorova rozvoje, dostaneme $\tilde{F}_i(\alpha) - \tilde{F}_i(0) = (1/2)\tilde{F}_i''(0)\alpha^2 + o(\alpha^2)$, takže

$$\frac{\tilde{F}_i(\alpha) - \tilde{F}_i(0)}{\tilde{Q}_i(\alpha)} = \frac{\tilde{F}_i''(0)\alpha^2(1 + o(1))}{\tilde{F}_i''(0)\alpha^2} = 1 + o(1).$$

V obou případech tedy platí $(\tilde{F}_i(\alpha) - \tilde{F}_i(0))/\tilde{Q}_i(\alpha) \to 1$, pokud $\alpha \to 0$. Existuje tedy číslo $\tilde{\alpha}_i > 0$ takové, že $\tilde{F}_i(\alpha_i) - \tilde{F}_i(0) < \varepsilon_1 \tilde{Q}_i(\alpha_i)$, pokud $\alpha < \tilde{\alpha}_i$. Jsou-li splněny podmínky lemmatu, je výraz $\tilde{Q}_i(\alpha)$ záporný pro libovolnou hodnotu $\alpha > 0$ a platí $\tilde{Q}_i(\alpha) \to -\infty$, pokud $\alpha \to \infty$. Pokud by platilo $\tilde{F}_i(\alpha) - \tilde{F}_i(0) < \varepsilon_1 \tilde{Q}_i(\alpha)$ $\forall \alpha > 0$, dostali bychom $\tilde{F}_i(\alpha) \to -\infty$ pro $\alpha \to \infty$, což je ve sporu s předpokladem F1. Jelikož funkce $\tilde{F}(\alpha)$ je spojitá, existuje délka kroku $\alpha_i \geq \tilde{\alpha}_i$ taková, že $\tilde{F}_i(\alpha_i) - \tilde{F}_i(0) = \varepsilon_1 \tilde{Q}_i(\alpha_i)$. Tato délka kroku splňuje zobecněnou Goldsteinovu podmínku (84). □

Jelikož $\tilde{F}_i(\alpha) = F(x_i(\alpha)) = F(x_i + \varphi_1(\alpha)s_i + \varphi_2(\alpha)z_i)$, platí

$$\begin{aligned} \tilde{F}_i'(0) &= g_i^T(\varphi_1'(0)s_i + \varphi_2'(0)z_i), \\ \tilde{F}_i''(0) &= g_i^T(\varphi_1''(0)s_i + \varphi_2''(0)z_i) + (\varphi_1'(0)s_i + \varphi_2'(0)z_i)^T G_i(\varphi_1'(0)s_i + \varphi_2'(0)z_i). \end{aligned}$$

Je-li x_i sedlovým bodem funkce F, platí $g_i = 0$ a $G_i \not\succeq 0$, takže $z_i^T G_i z_i < 0$ (definice 22). Podmínky lemmatu 11 jsou v tomto případě splněny, pokud $\varphi_1'(0) = 0$. Pak ale v libovolném bodě x_i platí

$$\begin{aligned} \tilde{F}_i'(0) &= \varphi_2'(0)g_i^T z_i, \\ \tilde{F}_i''(0) &= g_i^T(\varphi_1''(0)s_i + \varphi_2''(0)z_i) + (\varphi_2'(0))^2 z_i^T G_i z_i. \end{aligned}$$

Jestliže $g_i \neq 0$ a $G_i \succeq 0$, je podle definice 22 $g_i^T s_i < 0$, $g_i^T z_i \leq 0$ a $z_i^T G_i z_i = 0$, takže podmínky lemmatu 11 jsou splněny, pokud $\varphi_2'(0) > 0$ $\varphi_1''(0) > 0$ a $\varphi_2''(0) \geq 0$. Nejjednodušší křivkou tohoto typu je křivka

$$x_i(\alpha) = x_i + d_i(\alpha), \qquad d_i(\alpha) = \alpha^2 s_i + \alpha z_i, \tag{86}$$

pro kterou platí $\varphi_1'(0) = 0$, $\varphi_1''(0) = 1$, $\varphi_2'(0) = 1$, $\varphi_2''(0) = 0$. Křivku (86) lze s výhodou použít v okolí sedlového bodu. Pokud $G_i \succeq 0$, platí $z_i = 0$, takže $\tilde{F}'(0) = 0$, což znevýhodňuje použití této křivky v oblasti, kde je funkce F konvexní. V případě, že $G_i \succeq 0$, je výhodnější volit křivku

$$x_i(\alpha) = x_i + d_i(\alpha), \qquad d_i(\alpha) = \alpha s_i + \alpha^2 z_i. \tag{87}$$

Velmi praktickým způsobem použití směrů se zápornou křivostí, kdy není nutné definovat spádovou křivku, je volba mezi dvěma spádovými polopřímkami. V tomto případě pokládáme $x_i(\alpha) = x_i + d_i(\alpha)$, přičemž

$$d_i(\alpha) = \alpha s_i, \quad \text{pokud} \quad g_i^T z_i + \frac{1}{2} z_i^T G_i z_i > \underline{c} \frac{g_i^T s_i}{\|s_i\|}, \tag{88}$$

$$d_i(\alpha) = \alpha z_i, \quad \text{pokud} \quad g_i^T z_i + \frac{1}{2} z_i^T G_i z_i \leq \underline{c} \frac{g_i^T s_i}{\|s_i\|}, \tag{89}$$

kde $\underline{c} > 0$ je vhodná konstanta. Označíme-li $N_1 = \{i \in N : d_i(\alpha) = \alpha s_i\}$ a $N_2 = \{i \in N : d_i(\alpha) = \alpha z_i\}$, můžeme v dalších úvahách formálně pokládat $z_i = 0$, pokud $i \in N_1$, a $s_i = 0$, pokud $i \in N_2$.

K určení délky kroku $\alpha_i > 0$ pro spádové křivky (86), (87) a polopřímky (88), (89) se používá zobecněná Goldsteinova podmínka s funkcí

$$Q_i(d_i(\alpha)) = g_i^T d_i(\alpha) + \frac{1}{2} \min(0, d_i^T(\alpha) G_i d_i(\alpha)). \tag{90}$$

Definice 24. *Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje zobecněnou Goldsteinovu podmínku s funkcí (90), existují-li čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) taková, že*

$$\varepsilon_2 Q_i(d_i(\alpha_i)) \leq F(x_i + d_i(\alpha_i)) - F(x_i) \leq \varepsilon_1 Q_i(d_i(\alpha_i)) \tag{91}$$

Definice 25. *Optimalizační metodu, jejíž iterační krok má tvar $x_{i+1} = x_i + d_i(\alpha_i)$, kde vektor $d_i(\alpha)$ je určen podle (86) nebo (87) nebo (88)–(89), přičemž dvojice vektorů (s_i, z_i) je spádovým párem pro funkci $F \in \mathcal{C}^2 : \mathcal{D} \to R$ v bodě x_i, a kde délka kroku $\alpha_i > 0$ se vybírá tak, aby byla splněna zobecněná Goldsteinova podmínka (91), nazveme metodou spádových párů. Jsou-li spádové páry přijatelné, nazveme tuto metodu metodou přijatelných spádových párů.*

V dalším výkladu se omezíme na metodou spádových párů, kde vektor $d_i(\alpha)$ je určen podle (88)–(89). Budeme přitom používat výsledky uvedemé v práci [70]

Lemma 12. *Nechť dvojice vektorů (s_i, z_i) je spádovým párem (v bodě x_i) pro funkci $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňující předpoklad F1 a nechť vektor $d_i(\alpha)$ je určen podle (88)–(89). Pak pro libovolná čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ existuje délka kroku $\alpha_i > 0$ splňující zobecněnou Goldsteinovu podmínku (91).*

Důkaz (a) Nechť $i \in N_1$, takže $d_i(\alpha) = \alpha s_i$, kde $g_i^T s_i < 0$. Pak

$$Q_i(d_i(\alpha)) = \alpha g_i^T s_i + \frac{1}{2} \alpha^2 \min\left(0, s_i^T G_i s_i\right) = \alpha g_i^T s_i + o(\alpha)$$

a použijeme-li dva členy Taylorova rozvoje, dostaneme

$$F(x_i + d_i(\alpha)) - F(x_i) = \alpha g_i^T s_i + o(\alpha)$$

Platí tedy

$$\lim_{\alpha \to 0} \frac{F(x_i + d_i(\alpha)) - F(x_i)}{Q_i(d_i(\alpha))} = 1. \tag{92}$$

(b) Nechť $i \in N_2$, takže $d_i(\alpha) = \alpha z_i$, kde $z_i^T G_i z_i < 0$. Pak

$$Q_i(d_i(\alpha)) = \alpha g_i^T z_i + \frac{1}{2} \alpha^2 \min\left(0, z_i^T G_i z_i\right) = \alpha g^T z_i + \frac{1}{2} \alpha^2 z_i^T G_i z_i$$

a použijeme-li tři členy Taylorova rozvoje, dostaneme

$$F(x_i + d_i(\alpha)) - F(x_i) = \alpha g^T z_i + \frac{1}{2}\alpha^2 z_i^T G_i z_i + o(\alpha^2).$$

Platí tedy opět (92).

(c) Použijeme-li stejnou argumentaci jako v důkazu lemmatu 11, zjistíme, že existuje délka kroku $\alpha_i > 0$ splňující zobecněou Goldsteinovu podmínku (91).

□

Věta 27. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou přijatelných spádových párů (88)–(89). aplikovanou na funkce funkci $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňující předpoklady F1, F2, F4 a F6. Pak platí*

$$\lim_{i\to\infty} \|g(x_i)\| = 0, \qquad \lim_{i\to\infty} \min(0, \underline{\lambda}(G_i)) = 0.$$

Důkaz (a) Dokážeme nejprve, že $\|g_i\| \overset{N_1}{\to} 0$. Předpokládejme naopak, že existují číslo $\underline{\varepsilon} > 0$ a nekonečná množina $N_1' \subset N_1$ tak, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N_1'$. Použijeme-li (90) a (91), dostaneme

$$F(x_i + d_i(\alpha_i)) - F(x_i) \leq \varepsilon_1 \alpha_i g_i^T s_i$$

a podle předpokladu F1 platí, že $F(x_i + d_i(\alpha_i)) - F(x_i) \to 0$, takže nutně $\alpha_i |g_i^T s_i| \overset{N_1'}{\to} 0$. Jelikož (80) implikuje nerovnost $|g_i^T s_i| \geq \varepsilon_0 \|g_i\| \|s_i\| \geq \underline{s}\varepsilon_0\underline{\varepsilon}^2$, musí platit $\alpha_i \overset{N_1'}{\to} 0$. Použijeme-li větu o střední hodnotě pro $d_i(\alpha) = \alpha s_i$ spolu s předpokladem F4, dostaneme

$$F(x_i + d_i(\alpha_i)) - F(x_i) = \alpha_i g_i^T (s_i + \theta d_i(\alpha_i)) s_i \leq \alpha_i g_i^T s_i + \alpha_i \|s_i\| \overline{G} \|\theta \alpha_i s_i\|,$$

kde $0 \leq \theta \leq 1$, takže podle (91) platí

$$\varepsilon_2 \geq \frac{F(x_i + d_i(\alpha_i)) - F(x_i)}{Q_i(d_i(\alpha_i))} \geq 1 + \frac{\alpha_i \|s_i\| \overline{G} \|\theta \alpha_i s_i\|}{\alpha_i g_i^T s_i} \geq 1 - \frac{\alpha_i \overline{G} \overline{s}}{\varepsilon_0}$$

(neboť $Q_i(d_i(\alpha_i)) \leq \alpha_i g_i^T s_i$), což je ve sporu s tím, že $\alpha_i \overset{N_1'}{\to} 0$.

(b) Dokážeme nyní, že $\|g_i\| \overset{N_2}{\to} 0$. Předpokládejme naopak, že existují číslo $\underline{\varepsilon} > 0$ a nekonečná množina $N_2' \subset N_2$ tak, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N_2'$. Použijeme-li (90) a (91), dostaneme

$$F(x_i + d_i(\alpha_i)) - F(x_i) \leq \varepsilon_1 \left(\alpha_i g_i^T z_i + \frac{1}{2}\alpha_i^2 z_i^T G_i z_i \right)$$

a podle předpokladu F1 platí, že $F(x_i + d_i(\alpha_i)) - F(x_i) \to 0$, takže buď $\alpha_i \overset{N_2'}{\to} 0$ nebo $g_i^T z_i + (1/2) z_i^T G_i z_i \overset{N_2'}{\to} 0$ (oba dva členy $g_i^T z_i$ a $(1/2) z_i^T G_i z_i$ jsou nekladné). Podle (80) a (89) pro $i \in N_2$ platí

$$\left| g_i^T z_i + \frac{1}{2} z_i^T G_i z_i \right| \geq \underline{c} \frac{|g_i^T s_i|}{\|s_i\|} \geq \underline{c}\, \varepsilon_0 \|g_i\| \geq \underline{c}\, \varepsilon_0 \underline{\varepsilon},$$

takže nutně $\alpha_i \overset{N_2'}{\to} 0$. Použijeme-li větu o střední hodnotě pro $d_i(\alpha) = \alpha z_i$ spolu s předpokladem F6, dostaneme

$$F(x_i + d_i(\alpha)) - F(x_i) = \alpha g_i^T z_i + \frac{1}{2}\alpha^2 z_i^T G(x_i + \theta d_i(\alpha_i)) z_i \leq \alpha g_i^T z_i + \frac{1}{2}\alpha^2 z_i^T G_i z_i + \frac{1}{2}\alpha^2 \|z_i\|^2 \overline{L} \|\theta \alpha z_i\|,$$

kde $0 \leq \theta \leq 1$, takže podle (91) a (89) platí

$$\varepsilon_2 \geq \frac{F(x_i + d_i(\alpha_i)) - F(x_i)}{Q_i(d_i(\alpha_i))} \geq 1 + \frac{\frac{1}{2}\alpha_i^2 \|z_i\|^2 \overline{L} \|\theta \alpha_i z_i\|}{\alpha_i g_i^T z_i + \frac{1}{2}\alpha_i^2 z_i^T G_i z_i} \geq 1 + \frac{\alpha_i^3 \|z_i\|^3 \overline{L}}{2\alpha_i^2 (g_i^T z_i + \frac{1}{2} z_i^T G_i z_i)} \geq 1 - \frac{\alpha_i \overline{L} \overline{z}^3}{2\underline{c}\, \varepsilon_0 \underline{\varepsilon}}$$

(předpokládáme bez újmy na obecnosti, že $\alpha_i \leq 1$, takže $\alpha_i^2 \leq \alpha_i$), což je ve sporu s tím, že $\alpha_i \overset{N_2'}{\to} 0$.

(c) Nechť bod x^* je hromadným bodem posloupnosti x_i, $i \in N$ (podle předpokladu F2 takový bod existuje). Ukážeme, že $G(x^*) \succeq 0$. Nechť $x_i \overset{N'}{\to} x^*$, kde $N' \subset N$, a nechť $G(x^*) \not\succeq 0$, takže $\underline{\lambda}^* = \underline{\lambda}(G(x^*)) < 0$. Pak $\underline{\lambda}_i \overset{\Delta}{=} \underline{\lambda}(G_i) \overset{N'}{\to} \underline{\lambda}^* < 0$ a podle (a) a (b) platí $g_i \overset{N'}{\to} 0$. Existuje tedy index $k \in N'$ takový, že $\underline{\lambda}_i \leq \underline{\lambda}^*/2$ a $\underline{c}\|g_i\| \leq -\delta_0\underline{\lambda}^*/4$, pokud $i \in N'$ a $i \geq k$, takže podle (81) platí

$$\underline{c}\frac{g_i^T s_i}{\|s_i\|} \geq -\underline{c}\|g_i\| \geq \delta_0\underline{\lambda}^*/4 \geq \delta_0\underline{\lambda}_i/2 \geq (1/2)z_i^T G_i z_i \geq g_i^T z_i + (1/2)z_i^T G_i z_i,$$

neboli $i \in N_2$, pokud $i \in N'$ a $i \geq k$. Tak jako v (b) pro $i \in N' \cap N_2$ platí

$$\varepsilon_2 \geq \frac{F(x_i + d_i(\alpha_i)) - F(x_i)}{Q_i(d_i(\alpha_i))} \geq 1 + \frac{\alpha_i^3\|z_i\|^3\overline{L}}{2\alpha_i^2(g_i^T z_i + \frac{1}{2}z_i^T G_i z_i)} \geq 1 - \frac{\alpha_i\overline{L}\overline{z}^3}{|z_i^T G_i z_i|},$$

neboli

$$\alpha_i \geq \frac{(1-\varepsilon_2)|z_i^T G_i z_i|}{\overline{L}\overline{z}^3},$$

což spolu s (90) a (91) dává

$$F(x_{i+1}) - F(x_i) \leq \varepsilon_1\frac{1}{2}\alpha_i^2 z_i^T G_i z_i \leq \frac{\varepsilon_1(1-\varepsilon_2)^2(z_i^T G_i z_i)^3}{2\overline{L}^2\overline{z}^6} \leq \frac{\varepsilon_1(1-\varepsilon_2)^2(\delta_0\underline{\lambda}^*)^3\underline{z}^6}{16\overline{L}^2\overline{z}^6} \overset{\Delta}{=} -\Delta.$$

Jelikož $F(x_i) \overset{N'}{\to} F(x^*)$, existuje index $k' \in N'$, $k' \geq k$ takový, že $F(x_i) - F(x^*) \leq \Delta/2$, pokud $i \in N'$ a $i \geq k'$. Pak platí

$$F(x_{i+1}) - F(x^*) = F(x_{i+1}) - F(x_i) + F(x_i) - F(x^*) \leq -\Delta + \frac{\Delta}{2} = -\frac{\Delta}{2} < 0,$$

pokud $i \in N'$ a $i \geq k'$, takže bod x^* není hromadným bodem posloupnosti x_i, $i \in N$. □

2.7 Maticové rozklady pro symetrické indefinitní matice

Spádové páry lze určit pomocí rozkladů symetrických indefinitních matic. Tyto rozklady lze získat vhodnou modifikací Choleského rozkladu.

Definice 26. *Nechť B je symetrická pozitivně definitní matice. Choleského rozkladem matice B rozumíme vyjádření*

$$B = R^T R,$$

kde R je regulární horní trojúhelníková matice.

Choleského rozklad lze odvodit induktivně metodou vroubení.

Věta 28. *Nechť $B_{k-1} = R_{k-1}^T R_{k-1}$ a*

$$B_k = \begin{bmatrix} B_{k-1}, & b_k \\ b_k^T, & \beta_k \end{bmatrix}, \qquad R_k = \begin{bmatrix} R_{k-1}, & r_k \\ 0, & \rho_k \end{bmatrix},$$

kde matice B_k je pozitivně definitní. Pak $B_k = R_k^T R_k$ právě tehdy, když

$$R_{k-1}^T r_k = b_k, \qquad \rho_k^2 = \beta_k - r_k^T r_k > 0. \tag{93}$$

Nové prvky matice R (prvky vektoru r_k a číslo ρ_k) jsou shora omezeny číslem $\sqrt{\beta_k}$.

Důkaz Platí

$$R_k^T R_k = \begin{bmatrix} R_{k-1}^T, & 0 \\ r_k^T, & \rho_k \end{bmatrix} \begin{bmatrix} R_{k-1}, & r_k \\ 0, & \rho_k \end{bmatrix} = \begin{bmatrix} R_{k-1}^T R_{k-1}, & R_{k-1}^T r_k \\ r_k^T R_{k-1}, & r_k^T r_k + \rho_k^2 \end{bmatrix},$$

takže $R_k^T R_k = B_k$ právě tehdy, platí-li (93). Zbývá dokázat, že $\beta_k - r_k^T r_k > 0$. Z (93) plyne, že $\beta_k - r_k^T r_k = \beta_k - b_k^T B_{k-1}^{-1} b_k$. Lze snadno dokázat (například vynásobením matic B_k a B_k^{-1} nebo blokovou eliminací), že

$$B_k^{-1} = \begin{bmatrix} B_{k-1}, & b_k \\ b_k^T, & \beta_k \end{bmatrix}^{-1} = \begin{bmatrix} B_{k-1}^{-1} + \dfrac{B_{k-1}^{-1} b_k b_k^T B_{k-1}^{-1}}{\beta_k - b_k^T B_{k-1}^{-1} b_k}, & \dfrac{B_{k-1}^{-1} b_k}{\beta_k - b_k^T B_{k-1}^{-1} b_k} \\ \dfrac{b_k^T B_{k-1}^{-1}}{\beta_k - b_k^T B_{k-1}^{-1} b_k}, & \dfrac{1}{\beta_k - b_k^T B_{k-1}^{-1} b_k} \end{bmatrix}. \tag{94}$$

Jelikož matice B_k^{-1} je (stejně jako matice B_k) pozitivně definitní, musí být její poslední diagonální prvek kladný, takže $\beta_k - r_k^T r_k > 0$. Zbytek tvrzení plyne bezprostředně z (93). □

Poznámka 56. Choleského rozklad pozitivně definitní matice je stabilní v tom smyslu, že i když matice R_{k-1} může mít malé prvky na hlavní diagonále, jsou prvky matice R_k shora omezené odmocninou z maximálního diagonálního prvku matice B_k, takže není nutné provádět permutace řádků a sloupců.

Poznámka 57. Vztahy uvedené ve větě 28 lze použít rekurentně tak, že položíme $R_1 = [B_{11}]$ a pro $2 \leq k \leq n$ pokládáme

$$R_k = \begin{bmatrix} R_{k-1}, & r_k \\ 0, & \rho_k \end{bmatrix},$$

kde $R_{k-1}^T r_k = Y_{k-1}^T B e_k$ a $\rho_k^2 = e_k^T B e_k - r_k^T r_k$ (e_k je k-tý sloupec jednotkové matice řádu n a Y_{k-1} obsahuje prvních $k-1$ sloupců jednotkové matice řádu n).

Popíšeme nyní postup, který je pro praktické použití výhodnější než způsob popsaný v poznámce 57. Předně, tak jako v důkazu věty 28, lze psát $\rho_k^2 = e_k^T (B - B Y_{k-1} (Y_{k-1}^T B Y_{k-1})^{-1} Y_{k-1}^T B) e_k$, takže je vhodné počítat a ukládat matici $Z_{k-1}^T \bar{B}_k Z_{k-1}$, kde

$$\bar{B}_k = B - B Y_{k-1} (Y_{k-1}^T B Y_{k-1})^{-1} Y_k^T B, \tag{95}$$

a kde Z_{k-1} obsahuje posledních $n-k+1$ sloupců jednotkové matice řádu n (takže $I = [Y_{k-1}, Z_{k-1}]$).

Lemma 13. *Platí* $\bar{B}_1 = B$ *a*

$$\bar{B}_k = \bar{B}_{k-1} - \frac{\bar{B}_{k-1} e_k e_k^T \bar{B}_{k-1}}{e_k^T \bar{B}_{k-1} e_k}$$

pro $2 \leq k \leq n$.

Důkaz Označme $C_k = (Y_k^T B Y_k)^{-1}$. Pak podle (94) platí

$$C_k = \begin{bmatrix} C_{k-1} + \dfrac{C_{k-1} Y_{k-1}^T B e_k e_k^T B Y_{k-1} C_{k-1}}{e_k^T \bar{B}_{k-1} e_k}, & \dfrac{C_{k-1} Y_{k-1}^T B e_k}{e_k^T \bar{B}_{k-1} e_k} \\ \dfrac{e_k^T B Y_{k-1} C_{k-1}}{e_k^T \bar{B}_{k-1} e_k}, & \dfrac{1}{e_k^T \bar{B}_{k-1} e_k} \end{bmatrix}$$

a jelikož $Y_k = [Y_{k-1}, e_k]$, můžeme psát

$$\begin{aligned}
\bar{B}_k &= B - BY_kC_kY_k^TB \\
&= B - BY_{k-1}\left(C_{k-1} - \frac{C_{k-1}Y_{k-1}^TBe_ke_k^TBY_{k-1}C_{k-1}}{e_k^T\bar{B}_{k-1}e_k}\right)Y_{k-1}^TB \\
&+ \frac{BY_{k-1}C_{k-1}Y_{k-1}^TBe_ke_k^TB}{e_k^T\bar{B}_{k-1}e_k} + \frac{Be_ke_k^TBY_{k-1}C_{k-1}Y_{k-1}^TB}{e_k^T\bar{B}_{k-1}e_k} - \frac{Be_ke_k^TB}{e_k^T\bar{B}_{k-1}e_k} \\
&= \bar{B}_{k-1} - \frac{(B - BY_{k-1}C_{k-1}Y_{k-1}^TB)e_ke_k^T(B - BY_{k-1}C_{k-1}Y_{k-1}^TB)}{e_k^T\bar{B}_{k-1}e_k} \\
&= \bar{B}_{k-1} - \frac{\bar{B}_{k-1}e_ke_k^T\bar{B}_{k-1}}{e_k^T\bar{B}_{k-1}e_k}.
\end{aligned}$$

□

Pokud počítáme a ukládáme prvky matic $Z_{k-1}^T\bar{B}_kZ_{k-1}$, $1 \le k \le n$, je výhodné určovat prvky horní trojúhelníkové matice podle vzorců $\rho_k^2 = e_k^T\bar{B}_ke_k$, $e_k^TRe_k = \rho_k$ a $e_k^TRZ_k = \rho_k^{-1}e_k^T\bar{B}_kZ_k$ (počítá se k-tý řádek matice R). Výsledkem je následující algoritmus (prvky matic $\bar{B}$ a R lze ukládat na stejná místa, kde byly uloženy odpovídající prvky matice B).

Algoritmus 2.

Krok 1 Položíme $\bar{B} := B$ a $k := 1$.

Krok 2 Jestliže $k > n$, ukončíme výpočet.

Krok 3 Nechť $\rho_k^2 := \bar{B}_{kk}$. Položíme $R_{kk} := \rho_k$ a vypočteme hodnoty $R_{kj} := \bar{B}_{kj}/\rho_k$, $k+1 \le j \le n$, a $\bar{B}_{ij} := \bar{B}_{ij} - \bar{B}_{ik}R_{kj}$, $k+1 \le i \le n$, $k+1 \le j \le n$. Položíme $k := k+1$ a přejdeme na krok 2.

Poznámka 58. Místo rozkladu $B = R^TR$ se často používá ekvivalentní rozklad $B = LDL^T$, kde D je pozitivně definitní diagonální matice a L je dolní trojúhelníková matice s jednotkami na hlavní diagonále. Platí $R = D^{1/2}L^T$ a $L = R^TD^{-1/2}$, kde matice $D^{1/2}$ obsahuje diagonální prvky matice R. Algoritmus rozkladu $B = LDL^T$ se liší od algoritmu 2 pouze tím, že se v kroku 3 pokládá $D_{kk} := \rho_k^2$, $L_{ik} := \bar{B}_{ik}/\rho_k^2$, $k+1 \le i \le n$, a $\bar{B}_{ij} := \bar{B}_{ij} - L_{ik}\bar{B}_{kj}$, $k+1 \le i \le n$, $k+1 \le j \le n$.

Není-li matice B pozitivně definitní, není zaručeno, že $\beta_k - r_k^Tr_k > 0$ v (93) a Choleského rozklad nelze použít. Přičteme-li k matici B pozitivně semidefinitní diagonální matici E tak, aby matice $B+E$ byla pozitivně definitní, lze na tuto matici aplikovat Choleského rozklad. Proces, který automaticky určuje vhodnou matici E a provádí Choleského rozklad matice $B+E$ se nazývá Gillovým-Murrayovým rozkladem matice B (je popsán v práci [66]).

Definice 27. *Nechť B je symetrická matice. Gillův-Murrayův rozklad matice B má tvar*

$$B + E = R^TR,$$

kde R je regulární horní trojúhelníková matice a E je pozitivně semidefinitní diagonální matice (je-li matice B pozitivně definitní, lze položit $E = 0$).

Označme E_{k-1} matici, která má prvních $k-1$ diagonálních prvků stejných jako matice E a ostatní prvky jsou nulové (obsahuje tedy již spočtené prvky matice E). Gillův-Murrayův rozklad se liší od Choleského rozkladu pouze tím, že $\rho_k^2 = e_k^T\bar{B}_ke_k + E_{kk}$, kde nyní

$$\bar{B}_k = B + E_{k-1} - (B + E_{k-1})Y_{k-1}(Y_{k-1}^T(B + E_{k-1})Y_{k-1})^{-1}Y_k^T(B + E_{k-1})$$

(takže, jako v důkazu věty 28, platí $e_k^T\bar{B}_ke_k = B_{kk} - e_k^TR^TRe_k$) a E_{kk} je k-tý diagonální prvek matice E. Tento prvek určujeme tak, že pokládáme

$$\rho_k^2 = \max(|e_k^T\bar{B}_ke_k|, \gamma_k^2/\beta^2, \delta^2),$$

kde

$$\gamma_k = \max_{k+1 \le j \le n} |e_k^T \bar{B}_k e_j|.$$

Pak

$$E_{kk} = \rho_k^2 - e_k^T B_k e_k = \rho_k^2 - B_{kk} + e_k^T R^T R e_k.$$

Parametry β a δ vybíráme tak, aby platilo $\beta^2 > |B_{kk}|$, $1 \le k \le n$, a $\delta^2 \ge 0$ (obvykle $\delta^2 \approx \varepsilon_M$, kde ε_M je relativní přesnost zobrazení čísla v počítači). Pokud $\delta^2 = 0$, může se stát, že $\rho_k = 0$. V tom případě však platí $e_k^T \bar{B}_k Z_{k-1} = 0$, takže lze položit $e_k^T R Z_{k-1} = 0$ (k-tý řádek matice R je nulový).

Algoritmus 3. Data $\delta \ge 0$.

Krok 1 Položíme $\bar{B} := B$. Zvolíme číslo $\beta^2 > |\bar{B}_{kk}|$, $1 \le k \le n$, a položíme $k := 1$.

Krok 2 Jestliže $k > n$, ukončíme výpočet.

Krok 3 Určíme číslo $\gamma_k = \max_{k+1 \le j \le n} |\bar{B}_{kj}|$. Nechť $\rho_k^2 := \max(|\bar{B}_{kk}|, \gamma_k^2/\beta^2, \delta^2)$. Položíme $R_{kk} := \rho_k$ a vypočteme hodnoty $R_{kj} := \bar{B}_{kj}/\rho_k$, $k+1 \le j \le n$ (pokud $\rho_k = 0$, položíme $R_{kj} := 0$, $k+1 \le j \le n$) a $\bar{B}_{ij} := \bar{B}_{ij} - \bar{B}_{ki} R_{kj}$, $k+1 \le i \le n$, $k+1 \le j \le n$. Položíme $k := k+1$ a přejdeme na krok 2.

Spádový směr $s \in R^n$ lze určit řešením soustavy rovnic $R^T R s + g = 0$, neboť matice $R^T R$ je pozitivně definitní. Ukážeme nyní, jak lze Gillův-Murrayův rozklad použít k určení směru se zápornou křivostí.

Věta 29. *Nechť $B + E = R^T R$ je Gillův-Murrayův rozklad s parametry $\beta > |B_{kk}|$, $1 \le k \le n$, a $\delta = 0$. Nechť*

$$e_l^T \bar{B}_l e_l = \min_{1 \le k \le n} e_k^T \bar{B}_k e_k$$

a nechť $z \in R^n$ je vektor určený řešením rovnice $Rz = e_l$ (e_l je l-tý sloupec jednotkové matice). Není-li matice B pozitivně semidefinitní, platí

$$z^T B z = \frac{e_l^T \bar{B}_l e_l}{\rho_l^2} < 0.$$

Důkaz Z rovnice $Rz = e_l$ plyne, že $z_l = 1/\rho_l$. Platí tedy

$$\begin{aligned} z^T B z &= z^T (B+E) z - z^T E z \le z^T R^T R z - z_l^2 E_{ll} = \\ &= e_l^T e_l - E_{ll}/\rho_l^2 = \frac{\rho_l^2 - E_{ll}}{\rho_l^2} = \frac{e_l^T \bar{B}_l e_l}{\rho_l^2}. \end{aligned}$$

Není-li matice B pozitivně semidefinitní, musí existovat index $1 \le k \le n$ takový, že $E_{kk} \ne 0$, neboli $\rho_k^2 \ne e_k^T \bar{B}_k e_k$. Mohou nastat dva případy. Buď $\rho_k^2 = |e_k^T \bar{B}_k e_k| \ne e_k^T \bar{B}_k e_k$, takže $e_k^T \bar{B}_k e_k < 0$ a tedy $e_l^T \bar{B}_l e_l < 0$, nebo $\rho_k^2 = \gamma_k^2/\beta^2$. Ve druhém případě musí existovat index $k+1 \le j \le n$ takový, že $\gamma_k = |e_k^T \bar{B}_k e_j|$, takže

$$|R_{kj}| = \frac{|e_k^T \bar{B}_k e_j|}{\rho_k} = \frac{\gamma_k}{\gamma_k/\beta} = \beta,$$

což dává

$$e_k^T \bar{B}_k e_k = \rho_k^2 - E_{kk} = B_{kk} - e_k^T R^T R e_k \le B_{kk} - \beta^2 \le |B_{kk}| - \beta^2 < 0.$$

□

Nevýhodou Gillova-Murrayova rozkladu je skutečnost, že se mění matice B, takže nelze získat řešení původní soustavy rovnic, což je potřeba například při výpočtu optimálního lokálně omezeného kroku (oddíl 6.2). Také získaný spádový pár není přijatelný (neplatí (80)). Výhodnější vlastnosti má Bunchův-Parlettův rozklad, popsaný v práci [16].

Definice 28. *Bunchův-Parlettův rozklad matice B má tvar*

$$PBP^T = LDL^T,$$

kde P je permutační matice a

$$L = \begin{bmatrix} I, & 0, & \dots, & 0 \\ L_{21}, & I, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ L_{m1}, & L_{m2}, & \dots, & I \end{bmatrix}, \quad D = \begin{bmatrix} D_{11}, & 0, & \dots, & 0 \\ 0, & D_{22}, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0, & 0, & \dots, & D_{mm} \end{bmatrix}.$$

Tedy L je dolní trojúhelníková matice s jednotkovými bloky na diagonále a D je blokově diagonální matice (bloky mají rozměr 1×1 nebo 2×2).

Výhodou Bunchova-Parlettova rozkladu je skutečnost, že se rozkládá přímo matice B k níž se nepřičítá žádná korekční matice E. Nevýhodou je nutnost permutací, bez nichž již není možné zajistit, aby absolutní hodnoty nově vznikajících prvků byly shora omezené.

Předpokládejme nejprve, že předem známe permutace i velikosti jednotlivých bloků. Pak lze rozklad provést podle vzorců uvedených v poznámce 58. Tedy $\bar{B}^{(1)} = PBP^T$ a

$$\begin{aligned} D_{kk} &= \bar{B}^{(k)}_{kk} \\ L_{ik} &= \bar{B}^{(k)}_{ik} D^{-1}_{kk}, \qquad k+1 \le i \le m, \\ \bar{B}^{k+1}_{ij} &= \bar{B}^{(k)}_{ij} - L_{ik}\bar{B}^{(k)}_{kj}, \quad k+1 \le i \le m, \quad k+1 \le j \le m, \end{aligned} \tag{96}$$

kde horní index udává pořadové číslo kroku a dolní indexy odpovídají jednotlivým blokům velikosti 1×1 nebo 2×2 (matice $\bar{B}^{(k)}$ má podobný význam jako matice $\bar{B}_k$ v (95)).

Permutace i velikosti jednotlivých bloků se vybírají tak aby docházelo k co nejmenšímu nárůstu absolutních hodnot nově vznikajících prvků. Způsob, jakým se to provádí budeme demonstrovat v případě, že $k = 1$, přičemž horní index (1) budeme vynechávat a horní index (2) nahradíme symbolem $+$. Označme

$$\alpha = \frac{\beta}{\gamma}, \qquad \beta = \max_i |\bar{B}_{ii}|, \qquad \gamma = \max_{i,j} |\bar{B}_{ij}|. \tag{97}$$

Dolní indexy jsou nyní indexy prvků matice $\bar{B}$, nikoliv jejích bloků (rozdělení na bloky ještě neznáme).

Lemma 14. *Nechť $|\bar{B}_{kl}| = \gamma$. Pak platí*

$$|\bar{B}_{kk}\bar{B}_{ll} - \bar{B}^2_{kl}| \ge \gamma^2 - \beta^2.$$

Důkaz Zřejmě $\gamma^2 \ge \beta^2$, takže $\gamma^2 - \beta^2 \ge 0$. Jelikož podle předpokladu a podle (97) platí $|\bar{B}_{kl}| = \gamma$, $|\bar{B}_{kk}| \le \beta$, $|\bar{B}_{ll}| \le \beta$, dostaneme nerovnost uvedenou v lemmatu. □

Lemma 15. *Nechť $|\bar{B}_{kk}| = \beta$. Eliminujeme-li blok 1×1 obsahující prvek $\bar{B}_{kk}$, platí*

$$\frac{\gamma^+}{\gamma} \le 1 + \frac{1}{\alpha}$$

Důkaz Použijeme-li (96), dostaneme

$$|\bar{B}_{ij}^{+}| = |\bar{B}_{ij} - \bar{B}_{ik}\bar{B}_{kk}^{-1}\bar{B}_{kj}| \leq \gamma + \frac{\gamma^2}{\beta} = \gamma\left(1 + \frac{1}{\alpha}\right)$$

pro $i \neq k$, $j \neq k$, neboť $|\bar{B}_{kk}| = \beta$.

□

Lemma 16. *Nechť $|\bar{B}_{kl}| = \gamma > \beta$. Eliminujeme-li blok 2×2 obsahující prvky $\bar{B}_{kk}$, $\bar{B}_{kl}$, $\bar{B}_{lk}$, $\bar{B}_{ll}$, platí*

$$\frac{\gamma^{+}}{\gamma} \leq 1 + \frac{2}{1-\alpha}$$

Důkaz Použijeme-li (96) a Cramerovo pravidlo, dostaneme

$$L_{ik} = \frac{\bar{B}_{ik}\bar{B}_{ll} - \bar{B}_{il}\bar{B}_{lk}}{\bar{B}_{kk}\bar{B}_{ll} - (\bar{B}_{kl})^2}, \qquad L_{il} = \frac{\bar{B}_{il}\bar{B}_{kk} - \bar{B}_{ik}\bar{B}_{kl}}{\bar{B}_{kk}\bar{B}_{ll} - (\bar{B}_{kl})^2} \tag{98}$$

a

$$\bar{B}_{ij}^{+} = \bar{B}_{ij} - L_{ik}\bar{B}_{kj} - L_{il}\bar{B}_{lj}$$

pro $i \neq k$, $i \neq l$, $j \neq k$, $j \neq l$. Podle lemmatu 14 platí

$$|\bar{B}_{kk}\bar{B}_{ll} - (\bar{B}_{kl})^2| \geq \gamma^2 - \beta^2 > 0,$$

takže

$$|L_{ik}| \leq \frac{\gamma\beta + \gamma^2}{\gamma^2 - \beta^2} = \frac{\alpha + 1}{1 - \alpha^2} = \frac{1}{1-\alpha}, \qquad |L_{il}| \leq \frac{\gamma\beta + \gamma^2}{\gamma^2 - \beta^2} = \frac{1 + \alpha}{1 - \alpha^2} = \frac{1}{1-\alpha}$$

a

$$|\bar{B}_{ij}^{+}| \leq \gamma + \gamma(|L_{ik}| + |L_{il}|) \leq \gamma\left(1 + \frac{2}{1-\alpha}\right).$$

□

Označme

$$\varphi_1(\alpha) = 1 + \frac{1}{\alpha}, \qquad \varphi_2(\alpha) = 1 + \frac{2}{1-\alpha}.$$

Protože eliminace bloku 2×2 odpovídá dvěma eliminacím bloků 1×1, budeme blok 2×2 vybírat tehdy, když $\varphi_2(\alpha) < \varphi_1^2(\alpha)$. Funkce $\varphi_1^2(\alpha)$ je klesající na intervalu $0 < \alpha \leq 1$, funkce $\varphi_2(\alpha)$ je rostoucí na intervalu $0 \leq \alpha < 1$ a rovnice $\varphi_2(\alpha) = \varphi_1^2(\alpha)$, neboli

$$4\alpha^2 - \alpha - 1 = 0,$$

má řešení $\alpha^* = (1 + \sqrt{17})/8$. Pokud $\alpha < \alpha^*$, použijeme blok 2×2. V opačném případě použijeme blok 1×1.

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 4. Data $\alpha^* = (1 + \sqrt{17})/8$.

Krok 1 Položíme $\bar{B} := B$ a $k := 1$.

Krok 2 Jestliže $k > n$, ukončíme výpočet.

Krok 3 Vypočteme čísla $\beta_k = \max_{k \leq i \leq n} |\bar{B}_{ii}|$, $\gamma_k = \max_{k \leq i,j \leq n} |\bar{B}_{ij}|$ a položíme $\alpha_k = \beta_k/\gamma_k$. Pokud $\alpha_k < \alpha^*$, přejdeme na krok 5. V opačném případě přejdeme na krok 4.

Krok 4 Nechť i je index takový, že $|\bar{B}_{ii}| = \beta_k$. Upravíme matici $\bar{B}$ tak, že vyměníme sloupec a řádek s indexem i za sloupec a řádek s indexem k. Položíme $D_{kk} := \bar{B}_{kk}$, $L_{ik} := \bar{B}_{ik}/D_{kk}$, $k+1 \leq i \leq n$, a $\bar{B}_{ij} := \bar{B}_{ij} - L_{ik}\bar{B}_{kj}$, $k+1 \leq i \leq n$, $k+1 \leq j \leq n$. Položíme $k := k+1$ a přejdeme na krok 2.

Krok 5 Nechť i, j, jsou indexy takové, že $|\bar{B}_{ij}| = \gamma_k$. Upravíme matici $\bar{B}$ tak, že vyměníme sloupce a řádky s indexy i, j za sloupce a řádky s indexy k, $k+1$. Položíme $D_{kk} := \bar{B}_{kk}$, $D_{kl} := \bar{B}_{kl}$, $D_{lk} := \bar{B}_{lk}$, $D_{ll} := \bar{B}_{ll}$, vypočteme matice L_{ik}, L_{il}, $k+1 \leq i \leq n$, podle (98) a položíme $\bar{B}_{ij} := \bar{B}_{ij} - L_{ik}\bar{B}_{kj} - L_{il}\bar{B}_{kl}$, $k+1 \leq i \leq n$, $k+1 \leq j \leq n$. Položíme $k := k+2$ a přejdeme na krok 2.

Známe-li Bunchův-Parlettův rozklad matice B, lze řešení $s \in R^n$ soustavy rovnic $Bs + g = 0$ získat pomocí řešení Ps soustavy rovnic $LDL^T(Ps) + Pg = 0$ tak, že položíme $s = P^T(Ps)$. Změníme-li matici D tak, aby byla pozitivně definitní (například tak, že její diagonální prvky nahradíme maximy z jejich absulutních hodnot a čísla $\delta > 0$ a ostatní prvky nahradíme nulami), dostaneme řešením získané soustavy rovnic spádový směr $s \in R^n$. Ukážeme nyní, jak lze Bunchův-Parlettův rozklad použít k určení směru se zápornou křivostí.

Věta 30. *Nechť $LDL^T = PBP^T$ je Bunchův-Parlettův rozklad. Nechť $v_i = 0$, pokud $\underline{\lambda}(D_{ii}) \geq 0$, a nechť v_i je normalizovaný vlastní vektor příslušný $\underline{\lambda}(D_{ii})$, pokud $\underline{\lambda}(D_{ii}) < 0$. Nechť $L^TPz = v$, kde $v^T = [v_1, \ldots, v_m]$. Není-li matice B pozitivně semidefinitní, platí*

$$z^TBz = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) < 0.$$

Důkaz Z rovnice $L^TPz = v$ dostaneme

$$z^TBz = z^TP^TLDL^TPz = v^TDv = \sum_{i=1}^{m} v_i^TD_{ii}v_i = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}).$$

Není-li matice B pozitivně semidefinitní, existuje alespoň jeden blok D_{kk} matice D, který není pozitivně semidefinitní, takže $\underline{\lambda}(D_{kk}) < 0$. Platí tedy

$$z^TBz = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) \leq \underline{\lambda}(D_{kk}) < 0.$$

□

Zbývá ukázat, jak se počítají vlastní čísla $\underline{\lambda}(D_{ii})$ a vlastní vektory v_i, $1 \leq i \leq m$. Jestliže $D_{ii} = [a_i]$ (blok 1×1), platí $\underline{\lambda}(D_{ii}) = a_i$ a $v_i = [0]$, pokud $a_i > 0$, nebo $v_i = [1]$, pokud $a_i \leq 0$. Nechť

$$D_{ii} = \begin{bmatrix} a_i, & b_i \\ b_i, & c_i \end{bmatrix}$$

(blok 2×2). Pak platí $v_i \parallel w_i$, kde

$$\begin{aligned} \underline{\lambda}(D_{ii}) &= \frac{1}{2}\left(c_i + a_i - \sqrt{(c_i - a_i)^2 + 4b_i^2}\right), \\ w_i &= \frac{1}{2}\begin{bmatrix} 2b_i \\ c_i - a_i - \sqrt{(a_i - c_i)^2 + 4b_i} \end{bmatrix}, \end{aligned}$$

takže $v_i = 0$, pokud $a_i > 0$, $b_i > 0$, $a_ic_i - b_i^2 > 0$, nebo $v_i = w_i/\|w_i\|$, v opačném pvrípadě. Můžeme se o tom přesvědčit řešením charakteristické rovnice $\det(D_{ii} - \lambda I) = 0$ a následným řešením rovnice $D_{ii}w_i = \underline{\lambda}(D_{ii})w_i$, kde za první prvek vektoru w_i volíme číslo b_i.

Kromě rozkladů indefinitních matic lze k určení spádových párů použít i některé iterační metody, jak je ukázáno v oddílu 6.7.

2.8 Metody sdružených směrů

Ukazuje se, že je účelné navrhovat metody spádových směrů tak, aby nalezly minimum ryze konvexní kvadratické funkce po konečném počtu kroků (obvykle po n krocích, kde n je počet proměnných). Takové metody obvykle konvergují n-krokově Q-superlineárně a podle věty 9 též R-superlineárě. Tuto podmínku splňují metody založené na sdruženosti směrových vektorů.

Definice 29. *Nechť $G \in R^{n\times n}$ je symetrická pozitivně definitní matice. Jestliže $s_j^T G s_i = 0$, $1 \le j < i \le n$, řekneme, že vektory $s_i \in R^n$, $1 \le i \le n$, jsou vzájemně sdružené vzhledem k matici G (nebo vzájemně G–ortogonální).*

Lemma 17. *Nenulové vzájemně G–ortogonální vektory jsou lineárně nezávislé.*

Důkaz Nechť vektory $s_i \in R^n$, $1 \le i \le k$, kde $1 \le k \le n$, jsou vzájemně G–ortogonální a nechť

$$\sum_{j=1}^{k} \alpha_j s_j = 0.$$

Pak podle definice 29 pro libovolný index $1 \le i \le k$ platí

$$\alpha_i s_i^T G s_i = s_i^T G \sum_{j=1}^{k} \alpha_j s_j = 0$$

a jelikož $s_i \neq 0$ a $G \succ 0$, dostaneme $\alpha_i = 0$. □

Lemma 18. *Nechť $s_i \in R^n$, $1 \le i \le n$, jsou nenulové vzájemně G–ortogonální vektory. Pak $G^{-1} = H$, kde*

$$H = \sum_{j=1}^{n} \frac{s_j s_j^T}{s_j^T G s_j}.$$

Důkaz Podle definice 29 pro libovolný index $1 \le i \le n$ platí

$$HGs_i = \sum_{j=1}^{n} \frac{s_j^T G s_i}{s_j^T G s_j} s_j = s_i.$$

Jelikož vektory $s_i \in R^n$, $1 \le i \le n$, jsou podle lemmatu 17 lineárně nezávislé, dostaneme $HG = I$, neboli $G^{-1} = H$. □

Nyní ukážeme, jak lze použít nenulové vzájemně G-ortogonální spádové směry s_i, $1 \le i \le n$, k nalezení minima kvadratické funkce

$$Q(x) = \frac{1}{2}(x - x^*)^T G (x - x^*) \tag{99}$$

s pozitivně definitní maticí G.

Věta 31. *(Kvadratické ukončení) Nechť x_i, $i \in N$, je posloupnost generovaná metodou spádových směrů s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0$, $i \in N$) aplikovaná na ryze konvexní kvadratickou funkci (99). Pak, jsou-li vektory s_i, $1 \le i \le n$, vzájemně G-ortogonální, existuje index $m \le n$ takový, že $g_{m+1} = 0$ a $x_{m+1} = x^*$.*

Důkaz Předpokládejme, že $g_i \neq 0$, $1 \le i \le n$ (není-li tato podmínka splněna, platí $g_{m+1} = 0$ a $x_{m+1} = x^*$ pro nějaký index $m < n$). Dokážeme indukcí, že pro libovolný index $1 \le i \le n$ platí

$$s_j^T g_{i+1} = 0, \qquad 1 \le j \le i. \tag{100}$$

Nechť pro nějaký index $1 \le i < n$ platí $s_j^T g_i = 0$, $1 \le j < i$ (indukční předpoklad). Pak můžeme psát $s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = 0$ pro $1 \le j < i$ (neboť podle indukčního předpokladu platí $s_j^T g_i = 0$ a použijeme-li

G-ortogonalitu směrových vektorů, dostaneme $s_j^T y_i = s_j^T(g_{i+1} - g_i) = \alpha_i s_j^T G s_i = 0$). Pro přesný výběr délky kroku platí $s_i^T g_{i+1} = 0$, takže $s_j^T g_{i+1} = 0$ pro $1 \le j \le i$, čímž je indukční krok dokončen. Směrové vektory s_i, $1 \le i \le n$, jsou podle (S1a) nenulové a podle lemmatu 17 lineárně nezávislé, takže z (100) plyne $g_{n+1} = 0$ a jelikož pro kvadratickou funkci (99) platí $g_{n+1} = g(x_{n+1}) = G(x_{n+1} - x^*)$, dostaneme $x_{n+1} = x^*$. □

Nyní ukážeme, jak lze použít sdruženost směrových vektorů v případě, že minimalizovaná funkce není kvadratická. Zavedeme proto pojem cyklických (nebo cyklicky přerušovaných) metod sdružených směru. Jelikož budeme vyšetřovat asymptotickou rychlost konvergence, budeme předpokládat, že posloupnost x_i, $i \in N$, konverguje k bodu $x^* \in R^n$ (takže $\|e_i\| \to 0$, kde $e_i = x_i - x^*$) a označíme

$$M = \{l \in N : l = nk + \underline{l}, k \ge 0\},$$

kde $\underline{l} \in N$ (tak jako v poznámce 14 budeme často předpokládat, že $\underline{l} = 1$).

Definice 30. *Řekneme, že metoda spádových směrů je cyklickou metodou sdružených směrů, jestliže*

$$s_j^T y_i = y_j^T s_i = o(\|e_l\|^2) \quad \forall l \le j < i < l + n, \quad l \in M,$$

což znamená že

$$\lim_{l \overset{M}{\to} \infty} \frac{s_j^T y_i}{\|e_l\|^2} = \lim_{l \overset{M}{\to} \infty} \frac{y_j^T s_i}{\|e_l\|^2} = 0, \quad \forall l \le j < i < l + n.$$

Poznámka 59. V dalších úvahách se omezíme na metody, které (a) jsou metodami stejnoměrně spádových směrů (definice 15), (b) jsou metodami gradientního typu (definice 18), (c) jsou cyklickými metodami sdružených směrů (definice 30), (d) výběr délky kroku je asymptoticky přesný (definice 19), (e) aplikujeme-li tyto metody (odstartované z bodu x_l, $l \in M$) na ryze konvexní kvadratickou funkci (99) a používáme-li přesný výběr délky kroku, jsou směrové vektory s_i, $l \le i < l + n - 1$, G-ortogonální.

Poznámka 60. Při vyšetřování asymptotické rychlosti konvergence budeme porovnávat dva iterační procesy získané stejnou cyklickou metodou sdružených směrů, základní iterační proces

$$x_{i+1} = x_i + \alpha_i s_i, \quad i \in N,$$

s asymptoticky přesným výběrem délky kroku, použitý pro minimalizaci funkce $F(x)$, a referenční iterační proces

$$\bar{x}_{i+1} = \bar{x}_i + \bar{\alpha}_i \bar{s}_i, \quad i \in N,$$

s přesným výběrem délky kroku, použitý pro minimalizaci kvadratické funkce

$$Q(x) = F(x^*) + \frac{1}{2}(x - x^*)^T G^* (x - x^*), \tag{101}$$

která má v bodě x^* stejnou hodnotu, gradient a Hessovu matici jako funkce F. Veličiny spjaté se základním iteračním procesem budeme označovat prostými symboly x_i, g_i, α_i, s_i, $e_i = x_i - x^*$, $d_i = x_{i+1} - x_i = \alpha_i s_i$, $y_i = g_{i+1} - g_i$ a veličiny spjaté s referenčním itaračním procesem budeme označovat symboly s pruhem $\bar{x}_i$, $\bar{g}_i$, $\bar{\alpha}_i$, $\bar{s}_i$, $\bar{e}_i = \bar{x}_i - x^*$, $\bar{d}_i = \bar{x}_{i+1} - \bar{x}_i = \alpha_i \bar{s}_i$, $\bar{y}_i = \bar{g}_{i+1} - g_i$. Referenční proces budeme cyklicky restartovat v bodech $x_l \in R^n$, $l \in M$, tak, že $\bar{x}_l = x_l$, $\bar{e}_l = e_l$. Dále budeme předpokládat, že směrové vektory jsou vybírány takovým způsobem, že pro $l \in M$ a $l \le i < l + n$ platí

$$\bar{s}_i = s_i(1 + o(1)) \tag{102}$$

(v oddílu 3.3 je ukázáno, že tuto podmínku splňují metody sdružených gradientů).

Lemma 19. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná cyklickou metodou sdružených směrů, která vyhovuje podmínkám uvedeným v poznámce 59. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F : R^n \to R$ vyhovující předpokladům F4 a F5. Nechť $\bar{x}_i \in R^n$, $i \in N$, je posloupnost získaná referenčním iteračním procesem uvedeným v poznámce 60, pro který platí (102). Nechť $e_i \sim e_l$ pro $l \in M$ a $l \le i < l+n$. Pak $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \in M$ a $l \le i \le l + n$.*

Důkaz Důkaz provedeme indukcí. Dokážeme navíc, že pro $l \in M$ a $l \leq i < l+n$ platí

$$\bar{e}_i = e_i(1+o(1)), \quad \bar{g}_i = g_i(1+o(1)), \quad \bar{\alpha}_i = \alpha_i(1+o(1)). \tag{103}$$

Na začátku cyklu je $\bar{e}_l = e_l$, takže $\bar{e}_l = e_l(1+o(1))$ a $e_l - \bar{e}_l = o(\|e_l\|)$. Podle věty 5 platí $\bar{g}_l = g_l + o(\|e_l\|) = g_l(1+o(1))$ a použijeme-li (103) a lemma 7, můžeme psát

$$\bar{\alpha}_l = -\frac{\bar{s}_l^T \bar{g}_l}{\bar{s}_i^T G^* \bar{s}_i} = -\frac{s_l^T g_l (1+o(1))^2}{s_l^T G^* s_l (1+o(1))^2} = \alpha_l(1+o(1)),$$

neboť hodnota $\bar{\alpha}_l = -\bar{s}_l^T \bar{g}_l / \bar{s}_l^T G^* \bar{s}_l$ realizuje pro kvadratickou funkci $Q(x)$ přesný výběr délky kroku a z asymptotické přesnosti výběru délky kroku plyne $-s_l^T g_l / s_l^T G^* s_l = \alpha_l(1+o(1))$. Předpokládejme nyní, že $e_i - \bar{e}_i = o(\|e_l\|)$ a (103) platí pro $l \leq i < l+n-1$ (indukční předpoklad).

(a) Z předpokladu (b) uvedeného v poznámce 59 plyne $s_i \sim g_i$ a předpoklady F4, F5 spolu s větou 5 implikují $g_i \sim e_i$, což spolu s $e_i \sim e_l$ dává $s_i \sim g_i \sim e_i \sim e_l$. Podobně dostaneme $s_{i+1} \sim g_{i+1} \sim e_{i+1} \sim e_l$. Použijeme-li předpoklady F4, F5 a lemma 7, můžeme psát $\alpha_i \sim 1$. Podle indukčních předpokladů a (102) platí

$$\bar{e}_{i+1} = \bar{e}_i + \bar{\alpha}_i \bar{s}_i = e_i(1+o(1)) + \alpha_i s_i (1+o(1))^2 = e_{i+1}(1+o(1)).$$

Podobně dostaneme

$$\bar{g}_{i+1} = \bar{g}_i + \bar{\alpha}_i G^* \bar{s}_i = g_i(1+o(1)) + \alpha_i G^* s_i (1+o(1))^2 = (g_i + \alpha_i G^* s_i)(1+o(1)) = g_{i+1}(1+o(1)),$$

neboť podle věty 5 lze psát $g_{i+1} = g_i + \alpha_i G^* s_i + \alpha_i s_i o(1)$ a jak jsme ukázali platí $\alpha_i s_i \sim e_l \sim g_{i+1}$.

(b) Použijeme-li (103) a lemma 7, můžeme psát

$$\bar{\alpha}_{i+1} = -\frac{\bar{s}_{i+1}^T \bar{g}_{i+1}}{\bar{s}_{i+1}^T G^* \bar{s}_{i+1}} = -\frac{s_{i+1}^T g_{i+1} (1+o(1))^2}{s_{i+1}^T G^* s_{i+1} (1+o(1))^2} = -\frac{s_{i+1}^T g_{i+1}}{s_{i+1}^T G^* s_{i+1}}(1+o(1)) = \alpha_{i+1}(1+o(1)),$$

neboť $\bar{\alpha}_{i+1} = -\bar{s}_{i+1}^T \bar{g}_{i+1} / \bar{s}_{i+1}^T G^* \bar{s}_{i+1}$ realizuje pro kvadratickou funkci $Q(x)$ přesný výběr délky kroku a z asymptotické přesnosti výběru délky kroku plyne $-s_{i+1}^T g_{i+1} / s_{i+1}^T G^* s_{i+1} = \alpha_{i+1}(1+o(1))$.

(c) Nechť $l \leq i < l+n$. Pak podle indukčních předpokladů platí

$$\bar{e}_{i+1} = \bar{e}_i + \bar{\alpha}_i \bar{s}_i = e_i(1+o(1)) + \alpha_i s_i (1+o(1))^2 = (e_i + \alpha_i s_i)(1+o(1)) = e_{i+1} + o(\|e_l\|),$$

neboť $\|e_i\| \sim \|e_l\|$, $\alpha_i \sim 1$ a $s_i \sim g_i \sim e_i \sim e_l$. Všimněme si, že k důkazu vztahu $e_{l+n} - \bar{e}_{l+n} = o(\|e_l\|)$ nepotřebujeme, aby platilo $\|e_{l+n}\| \sim \|e_l\|$. □

Poznámka 61. Tvrzení lemmatu 19 platí, pokud $\|e_i\| \sim \|e_l\|$ pro $l \leq i < l+n$. Jestliže $\|e_i\| \sim \|e_l\|$ pouze pro $l \leq i < l+m$, kde $m < n$, můžeme psát $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \leq i \leq l+m$ (plyne to z induktivní povahy důkazu).

Věta 32. *(n-kroková superlineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná cyklickou metodou sdružených směrů vyhovující podmínkám uvedeným v poznámce 59, pro kterou platí (102). Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in C^2 : \mathcal{D} \to R$ vyhovující předpokladům F4 a F5. Pak platí*

$$\lim_{l \overset{M}{\to} \infty} \frac{\|x_{l+n} - x^*\|}{\|x_l - x^*\|} = 0.$$

Důkaz Uvažujme, tak jako v lemmatu 19, dva iterační procesy (základní a referenční). Ukážeme, že $e_{l+n} = o(\|e_l\|)$, $l \in M$. Mohou nastat dva případy:

(a) Nechť pro nějaké přirozené číslo $1 \leq m < n$ neplatí $e_{l+m} \sim e_l$, $l \in M$. Protože podle poznámky 35 platí $e_{l+m} = O(\|e_l\|)$, musí být splněna podmínka $e_{l+m} = o(\|e_l\|)$. Jelikož $e_{l+n} = O(\|e_{l+m}\|)$ (poznámka 35), můžeme psát $e_{l+n} = O(\|e_{l+m}\|) = o(\|e_l\|)$, $l \in M$.

(b) Podle podmínky (e) uvedené v poznámce 59 a podle věty 31 existuje přirozené číslo $1 \leq m \leq n$ takové, že $\bar{e}_{l+m} = 0$, $l \in M$. Použijeme-li tvrzení lemmatu 19 (které podle poznámky 61 platí pro $l \leq i \leq l+m$), dostaneme

$$\|e_{l+m}\| \leq \|\bar{e}_{l+m}\| + \|e_{l+m} - \bar{e}_{l+m}\| = o(\|e_l\|).$$

Jelikož $e_{l+n} = O(\|e_{l+m}\|)$ (poznámka 35), platí $\|e_{l+n}\| = O(\|e_{l+m}\|) = o(\|e_l\|)$, $l \in M$. □

Poznámka 62. Podle věty 9 a poznámky 15 je cyklická metoda sdružených směrů vyhovující předpokladům věty 32 R-superlineárně konvergentní.

Větu 32 lze použít pro různé třídy cyklických metod sdružených směrů, které vyhovují jejím předpokladům. V oddílu 3.3 (věta 47) je ukázáno, že těmto předpokladům vyhovují cyklicky přerušované metody sdružených gradientů. Větu 32 by bylo možné aplikovat i na cyklicky přerušované metody s proměnnou metrikou. To však není nutné, neboť metody s proměnnou metrikou jsou Q-superlineárně konvergentní i bez přerušování iteračního procesu (věta 103).

Věta 31 neplatí, nepoužíváme-li přesný výběr délky kroku. V tomto případě je možné metody sdružených směrů korigovat použitím následující věty.

Věta 33. *Nechť $\bar{x}_{i+1} = \bar{x}_i + \bar{\alpha}_i \bar{s}_i$, $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, jsou dva iterační procesy aplikované na kvadratickou funkci (99), které vycházejí ze stejného bodu $\bar{x}_1 = x_1$ a ve kterých se používají stejné nenulové G-ortogonální směrové vektory $\bar{s}_i = s_i$, $1 \leq i \leq n$, přičemž čísla $\bar{\alpha}_i$ se určují pomocí přesného výběru délky kroku a $\alpha_i \neq 0$. Pak pro $1 \leq i \leq n$ platí*

$$\bar{x}_{i+1} = x_{i+1} - \sum_{j=1}^{i} \frac{d_j^T g_{j+1}}{y_j^T d_j} d_j, \qquad \bar{d}_i = d_i - \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i = -\frac{d_i^T g_i}{y_i^T d_i} d_i, \tag{104}$$

$$\bar{g}_{i+1} = g_{i+1} - \sum_{j=1}^{i} \frac{d_j^T g_{j+1}}{y_j^T d_j} y_j, \qquad \bar{y}_i = y_i - \frac{d_i^T g_{i+1}}{y_i^T d_i} y_i = -\frac{d_i^T g_i}{y_i^T d_i} y_i. \tag{105}$$

Důkaz Jelikož směrové vektory s_i, $1 \leq i \leq n$, jsou nenulové a $\alpha_i \neq 0$, platí $d_i = \alpha_i s_i \neq 0$ takže $y_i^T d_i = d_i^T G d_i \neq 0$, $1 \leq i \leq n$, a jmenovatele v dokazovaných vztazích jsou nenulové. Jelikož pro kvadratickou funkci (99) platí $y_i = G d_i$, $1 \leq i \leq n$, a matice G je pozitivně definitní, jsou rovnosti (104) ekvivalentní rovnostem (105). Stačí tedy dokázat rovnosti (105). Důkaz provedeme indukcí. Předpokládejme, že pro nějaký index $i \leq n$ platí

$$\bar{g}_i = g_i - \sum_{j=1}^{i-1} \frac{d_j^T g_{j+1}}{y_j^T d_j} y_j$$

(platí to zcela jistě pro $i = 1$, neboť $\bar{x}_1 = x_1$ a tedy i $\bar{g}_1 = g_1$). Protože $\bar{s}_i = s_i$ a $d_i \neq 0$, existuje číslo λ_i takové, že $\bar{d}_i - d_i = \lambda_i d_i$ a tedy $\bar{y}_i - y_i = \lambda_i y_i$. Použijeme-li indukční předpoklad, dostaneme

$$\bar{g}_{i+1} = \bar{g}_i + \bar{y}_i = g_i - \sum_{j=1}^{i-1} \frac{d_j^T g_{j+1}}{y_j^T d_j} y_j + y_i + (\bar{y}_i - y_i) = g_{i+1} - \sum_{j=1}^{i-1} \frac{d_j^T g_{j+1}}{y_j^T d_j} y_j + \lambda_i y_i. \tag{106}$$

Jelikož parametr $\bar{\alpha}_i$ určujeme pomocí přesného výběru délky kroku, musí platit7 $\bar{s}_i^T \bar{g}_{i+1} = 0$, což spolu s $\bar{s}_i = s_i$ a (106) dává

$$\bar{s}_i^T \bar{g}_{i+1} = s_i^T g_{i+1} - \sum_{j=1}^{i-1} \frac{d_j^T g_{j+1}}{y_j^T d_j} s_i^T y_j + \lambda_i s_i^T y_i = s_i^T g_{i+1} + \lambda_i s_i^T y_i = 0,$$

neboť z G-ortogonality směrových vektorů plyne $s_i^T y_j = s_i^T G d_j = \alpha_j s_i^T G s_j = 0$, $1 \leq j \leq i-1$. Platí tedy

$$\lambda_i = -\frac{s_i^T g_{i+1}}{s_i^T y_i} = -\frac{d_i^T g_{i+1}}{y_i^T d_i},$$

neboť $d_i = \alpha_i s_i$ a $\alpha_i \neq 0$. Dosadíme-li tuto hodnotu do (106), dostaneme první rovnost v (105). Ze vztahu $\bar{y}_i - y_i = \lambda_i y_i$ pak plyne druhá rovnost v (105). □

Větu 33 můžeme využít dvojím způsobem. Jelikož většina metod sdružených směrů generuje G-ortogonální směrové vektory pomocí vektorů g_{i+1}, d_i a y_i, můžeme tyto vektory nahradit vektory $\bar{g}_{i+1}$, $\bar{d}_i$ a $\bar{y}_i$ určenými podle vzorců (104) a (105). Pak G-ortogonalita směrových vektorů nezávisí na výběru délky kroku. V případě metod s proměnnou metrikou z Broydenovy třídy, popsané v oddílu 4.1, není nutné nahražovat vektory d a y. Z vyjádření (286) a (306) plyne, že matice H_+ a B_+ se nezmění, vynásobíme-li vektory d a y stejným číslem a tedy také nahradíme-li je vektory $\bar{d}$ a $\bar{y}$. Matice H_+ a B_+ určené metodou s proměnnou metrikou z Broydenovy třídy jsou tedy invariantní vůči korekcím (104) a (105). Podobnou vlastnost má metoda sdružených gradientů Hestenese a Stiefela, popsaná v oddílu 3.1, kde vektor $(y_i^T g_{i+1}/y_i^T s_i)s_i = (y_i^T g_{i+1}/y_i^T d_i)d_i$, přičítaný k záporně vzatému gradientu, se nezmění, použijeme-li místo vektorů d a y vektory $\bar{d}$ a $\bar{d}$. V těchto případech není nutné korigovat vektory d a y, je však nutné korigovat vektor g_{i+1}, pokud chceme splnit předpoklady věty 33. Máme-li k disposici metodu sdružených směrů, která generuje G-ortogonální směrové vektory nezávisle na výběru délky kroku, například metodu s proměnnou metrikou z Davidonovy třídy vyšetřované v oddílu 4.9, stačí použít korekční krok tvaru (104). V tomto případě pokládáme $x_{n+2} = x_{n+1} + \alpha_{n+1}s_{n+1}$, kde $\alpha_{n+1} = 1$ a $s_{n+1} = \bar{x}_{n+1} - x_{n+1}$. Vektor s_{n+1} lze určit rekurentním postupem tak, že $s_{n+1} = v_{n+1}$, kde

$$v_1 = 0 \quad \text{a} \quad v_{i+1} = v_i - \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i \quad \text{pro} \quad 1 \leq i \leq n.$$

Poznamenejme, že tyto korekce mají spíše teoretický význam, neboť v případech kdy minimalizovaná funkce není kvadratická, nezlepšují efektivitu metod sdružených směrů.

Věta 31 zaručuje, že metody sdružených směrů s přesným výběrem délky kroku naleznou minimum ryze konvexní kvadratické funkce po konečném počtu kroků. Metody sdružených směrů lze však zobecnit tak, že naleznou minimum složitějších funkcí po konečném počtu kroků. Takovými funkcemi jsou například zobecněné kvadratické funkce.

Definice 31. *Zobecněnou kvadratickou funkcí nazveme funkci $F : R^n \to R$ určenou vzorcem*

$$F(x) = \varphi(Q(x)), \tag{107}$$

kde $Q : R^n \to R$ je kvadratická funkce s pozitivně definitní maticí $\tilde{G}$, tvaru

$$Q(x) = Q(x^*) + \frac{1}{2}(x - x^*)^T \tilde{G}(x - x^*), \tag{108}$$

a $\varphi : R \to R$ je spojitá funkce diferencovatelná pro $Q > Q(x^*)$ taková, že

$$Q > Q(x^*) \quad \Rightarrow \quad \frac{\partial \varphi(Q)}{\partial Q} > 0. \tag{109}$$

Zobecněná kvadratická funkce (107) má podobné vlastnosti jako kvadratická funkce $Q(x)$.

Věta 34. *Bod $x^* \in R^n$ je globálním minimem zobecněné kvadratické funkce (107).*

Důkaz. Jelikož $\tilde{G} > 0$, je bod $x^* \in R$ globálním minimem kvadratické funkce (108), takže $Q(x) \geq Q(x^*)$, $\forall x \in R^n$. Označme $Q^* = Q(x^*)$ a $F^* = F(x^*)$. Jestliže $Q = Q^*$, pak podle (107) je také $F = F^*$. Jestliže $Q > Q^*$ můžeme použít větu o střední hodnotě, takže

$$F - F^* = \frac{\partial \varphi\left(Q^* + \lambda(Q - Q^*)\right)}{\partial Q}\,(Q - Q^*),$$

kde $0 < \lambda < 1$. Jelikož $Q > Q^* = 0$ a $0 < \lambda < 1$, je také $Q^* + \lambda(Q - Q^*) > Q^* = 0$ takže podle (109) platí $\partial\varphi(Q^* + \lambda(Q - Q^*))/\partial Q > 0$, což spolu s $Q - Q^* > 0$ dává $F - F^* > O$. Dokázali jsme tedy, že

$F(x) \geq F(x^*)$, $\forall x \in R^n$, takže bod $x^* \in R^n$ je globálním minimem zobecněné kvadratické funkce (107). □

Při vyšetřování metod sdružených směrů, které naleznou minimum zobecněné kvadratické funkce (107) po konečném počtu kroků, budeme používat označení

$$\sigma(x) = \frac{\partial \varphi(Q(x))}{\partial Q}, \tag{110}$$

takže $\sigma(x) > 0$, pokud $x \neq x^*$. Derivujeme-li vztah (107) pro $x \neq x^*$, dostaneme

$$g(x) = \sigma(x)\tilde{g}(x), \qquad \tilde{g}(x) = \tilde{G}(x - x^*). \tag{111}$$

Známe-li hodnotu $\sigma(x)$, můžeme z gradientu $g(x)$ zobecněné kvadratické funkce $F(x)$ určit gradient $\tilde{g}(x)$ kvadratické funkce $Q(x)$. Navíc z $s^T g(x) = 0$ plyne $s^T \tilde{g}(x) = 0$, takže přesný výběr délky kroku realizovaný pomocí zobecněné kvadratické funkce $F(x)$ dává stejný výsledek jako přesný výběr délky kroku realizovaný pomocí kvadratické funkce $Q(x)$. Z těchto úvah plyne, že nahradíme-li ve vzorcích používaných v oddílech 3.1 a 4.1 vektor $g(x)$ vektorem $\tilde{g} = g(x)/\sigma(x)$, dostaneme metody sdružených směrů, které naleznou minimum kvadratické funkce $Q(x)$ po konečném počtu kroků. Podle věty 25 je minimum kvadratické funkce $Q(x)$ totožné s minimem zobecněné kvadratické funkce $F(x)$. Dostaneme tedy metody sdružených směrů, které naleznou minimum zobecněné kvadratické funkce $F(x)$ po konečném počtu kroků.

Hlavním problémem při realizaci metod sdružených směrů pro zobecněné kvadratické funkce je určit hodnotu $\sigma(x)$. Dříve než přejdeme k obecnému případu, vyšetříme dva speciální typy zobecněných kvadratických funkcí, pro které je možné určit podíl hodnot σ_+ a σ pomocí hodnot a gradientů funkce (107) v bodech x_+ a x. Budeme přitom používat označení $b_+ = (x_+ - x)^T g_+$, $b = (x_+ - x)^T g$ a

$$A = \frac{F_+ - F}{b}, \qquad B = \frac{b_+}{b}. \tag{112}$$

Věta 35. *Nechť Q, Q_+ a $\tilde{g}$, $\tilde{g}_+$ jsou hodnoty a gradienty kvadratické funkce (108) v bodech $x \in R^n$, $x_+ \in R^n$. Pak platí*

$$2(Q_+ - Q) = (x_+ - x)^T \tilde{g}_+ + (x_+ - x)^T \tilde{g}. \tag{113}$$

Důkaz. Podle (108) platí

$$Q = Q^* + \frac{1}{2}(x - x^*)^T \tilde{g}, \qquad Q_+ = Q^* + \frac{1}{2}(x_+ - x^*)^T \tilde{g}_+,$$

neboť $\tilde{g} = \tilde{G}(x - x^*)$ a $\tilde{g}_+ = \tilde{G}(x_+ - x^*)$. Můžeme tedy psát

$$\begin{aligned} 2(Q_+ - Q) &= (x_+ - x^*)^T \tilde{g}_+ - (x - x^*)^T \tilde{g} \\ &= (x_+ - x)^T \tilde{g}_+ + (x - x^*)^T \tilde{G}(x_+ - x^*) - (x - x_+)^T \tilde{g} - (x_+ - x^*)^T \tilde{G}(x - x^*) \\ &= (x_+ - x)^T \tilde{g}_+ + (x_+ - x)^T \tilde{g}. \end{aligned}$$

□

Nejprve budeme předpokládat, že

$$F(x) = F(x^*) + a_1 Q(x) + a_2 Q^2(x), \tag{114}$$

kde $a_1 \geq 0$ a $a_2 > 0$.

Věta 36. *Uvažujme zobecněnou kvadratickou funkci tvaru (114), kde $a_1 \geq 0$ a $a_2 > 0$. Pak podíl $t = \sigma_+/\sigma$ je kořenem kvadratické rovnice*

$$t^2 - (4A - B - 1)t + B = 0. \tag{115}$$

Důkaz. Derivujeme-li vztah (114), dostaneme $g(x) = (a_1 + 2a_2Q(x))\tilde{g}(x)$, takže podle (111) platí

$$\sigma(x) = a_1 + 2a_2Q(x), \tag{116}$$

což s použitím (114) dává

$$\sigma^2(x) = a_1^2 + 4a_1a_2Q(x) + 4a_2^2Q^2(x) = a_1^2 + 4a_2(a_1Q(x) + a_2Q^2(x)) = a_1^2 + 4a_2(F(x) - F(x^*)),$$

takže

$$\sigma_+^2 - \sigma^2 = 4a_2(F_+ - F) \quad \Rightarrow \quad a_2 = \frac{\sigma_+^2 - \sigma^2}{4(F_+ - F)}. \tag{117}$$

Podle (116) můžeme psát

$$Q(x) = \frac{\sigma(x) - a_1}{2a_2},$$

takže s použitím (113) dostaneme

$$2\left(\frac{\sigma_+ - a_1}{2a_2} - \frac{\sigma - a_1}{2a_2}\right) = \frac{b_+}{\sigma_+} + \frac{b}{\sigma} \quad \Rightarrow \quad \frac{\sigma_+ - \sigma}{a_2} = \frac{b_+}{\sigma_+} + \frac{b}{\sigma}.$$

Dosadíme-li do této rovnice vyjádření (117), můžeme psát

$$\frac{4(F_+ - F)}{\sigma_+ + \sigma} = \frac{b_+}{\sigma_+} + \frac{b}{\sigma},$$

neboli

$$4(F_+ - F) = \left(1 + \frac{\sigma}{\sigma_+}\right)\left(b_+ + \frac{\sigma_+}{\sigma}\, b\right),$$

což po dosazení $t = \sigma_+/\sigma$ a po úpravě využívající vztahy (112) dává (115). □

Při rozboru rovnice (115) se omezíme pouze na případ, kdy platí $b < 0 \le b_+$ a $F_+ < F$, takže $A > 0$ a $B \le 0$ (tyto podmínky lze obvykle splnit při výběru délky kroku). Pak obdržíme

$$(4A - B - 1)^2 - 4B \ge 0,$$

takže rovnice (115) má reálné kořeny. Rovnice (115) má právě jeden kladný kořen, pokud platí $B < 0$, nebo $B = 0$ a $4A - B - 1 > 0$.

Dále budeme předpokládat, že

$$F(x) = F(x^*) + a_1Q^p(x), \tag{118}$$

kde $a_1 > 0$ a $p > 1$.

Věta 37. *Uvažujme zobecněnou kvadratickou funkci tvaru (118), kde $a_1 > 0$ a $p > 1$. Označme $q = p/(p-1)$. Pak podíl $t = \sigma_+/\sigma$ je kořenem nelineární rovnice*

$$t^{q+1} + (B - 2Ap)t^q - (1 - 2Ap)t - B = 0. \tag{119}$$

Důkaz. Derivujeme-li vztah (118), dostaneme

$$g(x) = a_1pQ^{p-1}(x)\tilde{g}(x),$$

takže podle (111) platí

$$\sigma(x) = a_1p\,Q^{p-1}(x). \tag{120}$$

Hodnotu $Q(x)$ můžeme určit řešením rovnice (118). Platí

$$Q(x) = \left(\frac{F(x) - F(x^*)}{a_1}\right)^{1/p}.$$

Dosadíme-li tento vztah do (120), dostaneme

$$\sigma(x) = a_1 p \left(\frac{F(x) - F(x^*)}{a_1} \right)^{(p-1)/p},$$

takže

$$\left(\frac{\sigma_+}{\sigma}\right)^q = \frac{F_+ - F(x^*)}{F - F(x^*)}. \tag{121}$$

Porovnáme-li vztahy (118) a (120), můžeme psát

$$Q(x) = \frac{p\,(F(x) - F(x^*))}{\sigma(x)},$$

což spolu s (113) dává

$$2p \left(\frac{F_+ - F(x^*)}{\sigma_+} - \frac{F - F(x^*)}{\sigma} \right) = \frac{b_+}{\sigma_+} + \frac{b}{\sigma},$$

neboli

$$2p \left(\frac{F_+ - F(x^*)}{F - F(x^*)} - \frac{\sigma_+}{\sigma} \right) = \frac{1}{F - F(x^*)} \left(b_+ + b\frac{\sigma_+}{\sigma} \right). \tag{122}$$

Ale

$$\frac{F_+ - F(x^*)}{F - F(x^*)} = 1 + \frac{F_+ - F}{F - F(x^*)}, \quad \Rightarrow \quad \frac{F_+ - F}{F - F(x^*)} = \frac{F_+ - F(x^*)}{F - F(x^*)} - 1.$$

Dosadíme-li tento vztah spolu s (121) do (122), dostanemee

$$2p(F_+ - F) \left(\left(\frac{\sigma_+}{\sigma}\right)^q - \frac{\sigma_+}{\sigma} \right) = \left(\left(\frac{\sigma_+}{\sigma}\right)^q - 1 \right) \left(b_+ + b\frac{\sigma_+}{\sigma} \right),$$

což po dosazení $t = \sigma_+/\sigma$ a po úpravě využívající vztahy (112) dává (119). □

Rovnice (119) má triviální řešení $t = 1$. Nás zajímá pouze netriviální řešení $t \neq 1$. Omezíme se opět na případ, kdy platí $b < 0 \leq b_+$ a $F_+ < F$, takže $A > 0$ a $B \leq 0$. Označme $\psi(t)$ levou stranu rovnice (119). Pak platí $\psi'(1) = (1 + B - 2A)q$, kde $q > 1$, neboť $p > 1$. Předpokládejme, že buď $B < 0$, nebo $2Ap > 1$. Pak pro dostatečně malé i dostatečně velké hodnoty parametru t platí $\psi(t) > 0$. Mohou tedy nastat tyto případy:

(a) Jestliže $2A - B - 1 < 0$, má rovnice (119) netriviální řešení $0 < t < 1$.
(b) Jestliže $2A - B - 1 = 0$, není zaručena existence netriviálního řešení.
(c) Jestliže $2A - B - 1 > 0$, má rovnice (119) netriviální řešení $t > 1$.

Rovnici (119) musíme řešit numericky (například metodou půlení intervalu).

V obecném případě, kdy neznáme tvar funkce $\varphi : R \to R$ v (107), nemůžeme použít lemma 35, neboť neznáme explicitní vyjádření pro hodnotu kvadratické funkce $Q : R^n \to R$. V tomto případě nestačí k určení podílu hodnot σ_+ a σ pouze hodnoty a gradienty funkce $F : R^n \to R$ v bodech x_+ a x. Platí však tato věta.

Věta 38. *Uvažujme zobecněnou kvadratickou funkci definovanou vztahem (107). Nechť $x_- = x + \alpha_- s$ a $x_+ = x + \alpha_+ s$ jsou dva různé body takové, že gradienty g_- a g_+ jsou lineárně nezávislé. Pak platí*

$$\frac{\sigma}{\sigma_-} = \frac{\|g_+\|^2 g^T g_- - g_-^T g_+ g^T g_+}{\|g_+\|^2 - (g_-^T g_+)^2} \frac{\alpha_+ - \alpha_-}{\alpha_+}, \tag{123}$$

$$\frac{\sigma}{\sigma_+} = \frac{\|g_-\|^2 g^T g_+ - g_-^T g_+ g^T g_-}{\|g_+\|^2 - (g_-^T g_+)^2} \frac{\alpha_- - \alpha_+}{\alpha_-}. \tag{124}$$

Důkaz. Použijeme-li vzorec (111), můžeme psát $g = \sigma\tilde{g}$ a

$$g_- = \sigma_-\tilde{g}_- = \sigma_-(\tilde{g} + \alpha_-\tilde{G}s)\,, \qquad g_+ = \sigma_+\tilde{g}_+ = \sigma_+(\tilde{g} + \alpha_+\tilde{G}s)\,.$$

Platí tedy

$$\frac{g_-}{\sigma_-} - \frac{g}{\sigma} = \alpha_-\tilde{G}s\,, \qquad \frac{g_+}{\sigma_+} - \frac{g}{\sigma} = \alpha_+\tilde{G}s\,,$$

takže

$$\left(\frac{g_-}{\sigma_-} - \frac{g}{\sigma}\right)\alpha_+ = \left(\frac{g_+}{\sigma_+} - \frac{g}{\sigma}\right)\alpha_-\,. \tag{125}$$

Vynásobíme-li tuto rovnost skalárně vektory g_- a g_+, dostaneme soustavu rovnic

$$\begin{aligned}
\|g_-\|^2\frac{\alpha_+}{\sigma_-} - (g_-)^Tg_+\frac{\alpha_-}{\sigma_+} &= \frac{\alpha_+ - \alpha_-}{\sigma}(g_-)^Tg\,,\\
-(g_+)^Tg_-\frac{\alpha_+}{\sigma_-} + \|g_+\|^2\frac{\alpha_-}{\sigma_+} &= \frac{\alpha_- - \alpha_+}{\sigma}(g_+)^Tg\,,
\end{aligned}$$

která má řešení (123)–(124). Determinant této soustavy je nenulový, neboť vektory g_- a g_+ jsou lineárně nezávislé. □

Poznámka 63. Rovnice (125) ukazuje, že pro zobecněnou kvadratickou funkci jsou gradienty g, g_- a g_+, určené v bodech x, x_- a x_+ ležících na přímce, lineárně závislé. Podmínka

$$\det\left([g, g_-, g_+]^T[g, g_-, g_+]\right) \approx 0$$

může sloužit k ověření vhodnosti zobecněné kvadratické funkce jako modelu pro obecnou nelineární funkci. Vzorce (123)–(124) můžeme použít pouze tehdy, platí-li $\sigma/\sigma_- > 0$ a $\sigma/\sigma_+ > 0$.

Pro zobecněnou kvadratickou funkci můžeme snadno realizovat přesný výběr délky kroku.

Věta 39. *Nechť jsou splněny předpoklady věty 38, přičemž $s^Tg_-/\sigma_- \neq s^Tg/\sigma \neq 0$. Pak podmínka $s^Tg_+ = 0$ je splněna právě tehdy, když*

$$\frac{\alpha_+}{\alpha_-} = \frac{1}{1 - \dfrac{\sigma}{\sigma_-}\dfrac{s^Tg_-}{s^Tg}}. \tag{126}$$

Důkaz. Vynásobíme-li rovnici (125) skalárně vektorem s, dostaneme

$$\left(\frac{s^Tg_-}{\sigma_-} - \frac{s^Tg}{\sigma}\right)\alpha_+ = \left(\frac{s^Tg_+}{\sigma_+} - \frac{s^Tg}{\sigma}\right)\alpha_-,$$

takže podmínka $s^Tg_+ = 0$ je splněna právě tehdy, platí-li (126). □

Metody sdružených směrů pro zobecněné kvadratické funkce jsou účinné, je-li minimalizovaná funkce zobecněnou kvadratickou funkcí. Pro obecné minimalizované funkce se tato výhoda obvykle neprojeví a vzhledem k tomu, že rovnice (123)–(124) obsahují několik skalárních součinů, je výhodnější používat standardní metody sdružených směrů. Totéž platí i pro další metody sdružených směrů, které umožňují nalézd minimum ještě složitějších funkcí po konečném počtu kroků. Například v práci [37], je popsána metoda, která nalezne minimum konické funkce po konečném počtu kroků a v práci [95] metoda, která nalezne minimum zobecněné konické funkce po konečném počtu kroků.

3 Metody sdružených gradientů

Metody sdružených gradientů jsou nejjednoduššími metodami spádových směrů založenými na principu konjugovanosti. Princip konjugovanosti, použitý také u metod s proměnnou metrikou, má za následek podstatné urychlení konvergence oproti metodě největšího spádu. První lineární metodu sdružených gradientů pro řešení soustavy lineárních rovnic se symetrickou pozitivně definitní maticí (neboli pro minimalizaci ryze konvexní kvadratické funkce) popsali Hestenes a Stiefel v práci [81]. První nelineární metodu sdružených gradientů publikovali Fletcher a Reeves v práci [57]. Nelineární metoda sdružemých gradientů se od lineární metody liší tím, že se místo přesného výběru délky kroku, definovaného vzorcem v němž se vyskytuje Hessova matice, používá nepřesný výběr délky kroku, založený na splnění některé z Wolfeho podmínek (poznámka 22), a tím, že se gradienty minimalizované funkce vyčíslují, místo toho, aby se počítaly rekurentné. Nepřesný výběr délky kroky způsobuje, že se porušuje sdruženost směrových vektorů, což otvírá možnosti pro různé modifikace základních vzorců.

3.1 Základní metody sdružených gradientů

Definice 32. *Řekneme, že metoda spádových směrů (definice 17) je metodou sdružených gradientů, jestliže*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{127}$$

kde parametr β_i se vybírá tak, aby směrové vektory s_i, $1 \leq i \leq n$, byly sdružené (nebo G–ortogonální, podmínka (131)), aplikujeme-li tuto metodu na ryze konvexní kvadratickou funkci (99) a používáme-li přesný výběr délky kroku.

Označme $d_i = x_{i+1} - x_i = \alpha_i s_i$ a $y_i = g_{i+1} - g_i$. Pak pro kvadratickou funkci (99) platí $y_i = Gd_i$ a podmínku G–ortogonality vektorů s_i, s_{i+1} lze zapsat ve tvaru $\alpha_i s_i^T G s_{i+1} = y_i^T s_{i+1} = 0$ (předpokládáme, že $\alpha_i \neq 0$). Odtud prostřednictvím (127) dostaneme rovnici $\beta_i y_i^T s_i - y_i^T g_{i+1} = 0$, neboli

$$\beta_i = \frac{y_i^T g_{i+1}}{y_i^T s_i}. \tag{128}$$

Ukážeme, že tato volba již zaručuje vzájemnou G–ortogonalitu směrových vektorů s_i, $1 \leq i \leq n$, a nalezení minima ryze konvexní kvadratické funkce (99) po konečném počtu kroků (je-li výběr délky kroku přesný).

Věta 40. *(Kvadratické ukončení) Nechť x_i, $i \in N$, je posloupnost generovaná metodou sdružených gradientů (127) a (128) s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0$, $i \in N$) aplikovaná na ryze konvexní kvadratickou funkci (99). Pak existuje index $m \leq n$ takový, že $g_{m+1} = 0$ a $x_{m+1} = x^*$.*

Důkaz Předpokládejme, že $g_i \neq 0$, $1 \leq i \leq n$ (není-li tato podmínka splněna, platí $g_{m+1} = 0$ a $x_{m+1} = x^*$ pro nějaký index $m < n$). Dokážeme indukcí, že pro libovolný index $1 \leq i \leq n$ je $s_i \neq 0$, $\alpha_i > 0$, přičemž platí

$$s_j^T g_{i+1} = 0, \qquad 1 \leq j \leq i, \tag{129}$$

$$g_j^T g_{i+1} = 0, \qquad 1 \leq j \leq i, \tag{130}$$

$$s_j^T G s_i = 0, \qquad 1 \leq j < i, \tag{131}$$

$$s_j^T y_i = y_j^T s_i = 0 \qquad 1 \leq j < i. \tag{132}$$

Rovnosti (131) a (132) jsou ekvivalentní, neboť pro kvadratickou funkci (99) platí

$$y_i = g_{i+1} - g_i = G(x_{i+1} - x_i) = Gd_i = \alpha_i G s_i$$

a $\alpha_i > 0$ podle předpokladu. Z (130) plyne, že nenulové gradienty g_i, $1 \leq i \leq n$, jsou vzájemně ortogonální, tudíž lineárně nezávislé, takže nutně $g_{n+1} = 0$ a $x_{n+1} = x^*$. Pro $i = 1$ je $s_1^T g_1 = -g_1^T g_1 < 0$, takže $s_1 \neq 0$

a $\alpha_1 > 0$. Nechť pro nějaký index $1 \leq i < n$ platí $s_j \neq 0$, $\alpha_j > 0$, pro $1 \leq j \leq i$, a $s_j^T g_i = 0$, $g_j^T g_i = 0$, $s_j^T G s_i = s_j^T y_i = y_j^T s_i = 0$, pro $1 \leq j < i$ (indukční předpoklad).

(a) Zřejmě $s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = 0$ pro $1 \leq j < i$ (neboť podle indukčního předpokladu pro $1 \leq j < i$ platí $s_j^T g_i = 0$ a $s_j^T y_i = 0$). Z přesného výběru délky kroku plyne, že $s_i^T g_{i+1} = 0$. Platí tedy $s_j^T g_{i+1} = 0$ pro $1 \leq j \leq i$.

(b) Použijeme-li (127), dostaneme

$$\begin{aligned} g_1 &= -s_1, \\ g_j &= -s_j + \beta_{j-1} s_{j-1}, \qquad 1 < j \leq i, \end{aligned}$$

takže podle (a) platí

$$\begin{aligned} g_1^T g_{i+1} &= -s_1^T g_{i+1} = 0, \\ g_j^T g_{i+1} &= -s_j^T g_{i+1} + \beta_{j-1} s_{j-1}^T g_{i+1} = 0, \qquad 1 < j \leq i. \end{aligned}$$

(c) Použijeme-li (127) a (a) dostaneme

$$s_{i+1}^T g_{i+1} = -g_{i+1}^T g_{i+1} + \beta_i s_i^T g_{i+1} = -g_{i+1}^T g_{i+1} < 0,$$

takže $s_{i+1} \neq 0$ a $\alpha_{i+1} > 0$. Z (127) a (b) plyne, že

$$y_j^T s_{i+1} = -y_j^T g_{i+1} + \beta_i y_j^T s_i = -y_j^T g_{i+1} = -(g_{j+1} - g_j)^T g_{i+1} = 0$$

pro $1 \leq j < i$ (neboť podle indukčního předpokladu pro $1 \leq j < i$ platí $y_j^T s_i = 0$). Dále podle (127) a (128) platí

$$y_i^T s_{i+1} = -y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} y_i^T s_i = 0,$$

takže $y_j^T s_{i+1} = 0$ (a tedy také $s_j^T y_{i+1} = 0$ a $s_j^T G s_{i+1} = 0$) pro $1 \leq j \leq i$. Tím je indukční krok dokončen □

Poznámka 64. Z rovností $s_j^T g_{i+1} = 0$, $1 \leq j \leq i$, vyplývá, že bod x_{i+1} realizuje minimum ryze konvexní kvadratické funkce (99) na podprostoru generovaném vektory s_j, $1 \leq j \leq i$.

Poznámka 65. Používáme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i$, takže

$$(y_i - d_i)^T g_{i+1} = y_i^T g_{i+1} - \alpha_i s_i^T g_{i+1} = y_i^T g_{i+1},$$

a podle (127) můžeme psát

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i = -g_i^T s_i = g_i^T g_i - \beta_{i-1} g_i^T s_{i-1} = g_i^T g_i$$

(jelikož předpokládáme, že $g_i^T s_i < 0$, píšeme obvykle $|g_i^T s_i|$ místo $-g_i^T s_i$). Je-li navíc minimalizovaná funkce kvadratická, platí (130), takže

$$y_i^T g_{i+1} = g_{i+1}^T g_{i+1} - g_i^T g_{i+1} = g_{i+1}^T g_{i+1}.$$

Odtud plyne, že ve vzorci (128) můžeme použít tři různé jmenovatele a tři různé čitatele, aniž bychom porušili platnost věty 40. Dostaneme tak devět základních metod sdružených gradientů.

$$\beta_i^{HS} = \frac{y_i^T g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PR} = \frac{y_i^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{LS} = \frac{y_i^T g_{i+1}}{|g_i^T s_i|} \tag{133}$$

(HS – Hestenes a Stiefel [81], PR – Polak, Ribiére [134] a Polyak [135], LS – Liu a Storey [90]),

$$\beta_i^{DY} = \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \quad \beta_i^{FR} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{CD} = \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|} \tag{134}$$

(DY – Dai a Yuan [34], FR – Fletcher a Reeves [57], CD – conjugate descent [53]),

$$\beta_i^{HP} = \frac{(y_i - d_i)^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{PP} = \frac{(y_i - d_i)^T g_{i+1}}{g_i^T g_i}, \qquad \beta_i^{LP} = \frac{(y_i - d_i)^T g_{i+1}}{|g_i^T s_i|} \tag{135}$$

(Perryho verze metod HS, PR, LS [133], jejichž důmyslnější odvození je uvedeno v oddílu 3.4). Tyto metody můžeme rozdělit do tří skupin podle použitého čitatele. Metody první skupiny (HS, PR, LS) jsou výhodnější pro praktické použití, ale nejsou bez nutných úprav globálně konvergentní. Metody druhé skupiny (DY, FR, CD) jsou za určitých předpokladů (kladených na výběr délky kroku) globálně konvergentní, ale hůře zachovávají sdruženost směrových vektorů v případě, že nepoužíváme přesný výběr délky kroku a minimalizovaná funkce není kvadratická. Metody třetí skupiny (HP, PP, LP) mají podobné vlastnosti jako metody první skupiny. Metody patřící do téže skupiny se svými vlastnostmi příliš neliší až na to, že jmenovatel $g_i^T g_i$ není obecně souměřitelný s ostatními jmenovateli, což činí potíže v důkazech globální konvergence. Poznamenejme, že je-li splněna slabá Wolfeho podmínka, jsou hodnoty (134) kladné.

Poznámka 66. Nechť H je symetrická pozitivně definitní matice. Položme $\tilde{x} = H^{-1/2}x$ a $\tilde{F}(\tilde{x}) = F(x)$, takže $\tilde{g}(\tilde{x}) = H^{1/2}g(x)$ a $\tilde{G}(\tilde{x}) = H^{1/2}G(x)H^{1/2}$. Aplikujeme-li metodu sdružených gradientů na funkci $\tilde{F}(\tilde{x})$ a vrátíme-li se k původním proměnným, dostaneme

$$s_1 = -Hg_1 \quad \text{a} \quad s_{i+1} = -Hg_{i+1} + \beta_i s_i \quad \text{pro} \quad i \in N,$$

kde

$$\beta_i^{PHS} = \frac{y_i^T H g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PPR} = \frac{y_i^T H g_{i+1}}{g_i^T H g_i}, \quad \beta_i^{PLS} = \frac{y_i^T H g_{i+1}}{|g_i^T s_i|},$$

$$\beta_i^{PDY} = \frac{g_{i+1}^T H g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PFR} = \frac{g_{i+1}^T H g_{i+1}}{g_i^T H g_i}, \quad \beta_i^{PCD} = \frac{g_{i+1}^T H g_{i+1}}{|g_i^T s_i|},$$

$$\beta_i^{PHP} = \frac{(Hy_i - d_i)^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{PPP} = \frac{(Hy_i - d_i)^T g_{i+1}}{g_i^T H g_i}, \qquad \beta_i^{PLP} = \frac{(Hy_i - d_i)^T g_{i+1}}{|g_i^T s_i|},$$

Metody, které používají tyto vzorce se nazývají předpodmíněnými metodami sdružených gradientů. Pro tyto metody platí všechny věty, které jsme zatím dokázali (splňuje-li funkce $F(x)$ předpoklady F1–F3, případně F4–F6, splňuje tyto předpoklady i funkce $\tilde{F}(\tilde{x})$). Je však třeba psát $\tilde{g} = H^{1/2}g$ místo g a $\tilde{s} = H^{-1/2}s$ místo s, takže vzorce (129), (131), (132) zůstanou beze změny, ale pro $1 < i \leq n$ místo (130) platí

$$g_j^T H g_{i+1} = 0, \qquad 1 \leq j < i.$$

Poznámka 67. Podobný postup lze použít k odvození škálovaných metod sdružených gradientů. Nahradíme-li v předchozích vzorcích matici H maticí $\gamma_i I$, dostaneme

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -\gamma_i(g_{i+1} - \beta_i s_i) \quad \text{pro} \quad i \in N,$$

kde

$$\beta_i^{SHS} = \frac{y_i^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{SPR} = \frac{y_i^T g_{i+1}}{\gamma_{i-1} g_i^T g_i}, \qquad \beta_i^{SLS} = \frac{y_i^T g_{i+1}}{|g_i^T s_i|},$$

$$\beta_i^{SDY} = \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{SFR} = \frac{g_{i+1}^T g_{i+1}}{\gamma_{i-1} g_i^T g_i}, \qquad \beta_i^{SCD} = \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|},$$

$$\beta_i^{SHP} = \frac{p_i^T g_{i+1}}{y_i^T s_i}, \qquad \beta_i^{SPP} = \frac{p_i^T g_{i+1}}{\gamma_{i-1} g_i^T g_i}, \qquad \beta_i^{SLP} = \frac{p_i^T g_{i+1}}{|g_i^T s_i|}.$$

kde $p_i = y_i - d_i/\gamma_i$. Vzorce ve jmenovatelích plynou z toho, že v případě přesného výběry délky kroku platí

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i = -g_i^T s_i = \gamma_{i-1} g_i^T (g_i - \beta_{i-1} g_i^T s_{i-1}) = \gamma_{i-1} g_i^T g_i.$$

Poznámka 68. Abychom dokázali globální konvergenci škálovaných metod sdružených gradientů, předpokládáme existenci čísel $0 < \underline{\gamma} \leq \overline{\gamma}$ takových, že $\underline{\gamma} \leq \gamma_i \leq \overline{\gamma}$. Obvykle tato čísla zvolíme předem a pokud hodnota škálovacího paranetru uvedené nerovnosti nesplňuje, pokládáme $\gamma_i = 1$. Pro škálování se nejčastěji používají hodnoty

$$\gamma_i = \frac{y_i^T d_i}{y_i^T y_i}, \qquad \gamma_i = \frac{d_i^T d_i}{y_i^T d_i}, \qquad \gamma_i = \sqrt{\frac{d_i^T d_i}{y_i^T y_i}},$$

(poznámka 156). Takto škálované metody se nazývají spektrálními metodami sdružených gradientů [6]. Jelikož $y_i = g(x_{i+1}) - g(x_i)$ a $d_i = x_{i+1} - x_i$, platí $y_i = \tilde{G}_i d_i$ a $d_i = \tilde{G}_i^{-1} y_i$, kde

$$\tilde{G}_i = \int_0^1 G(x_i + \lambda d_i) \mathrm{d}\lambda. \tag{136}$$

Jsou-li splněny předpoklady F4 a F5, můžeme psát

$$\underline{G}\|y_i\|^2 \leq y_i^T d_i \leq \overline{G}\|y_i\|^2,$$
$$\frac{1}{\overline{G}}\|d_i\|^2 \leq y_i^T d_i \leq \frac{1}{\underline{G}}\|d_i\|^2,$$

takže spektrální škálovací parametry vyhovují nerovnostem $\underline{G} \leq \gamma_i \leq \overline{G}$.

3.2 Globální konvergence základních metod sdružených gradientů

Při vyšetřování globální konvergence metod sdružených gradientů lze místo podmínky (38) použít podmínku (37) (neboli $\sum_{i=1}^{\infty} 1/\|s_i\|^2 = \infty$). Platí tato obecná věta [29].

Věta 41. *Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda sdružených gradientů (127), která generuje směrové vektory splňující podmínku (S1a) a používá délky kroku splňující zobecněnou Wolfeho podmínku je globálně konvergentní, platí-li (37).*

Důkaz Vztah (127) můžeme zapsat ve tvaru $s_{i+1} + g_{i+1} = \beta_i s_i$, což po umocnění dává

$$\|s_{i+1}\|^2 + 2g_{i+1}^T s_{i+1} + \|g_{i+1}\|^2 = \beta_i^2 \|s_i\|^2$$

a jelikož podle (S1a) platí $g_{i+1}^T s_{i+1} < 0$, dostaneme

$$\|s_{i+1}\|^2 \geq \beta_i^2 \|s_i\|^2 - \|g_{i+1}\|^2. \tag{137}$$

Použijeme-li znovu (127) můžeme psát

$$g_{i+1}^T s_{i+1} - \beta_i g_{i+1}^T s_i = -\|g_{i+1}\|^2,$$

což spolu s (S2a), kde $\varepsilon_3 \geq \varepsilon_2$, dává

$$|g_{i+1}^T s_{i+1}| + \varepsilon_3 |\beta_i| |g_i^T s_i| \geq \|g_{i+1}\|^2$$

a po umocnění dostaneme

$$\begin{aligned} \|g_{i+1}\|^4 &\leq (g_{i+1}^T s_{i+1})^2 + 2\varepsilon_3 |\beta_i| |g_{i+1}^T s_{i+1}| |g_i^T s_i| + \varepsilon_3^2 \beta_i^2 (g_i^T s_i)^2 \\ &\leq (g_{i+1}^T s_{i+1})^2 + \beta_i^2 (g_i^T s_i)^2 + \varepsilon_3^2 (g_{i+1}^T s_{i+1})^2 + \varepsilon_3^2 \beta_i^2 (g_i^T s_i)^2, \end{aligned}$$

neboť pro libovolná dvě reálná čísla a a b platí $a^2 + b^2 \geq 2ab$. Poslední nerovnost převedeme na tvar

$$(g_{i+1}^T s_{i+1})^2 + \beta_i^2 (g_i^T s_i)^2 \geq \frac{1}{1 + \varepsilon_3^2} \|g_{i+1}\|^4,$$

což spolu s (137) dává

$$\begin{aligned}\frac{(g_{i+1}^T s_{i+1})^2}{\|s_{i+1}\|^2} + \frac{(g_i^T s_i)^2}{\|s_i\|^2} &= \frac{1}{\|s_{i+1}\|^2}\left((g_{i+1}^T s_{i+1})^2 + \frac{\|s_{i+1}\|^2}{\|s_i\|2}(g_i^T s_i)^2\right)\\ &\geq \frac{1}{\|s_{i+1}\|^2}\left((g_{i+1}^T s_{i+1})^2 + \beta_i^2(g_i^T s_i)^2 - \frac{(g_i^T s_i)^2}{\|s_i\|^2}\|g_{i+1}\|^2\right)\\ &\geq \frac{1}{\|s_{i+1}\|^2}\left(\frac{1}{1+\varepsilon_3^2}\|g_{i+1}\|^4 - \frac{(g_i^T s_i)^2}{\|s_i\|^2}\|g_{i+1}\|^2\right)\\ &= \frac{\|g_{i+1}\|^2}{\|s_{i+1}\|^2}\left(\frac{1}{1+\varepsilon_3^2}\|g_{i+1}\|^2 - \frac{(g_i^T s_i)^2}{\|s_i\|^2}\right).\end{aligned}$$

Předpokládejme, že platí (37) a neplatí $\liminf_{i\to\infty}\|g_i\| = 0$. Pak existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g_i\| \geq \underline{\varepsilon}$, $i \in N$, a předchozí nerovnost lze zapsat ve tvaru

$$\frac{(g_{i+1}^T s_{i+1})^2}{\|s_{i+1}\|^2} + \frac{(g_i^T s_i)^2}{\|s_i\|^2} \geq \frac{\underline{\varepsilon}^2}{\|s_{i+1}\|^2}\left(\frac{\underline{\varepsilon}^2}{1+\varepsilon_3^2} - \frac{(g_i^T s_i)^2}{\|s_i\|^2}\right).$$

Protože podle (33) platí $(g_i^T s_i)^2/\|s_i\|^2 \to 0$, existuje index $k \in N$ takový, že $(g_i^T s_i)^2/\|s_i\|^2 \leq \underline{\varepsilon}^2/(2+2\varepsilon_3^2)$ což spolu s předchozí nerovností a (33) dává

$$\frac{\underline{\varepsilon}^4}{2(1+\varepsilon_3^2)}\sum_{i=k}^{\infty}\frac{1}{\|s_{i+1}\|^2} < \infty.$$

To je však ve sporu s předpokladem (37). □

Poznámka 69. Věta 41 je velmi obecná, ale vyžaduje platnost podmínky (S1a), kterou metody sdružených gradientů obecně nesplňují. Vztah (S1a) se obvykle dokazuje tak, že se dokáže silnější nerovnost (34). Pak, důkaz toho, že (37) implikuje globální konvergenci je triviální (poznámka 28). Platí-li navíc (36), plyne globální konvergence z poznámky 27 a není třeba používat větu 41.

Použití věty 41 pro konkrétní metody sdružených gradientů není zcela přímočaré a je třeba vyšetřovat jednotlivé případy samostatně. Jak již bylo poznamenáno (poznámka 65), jsou metody (134) za určitých předpokladů (kladených na výběr délky kroku) globálně konvergentní bez jakýchkoliv úprav. Nejprve dokážeme globální konvergenci metody DY. Větu zformulujeme tak, aby zahrnovala poněkud širší třídu metod sdružených gradientů.

Věta 42. *(Globální konvergence metody DY [34]). Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda sdružených gradientů (127) s výběrem délky kroku splňujícím slabou Wolfeho podmínku je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \beta_i^{DY} = \lambda_i \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \qquad -\frac{1-\varepsilon_2}{1+\varepsilon_2} \leq \lambda_i \leq 1, \qquad i \in N. \tag{138}$$

Důkaz (a) Dokážeme nejprve, že

$$|\beta_i| \leq \frac{g_{i+1}^T s_{i+1}}{g_i^T s_i} \tag{139}$$

$\forall i \in N$. Použijeme-li (127) a vztah $y_i = g_{i+1} - g_i$, můžeme psát

$$\begin{aligned} g_{i+1}^T s_{i+1} &= -g_{i+1}^T g_{i+1} + \beta_i g_{i+1}^T s_i\\ &= \frac{-g_{i+1}^T g_{i+1}(g_{i+1}-g_i)^T s_i + \lambda_i g_{i+1}^T g_{i+1} g_{i+1}^T s_i}{y_i^T s_i}\\ &= -(1-\lambda_i)\frac{g_{i+1}^T g_{i+1} g_{i+1}^T s_i}{y_i^T s_i} + \frac{g_{i+1}^T g_{i+1} g_i^T s_i}{y_i^T s_i}, \end{aligned}$$

což s použitím (S3a), kde $\varepsilon_3 = \infty$, dává

$$\begin{aligned}\frac{g_{i+1}^T s_{i+1}}{g_i^T s_i} &= |\lambda_i| \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i} + (1 - |\lambda_i|) \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i} - (1 - \lambda_i) \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i} \frac{g_{i+1}^T s_i}{g_i^T s_i} \\ &\geq |\beta_i| + \left(1 - |\lambda_i| - (1 - \lambda_i)\varepsilon_2 \frac{g_i^T s_i}{g_i^T s_i}\right) \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i} \geq |\beta_i|,\end{aligned}$$

neboť $y_i^T s_i > 0$ a pro $-(1 - \varepsilon_2)/(1 + \varepsilon_2) \leq \lambda_i \leq 1$ platí $(1 - |\lambda_i| - (1 - \lambda_i)\varepsilon_2) \geq 0$. Z (139) plyne indukcí, že směrové vektory s_i, $i \in N$, jsou spádové (pokud gradienty g_i, $i \in N$, jsou nenulové). Platí totiž $g_1^T s_1 = -g_1^T g_1 < 0$ a předpokládáme-li, že $g_i^T s_i < 0$, dává (139) $g_{i+1}^T s_{i+1} \leq |\beta_i| g_i^T s_i < 0$, pokud $\beta_i \neq 0$. Jestliže $\beta_i = 0$, dostaneme podle (127) $g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1} < 0$.

(b) Zapíšeme-li (127) ve tvaru $s_{i+1} + g_{i+1} = \beta_i s_i$, dostaneme umocněním, převedením dvou členů na pravou stranu a použitím nerovnosti (139) vztah

$$\|s_{i+1}\|^2 = \beta_i^2 \|s_i\|^2 - 2 g_{i+1}^T s_{i+1} - \|g_{i+1}\|^2 \leq \left(\frac{g_{i+1}^T s_{i+1}}{g_i^T s_i}\right)^2 \|s_i\|^2 - 2 g_{i+1}^T s_{i+1} - \|g_{i+1}\|^2,$$

neboli

$$\begin{aligned}\frac{\|s_{i+1}\|^2}{(g_{i+1}^T s_{i+1})^2} &= \frac{\|s_i\|^2}{(g_i^T s_i)^2} - \frac{2}{g_{i+1}^T s_{i+1}} - \frac{\|g_{i+1}\|^2}{(g_{i+1}^T s_{i+1})^2} \\ &= \frac{\|s_i\|^2}{(g_i^T s_i)^2} - \left(\frac{1}{\|g_{i+1}\|} + \frac{\|g_{i+1}\|}{g_{i+1}^T s_{i+1}}\right)^2 + \frac{1}{\|g_{i+1}\|^2} \\ &\leq \frac{\|s_i\|^2}{(g_i^T s_i)^2} + \frac{1}{\|g_{i+1}\|^2}.\end{aligned}$$

Protože $\|s_1\|^2/(g_1^T s_1)^2 = 1/\|g_1\|^2$, dává předchozí nerovnost

$$\frac{\|s_i\|^2}{(g_i^T s_i)^2} \leq \sum_{j=1}^{i} \frac{1}{\|g_j\|^2} \quad \forall i \in N.$$

Předpokládejme, že neplatí $\liminf_{i \to \infty} \|g_i\| = 0$. Pak nutně existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$, takže

$$\frac{\|s_i\|^2}{(g_i^T s_i)^2} \leq \frac{i}{\underline{\varepsilon}^2} \quad \forall i \in N,$$

neboli

$$\sum_{i=1}^{\infty} \frac{(g_i^T s_i)^2}{\|s_i\|^2} \geq \sum_{i=1}^{\infty} \frac{\underline{\varepsilon}^2}{i} = \infty,$$

neboť harmonická řada je divergentní. To je však ve sporu s nerovností (33) uvedenou v poznámce 26. □

Nyní se budeme zabývat důkazem globální konvergence metody FR. Opět budeme vyšetřovat poněkud širší třídu metod sdružených gradientů.

Věta 43. *(Globální konvergence metody FR [1]). Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda sdružených gradientů (127) s výběrem délky kroku splňujícím silnou Wolfeho podmínku je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \beta_i^{FR} = \lambda_i \frac{\|g_{i+1}\|^2}{\|g_i\|^2}, \qquad |\lambda_i| \leq 1, \qquad i \in N. \tag{140}$$

Důkaz (a) (Al-Baali) Dokážeme indukcí nerovnost

$$0 < 1 - \frac{\varepsilon_2}{1-\varepsilon_2} \leq -\frac{g_i^T s_i}{\|g_i\|^2} \leq 1 + \frac{\varepsilon_2}{1-\varepsilon_2}. \tag{141}$$

Pro $i = 1$ nerovnost platí, neboť $s_1 = -g_1$ a tedy $-g_1^T s_1/\|g_1\|^2 = 1$. Předpokládejme, že nerovnost platí pro nějaký index $i \in N$. Zapíšeme-li (127) ve tvaru $s_{i+1} + g_{i+1} = \beta_i s_i$, můžeme psát

$$\frac{g_{i+1}^T s_{i+1}}{\|g_{i+1}\|^2} + 1 = \beta_i \frac{g_{i+1}^T s_i}{\|g_{i+1}\|^2} = \lambda_i \frac{g_{i+1}^T s_i}{\|g_i\|^2}.$$

Podle (S3a) platí $|g_{i+1}^T s_i| \leq -\varepsilon_2 g_i^T s_i$ a z indukčního předpokladu (pravá část nerovnosti) plyne $-g_i^T s_i/\|g_i\|^2 \leq 1 + \varepsilon_2/(1-\varepsilon_2)$. Použijeme-li tyto vztahy spolu s předchozí rovností, dostaneme

$$\left| \frac{g_{i+1}^T s_{i+1}}{\|g_{i+1}\|^2} + 1 \right| \leq -\varepsilon_2 |\lambda_i| \frac{g_i^T s_i}{\|g_i\|^2} \leq -\varepsilon_2 \frac{g_i^T s_i}{\|g_i\|^2} \leq \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) = \frac{\varepsilon_2}{1-\varepsilon_2}$$

(první nerovnost plyne z (S3a), druhá z toho, že $|\lambda_i| \leq 1$ a třetí z indukčního předpokladu). Tím je indukční krok dokončen (stačí odstranit absolutní hodnotu). Snadno se přesvědčíme, že platí $1 - \varepsilon_2/(1-\varepsilon_2) > 0$, pokud $0 < \varepsilon_2 < 1/2$, takže směrové vektory s_i, $i \in N$, jsou spádové a platí (34) s $\underline{s} = (1-2\varepsilon_2)/(1-\varepsilon_2)$, což podle poznámky 27 implikuje nerovnost (35).

(b) Použijeme-li levou část podmínky (S3a) a levou část nerovnosti (141), dostaneme

$$|s_i^T g_{i+1}| \leq -\varepsilon_2 s_i^T g_i \leq \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) \|g_i\|^2 = \frac{\varepsilon_2}{1-\varepsilon_2} \|g_i\|^2.$$

Použijeme-li tuto nerovnost spolu s (127), můžeme psát

$$\begin{aligned}
\|s_{i+1}\|^2 &\leq \|g_{i+1}\|^2 + 2|\beta_i||s_i^T g_{i+1}| + \beta_i^2 \|s_i\|^2 \\
&\leq \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2} |\beta_i| \|g_i\|^2 + \beta_i^2 \|s_i\|^2 \\
&= \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2} |\lambda_i| \|g_{i+1}\|^2 + \lambda_i^2 \frac{\|g_{i+1}\|^4}{\|g_i\|^4} \|s_i\|^2 \\
&\leq \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2} \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{\|g_i\|^4} \|s_i\|^2 \\
&= \frac{1+\varepsilon_2}{1-\varepsilon_2} \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{\|g_i\|^4} \|s_i\|^2,
\end{aligned}$$

neboť $|\lambda_i| \leq 1$. Předpokládejme, že neplatí

$$\liminf_{i \to \infty} \|g_i\| = 0.$$

Pak existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$, takže z předchozí nerovnosti plyne

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \frac{1}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \frac{i+1}{\underline{\varepsilon}^2}$$

(předpokládáme bez újmy na obecnosti, že $\underline{\varepsilon}^2 \|s_1\|^2/\|g_1\|^4 \leq (1+\varepsilon_2)/(1-\varepsilon_2)$). Můžeme tedy psát

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} \geq \frac{1-\varepsilon_2}{1+\varepsilon_2} \underline{\varepsilon}^2 \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

což je ve sporu s nerovností (35) uvedenou v poznámce 27. □

Poznámka 70. Věta 43 vyžaduje silnější předpoklady než věta 42. Je třeba, aby byla splněna silná Wolfeho podmínka a aby navíc platilo $\varepsilon_2 < 1/2$. Samotnou nerovnost (141) však můžeme zobecnit tak, že ji lze použít i za poněkud slabších předpokladů. Jestliže $|\lambda_i| \leq \bar{\varepsilon}_2/\varepsilon_2$, kde $0 < \varepsilon_2 < \bar{\varepsilon}_2 < 1/2$, postupem použitým v části (a) důkazu věty 43 dostaneme

$$0 < 1 - \frac{\bar{\varepsilon}_2}{1-\bar{\varepsilon}_2} \leq -\frac{g_i^T s_i}{\|g_i\|^2} \leq 1 + \frac{\bar{\varepsilon}_2}{1-\bar{\varepsilon}_2}. \tag{142}$$

Pokud $\varepsilon_2 \approx 1/10$ (což je doporučená hodnota) a $\bar{\varepsilon}_2 \approx 1/2$, platí tato nerovnost i pro $|\lambda_i| \approx 5$ (λ_i je koeficient v (140)). V důkazu globální konvergence však nerovnost $|\lambda_i| \leq \bar{\varepsilon}_2/\varepsilon_2$ použít nelze. Chceme-li připustit hodnoty $|\lambda_i| > 1$, je třeba pro $i \in N$ spolu s $|\lambda_i| \leq \bar{\varepsilon}_2/\varepsilon_2$ splnit podmínku

$$\|g_i\|^2 \sum_{j=1}^{i-1} \prod_{k=j}^{i-1} \lambda_k^2 \leq \bar{c}\, i$$

pro nějakou hodnotu $\bar{c} > 0$ [28].

V předpokladech věty 43 můžeme silnou Wolfeho podmínku nahradit zobecněnou Wolfeho podmínkou s $0 \leq \varepsilon_3 < 1/2$. Pro metodu FR to nemá žádný praktický význam. Zobecněná Wolfeho podmínka s $\varepsilon_3 = 0$ je však podstatná pro důkaz globální konvergence metody CD.

Věta 44. *(Globální konvergence metody CD [33]). Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak metoda sdružených gradientů (127) s výběrem délky kroku splňujícím zobecněnou Wolfeho podmínku s $\varepsilon_3 = 0$, je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \beta_i^{CD} = \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|}, \qquad 0 \leq \lambda_i \leq 1, \qquad i \in N. \tag{143}$$

Důkaz (a) Použijeme-li (127), (143) a (S3a) s $\varepsilon_3 = 0$ (takže $g_{i+1}^T s_i \leq 0$), dostaneme

$$-g_{i+1}^T s_{i+1} = g_{i+1}^T g_{i+1} - \lambda_i \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|} g_{i+1}^T s_i \geq g_{i+1}^T g_{i+1}, \tag{144}$$

takže směrové vektory s_i, $i \in N$, jsou spádové a platí (34) s $\underline{s} = 1$, což podle poznámky 27 implikuje nerovnost (35).

(b) Použijeme-li vztahy (128), (143) a podmínku $0 \leq \lambda_i \leq 1$, dostaneme

$$\begin{aligned}
\|s_{i+1}\|^2 &= \left(-g_{i+1} + \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right)^T \left(-g_{i+1} + \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right) \\
&= \|g_{i+1}\|^2 - 2\lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} g_{i+1}^T s_i + \lambda_i^2 \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2 \\
&\leq \|g_{i+1}\|^2 + 2\varepsilon_2 \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2
\end{aligned}$$

a protože podle (144) platí $|g_i^T s_i| \geq \|g_i\|^2$, dostaneme

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{1 + 2\varepsilon_2}{\|g_{i+1}\|^2} + \frac{\|s_i\|^2}{\|g_i\|^4}. \tag{145}$$

Předpokládejme, že neplatí

$$\liminf_{i \to \infty} \|g_i\| = 0.$$

Pak existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$, takže z předchozí nerovnosti plyne

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{1+2\varepsilon_2}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \leq \frac{1+2\varepsilon_2}{\underline{\varepsilon}^2}(i+1)$$

(předpokládáme bez újmy na obecnosti, že $\underline{\varepsilon}^2\|s_1\|^2/\|g_1\|^4 \leq 1+2\varepsilon_2$). Můžeme tedy psát

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} \geq \frac{\underline{\varepsilon}^2}{1+2\varepsilon_2} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

což je ve sporu s nerovností (35) uvedenou v poznámce 27. □

Poznámka 71. Podmínka $\varepsilon_3 = 0$ v (S3a) je nutná. Pro libovolnou hodnotu $\varepsilon_3 > 0$ lze nalézt funkci $F : R^n \to R$ a počáteční bod $x_1 \in R^n$ tak, že metoda CD nekonverguje [33]. Potíž spočívá v tom, že metoda CD s $g_{i+1}^T s_i > 0$ sice splňuje nerovnost (34) s $\underline{s} = 1-\varepsilon_2$, ale druhý člen na pravé straně nerovnosti (145) je třeba vynásobit koeficientem $1/(1-\varepsilon_2) > 1$. Posloupnost $\|s_i\|^2/\|g_i\|^4$ pak pro $\|g_i\| \geq \underline{\varepsilon}$ roste rychleji než geometrická posloupnost s koeficientem větším než jedna, takže platí (35).

Poznámka 72. Jak již bylo zmíněno v poznámce 65, dávají metody (133) a (135) lepší praktické výsledky než metody (134). Vlastnosti metod (133) a (135) lze zlepšit tím, že vyloučíme záporné hodnoty, takže dostaneme

$$\beta_i^{HS+} = \max(0, \beta_i^{HS}), \quad \beta_i^{PR+} = \max(0, \beta_i^{PR}), \quad \beta_i^{LS+} = \max(0, \beta_i^{LS}). \tag{146}$$

$$\beta_i^{HP+} = \max(0, \beta_i^{HP}), \quad \beta_i^{PP+} = \max(0, \beta_i^{PP}), \quad \beta_i^{LP+} = \max(0, \beta_i^{LP}). \tag{147}$$

Teoretické důvody pro vyloučení záporných hodnot jsou vysvětleny v oddílu 3.5. Metody (133) a (135) lze též kombinovat s metodami (134) tak, že je použijeme pouze tehdy splňují-li předpoklady vět o globální konvergenci, tedy

$$-\frac{1-\varepsilon_2}{1+\varepsilon_2} \leq \frac{\beta_i^{HS}}{\beta_i^{DY}} \leq 1, \quad -1 \leq \frac{\beta_i^{PR}}{\beta_i^{FR}} \leq 1, \quad 0 \leq \frac{\beta_i^{LS}}{\beta_i^{CD}} \leq 1,$$

$$-\frac{1-\varepsilon_2}{1+\varepsilon_2} \leq \frac{\beta_i^{HP}}{\beta_i^{DY}} \leq 1, \quad -1 \leq \frac{\beta_i^{PP}}{\beta_i^{FR}} \leq 1, \quad 0 \leq \frac{\beta_i^{LP}}{\beta_i^{CD}} \leq 1$$

(je-li splněna slabá Wolfeho podmínka, jsou hodnoty (134) kladné). Nejsou-li tyto nerovnosti splněny použijeme odpovídající metodu z (134), tedy β_i^{DY}, β_i^{FR}, β_i^{CD} . Také lze položit

$$\begin{aligned}
\beta_i^{HSC+} &= \max(0, \min(\beta_i^{HS}, \beta_i^{DY})), & \beta_i^{HPC+} &= \max(0, \min(\beta_i^{HP}, \beta_i^{DY})), \\
\beta_i^{PRC+} &= \max(0, \min(\beta_i^{PR}, \beta_i^{FR})), & \beta_i^{PPC+} &= \max(0, \min(\beta_i^{PP}, \beta_i^{FR})), \\
\beta_i^{LSC+} &= \max(0, \min(\beta_i^{LS}, \beta_i^{CD})), & \beta_i^{LPC+} &= \max(0, \min(\beta_i^{LP}, \beta_i^{CD})).
\end{aligned} \tag{148}$$

Z předchozích vět je zřejmé, že kombinované metody jsou globálně konvergentní za stejných předpokladů jako metody DY, FR, CD.

Poznámka 73. Globální konvergenci metod (133) a (135) lze zajistit pomocí přerušování iteračního procesu. Pokud není splněna podmínka $-g_{i+1}^T s_{i+1} \geq \varepsilon_0 \|g_{i+1}\| \|s_{i+1}\|$, kde $\varepsilon_0 > 0$, pokládáme $s_{i+1} = -g_{i+1}$ (což odpovídá hodnotě $\beta_i = 0$). Podle poznámky 32 jsou takto upravené metody globálně konvergentní. Praktické zkušenosti ukazují, že zvolíme li číslo ε_0 dostatečně malé, dochází k přerušování iteračního procesu pouze výjimečně a takto upravené metody jsou velmi efektivní.

3.3 Asymptotická rychlost konvergence

Metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce po nejvýše n krocích (věta 40). Neplatí to však jestliže (a) výběr délky kroku není přesný, (b) funkce není kvadratická, (c) Hessova matice je špatně podmíněná a projevují se zaokrouhlovací chyby. Pak je třeba pokračovat ve výpočtu. Aby byly i nadále splněny předpoklady věty 40, je třeba iterační proces přerušit (položit $s_{n+1} = -g_{n+1}$). Přerušování iteračního procesu je též nutné pro vyšetřování asymptotické rychlosti konvergence. V dalších úvahách se budeme zabývat cyklicky přerušovanými metodami sdružených gradientů, pro které jsou splněny předpoklady věty 32.

Definice 33. *Řekneme, že základní optimalizační metoda je cyklicky přerušovanou metodou sdružených gradientů, jestliže platí $s_i = -g_i$ pro $i \in M$ a $s_i = -g_i + \beta_{i-1}s_{i-1}$ pro $i \notin M$, kde parametr β_i je určen některým ze vzorců (133)–(135) a $M = \{l \in N : l = nk + \underline{l}, k \geq 0\}$ a $\underline{l} \in N$.*

Nejprve ukážeme, že cyklicky přerušovaná metoda sdružených gradientů, kde parametr β_i se vybírá tak, aby byla splněna nerovnost (142), je metodou stejnoměrně spádových směrů gradientního typu.

Věta 45. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná cyklicky přerušovanou metodou sdružených gradientů s výběrem délky kroku splňujícím silnou Wolfeho podmínku, přičemž platí*

$$|\beta_i| \leq \frac{\overline{\varepsilon}_2}{\varepsilon_2}\frac{\|g_{i+1}\|^2}{\|g_i\|^2}, \tag{149}$$

kde $0 < \varepsilon_2 < \overline{\varepsilon}_2 < 1/2$. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F : \mathcal{D} \to R$ vyhovující předpokladům F4 a F5. Pak jsou směrové vektory s_i, $i \in N$, stejnoměrně spádové a platí $s_i \sim g_i$.

Důkaz Připomeňme, že je-li splněna podmínka (149), platí podle poznámky 70 nerovnost (142).
(a) Zřejmě $\|e_i\| = O(\|e_{i-1}\|)$ (poznámka 35) a $\|g_{i-1}\| \sim \|e_{i-1}\|$ (věta 5), takže $\|g_i\| = O(\|g_{i-1}\|)$. Existuje tedy konstanta $c > 0$ taková, že

$$\frac{\|g_i\|}{\|g_{i-1}\|} \leq c\frac{\varepsilon_2}{\overline{\varepsilon}_2} \quad \forall i \notin M.$$

Nechť $i \notin M$. Pak podle (149) platí

$$\|s_i\| \leq \|g_i\| + |\beta_{i-1}|\|s_{i-1}\| \leq \|g_i\| + \frac{\overline{\varepsilon}_2}{\varepsilon_2}\frac{\|g_i\|^2}{\|g_{i-1}\|^2}\|s_{i-1}\|,$$

takže

$$\frac{\|s_i\|}{\|g_i\|} \leq 1 + \frac{\overline{\varepsilon}_2}{\varepsilon_2}\frac{\|g_i\|}{\|g_{i-1}\|}\frac{\|s_{i-1}\|}{\|g_{i-1}\|} \leq 1 + c\frac{\|s_{i-1}\|}{\|g_{i-1}\|}.$$

Nechť $k = \sup\{j \in M, j \leq i\}$. Protože $s_k = -g_k$, platí $\|s_k\|/\|g_k\| = 1$, takže rekurentním použitím poslední nerovnosti dostaneme

$$\frac{\|s_i\|}{\|g_i\|} \leq \sum_{j=0}^{i-k} c^j \leq \sum_{j=0}^{n} c^j \triangleq \overline{c}.$$

(b) Použijeme-li nerovnost (142) (pravou část) dostaneme

$$-\frac{s_i^T g_i}{\|g_i\|^2} \geq 1 - \frac{\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2} = \frac{1-2\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2},$$

což spolu s (a) dává

$$-\frac{s_i^T g_i}{\|s_i\|\|g_i\|} = -\frac{s_i^T g_i}{\|g_i\|^2}\frac{\|g_i\|}{\|s_i\|} \geq -\frac{1}{\overline{c}}\frac{s_i^T g_i}{\|g_i\|^2} \geq \frac{1}{\overline{c}}\frac{1-2\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2} = \frac{\underline{c}}{\overline{c}},$$

kde $\underline{c} = (1-2\overline{\varepsilon}_2)/(1-\overline{\varepsilon}_2) > 0$, takže $-s_i^T g_i \geq \varepsilon_0\|s_i\|/\|g_i\|$, kde $\varepsilon_0 = \underline{c}/\overline{c} > 0$.

(c) Použitím nerovnosti (142) a Schwarzovy nerovnosti dostaneme

$$\|s_i\|\|g_i\| \geq -s_i^T g_i \geq \frac{1-2\bar{\varepsilon}_2}{1-\bar{\varepsilon}_2}\|g_i\|^2,$$

což dává $\|s_i\| \geq \underline{c}\|g_i\|$. Jelikož z (a) plyne $\|s_i\| \leq \bar{c}\|g_i\|$, platí $s_i \sim g_i$. □

Nyní budeme vyšetřovat cyklicky přerušované metody sdružených gradientů s asymptoticky přesným výběrem délky kroku takové, že

$$\beta_i = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}(1+o(1)) \tag{150}$$

(pokud $\beta_i \neq 0$). Z (150) plyne existence indexu $\underline{l} \in M$ takového, že nerovnost (149) platí pro $i \geq \underline{l}$. Protože vyšetřujeme asymptotické chování iteračního procesu, budeme pro jednoduchost předpokládat, že $\underline{l} = 1$ (v opačném případě lze posunout indexy, aniž by se změnilo asymptotické chování uvažované posloupnosti). Pak jsou splněny předpoklady věty 45 a uvažovaná metoda je metodou stejnoměrně spádových směrů gradientního typu.

V dalším výkladu budeme předpokládat, že $e_i \neq 0$ a $g_i \neq 0$, $i \in N$, neboť v opačném případě iterační proces končí ve stacionárním bodě. Dále budeme předpokládat, že

$$\|e_i\| \sim \|e_l\| \quad \forall l \leq i < l+n, \quad l \in M.$$

Pokud pro nějaký index $l \leq i < l+n$ neplatí $\|e_i\| \sim \|e_l\|$, pak nutně $\|e_i\| = o(\|e_l\|)$ (jelikož podle poznámky 35 je $\|e_i\| = O(\|e_l\|)$), takže rychlost konvergence je vyšší než lineární (tato úvaha je precizována v důkazu věty 32).

Věta 46. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná cyklicky přerušovanou metodou sdružených gradientů s asymptoticky přesným výběrem délky kroku, pro kterou platí (150). Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in C^2 : \mathcal{D} \to R$ vyhovující předpokladům F4 a F5. Nechť $\|e_i\| \sim \|e_l\|$ pro $l \in M$ a $l \leq i < l+n$. Pak jsou splněny podmínky (a)-(e) uvedené v poznámce 59.*

Důkaz Podle předpokladu platí $e_i \neq 0$, $g_i \neq 0$ a $e_i \sim e_l$ pro $l \in M$ a $l \leq i < l+n$ a z předpokladů F4, F5 a věty 5 plyne, že $g_i \sim e_i$ pro $l \in M$ a $l \leq i < l+n$. Důkaz provedeme indukcí. Ukážeme, že pro $l \in M$ a $l \leq i < l+n$ platí

$$s_i \sim g_i, \quad \alpha_i \sim 1, \quad \beta_i \sim 1, \tag{151}$$

$$-s_i^T g_i = g_i^T g_i(1+o(1)) \tag{152}$$

(ze vztahů (151) a (152) plyne platnost podmínek (a) a (b) uvedených v poznámce 59), a že pro $l \in M$ a $l \leq j < i < l+n$ platí

$$s_j^T g_i = o(\|e_l\|^2), \tag{153}$$

$$g_j^T g_i = o(\|e_l\|^2), \tag{154}$$

$$s_j^T G^* s_i = o(\|e_l\|^2), \tag{155}$$

$$s_j^T y_i = y_j^T s_i = o(\|e_l\|^2), \tag{156}$$

(z (156) plyne platnost podmínky (c) uvedené v poznámce 59, podmínka (d) je předpokladem dokazované věty a podmínka (e) plyne z věty 40). Na začátku cyklu platí $s_l = -g_l \sim g_l$ a $-s_l^T g_l = g_l^T g_l = g_l^T g_l(1+o(1))$. Z předpokladů F4, F5 a lemmatu 7 plyne, že $\alpha_l \sim 1$. Dále není co dokazovat. Nechť (151)–(156) platí pro $l \leq i < l+n-1$ (indukční předpoklad).

(a) Důkaz vztahu (153). Podle (153) a (156) platí

$$s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = o(\|e_l\|^2)$$

pro $l \leq j < i$. Z definice 19 a z (152) plyne, že

$$s_i^T g_{i+1} = s_i^T g_i o(1) = o(\|g_i\|^2) = o(\|e_i\|^2) = o(\|e_l\|^2).$$

Platí tedy $s_j^T g_{i+1} = o(\|e_l\|^2)$ pro $l \le j \le i$.

(b) Důkaz vztahu (154). Zřejmě

$$\begin{aligned} g_l &= -s_l, \\ g_j &= -s_j + \beta_{j-1} s_{j-1} \qquad \forall l < j \le i, \end{aligned}$$

takže podle (a) a (150) platí

$$\begin{aligned} g_l^T g_{i+1} &= -s_l^T g_{i+1} = o(\|e_l\|^2), \\ g_j^T g_{i+1} &= -s_j^T g_{i+1} + \beta_{j-1} s_{j-1}^T g_{i+1} = o(\|e_l\|^2) \qquad \forall l < j \le i. \end{aligned}$$

(c) Důkaz vztahů (151). Z relací $g_{i+1} \sim e_{i+1} \sim e_l$ a $g_i \sim e_i \sim e_l$ plyne $g_{i+1} \sim g_i$, takže podle (150) platí $\beta_i \sim 1$. Z definice 19 a z (151)–(152) pak dostaneme

$$\begin{aligned} s_{i+1}^T s_{i+1} &= (-g_{i+1} + \beta_i s_i)^T (-g_{i+1} + \beta_i s_i) = g_{i+1}^T g_{i+1} - 2\beta_i g_{i+1}^T s_i + \beta_i^2 s_i^T s_i \\ &\sim g_{i+1}^T g_{i+1} + g_i^T g_i o(1) + g_i^T g_i \sim g_{i+1}^T g_{i+1} \end{aligned}$$

(používáme relaci $g_{i+1} \sim g_i$). Z předpokladů F4, F5 a lemmatu 7 plyne, že $\alpha_{i+1} \sim \|g_{i+1}\|/\|s_{i+1}\|$, což spolu s $s_{i+1} \sim g_{i+1}$ dává $\alpha_{i+1} \sim 1$.

(d) Důkaz vztahu (152). Z definice 19 a vztahů (151), (152) plyne, že

$$-g_{i+1}^T s_{i+1} = g_{i+1}^T g_{i+1} - \beta_i g_{i+1}^T s_i = g_{i+1}^T g_{i+1} + g_i^T s_i o(1) = g_{i+1}^T g_{i+1} + o(\|g_i\|^2) = g_{i+1}^T g_{i+1}(1 + o(1))$$

(používáme relaci $g_{i+1} \sim g_i$).

(e) Důkaz vztahů (155) a (156). Použijeme-li (151), (156) a (b), dostaneme

$$y_j^T s_{i+1} = \beta_i y_j^T s_i - y_j^T g_{i+1} = o(\|e_l\|^2) + (g_j - g_{j+1})^T g_{i+1} = o(\|e_l\|^2)$$

pro $1 \le j < i$ a podle (133) platí

$$y_i^T s_{i+1} = -y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} y_i^T s_i = 0,$$

což dohromady dává $y_j^T s_{i+1} = o(\|e_l\|^2)$ pro $1 \le j \le i$. Použijeme-li větu 5, můžeme pro $1 \le j \le i$ psát

$$y_j = g_{j+1} - g_j = G^* d_j + o(\|d_j\|) = \alpha_j G^* s_j + \alpha_j o(\|s_j\|) = \alpha_j G^* s_j + o(\|e_l\|),$$

takže

$$s_j^T G^* s_{i+1} = \frac{1}{\alpha_j} y_j^T s_{i+1} + \frac{\|s_{i+1}\|}{\alpha_j} o(\|e_l\|) = o(\|e_l\|^2),$$

neboť podle (151) platí $\alpha_j \sim 1$ a podle (d) je $s_{i+1} \sim g_{i+1} \sim e_{i+1} \sim e_l$. Použijeme-li znovu větu 5, dostaneme

$$y_{i+1} = g_{i+2} - g_{i+1} = G^* d_{i+1} + o(\|d_{i+1}\|) = \alpha_{i+1} G^* s_{i+1} + \alpha_{i+1} o(\|s_{i+1}\|) = \alpha_{i+1} G^* s_{i+1} + o(\|e_l\|),$$

takže

$$s_j^T y_{i+1} = \alpha_{i+1} s_j^T G^* s_{i+1} + \|s_j\| o(\|e_l\|) = o(\|e_l\|^2),$$

neboť podle (151) platí $s_j \sim g_j \sim e_j \sim e_l$ a podle (d) je $\alpha_{i+1} \sim 1$. □

Poznámka 74. Podmínka (150) je zcela jistě splněna pro metodu FR. Platí však i pro jiné metody sdružených gradientů. Dokážeme její platnost pro metodu HS. Protože $g_{i+1} \sim e_{i+1} \sim e_l$, můžeme podle části (b) předchozího důkazu psát

$$y_i^T g_{i+1} = g_{i+1}^T g_{i+1} - g_i^T g_{i+1} = g_{i+1}^T g_{i+1} + o(\|e_l\|^2) = g_{i+1}^T g_{i+1} + o(\|g_{i+1}\|^2) = g_{i+1}^T g_{i+1}(1 + o(1)).$$

Z definice 19 a z (152) plyne, že $y_i^T s_i = -g_i^T s_i(1+o(1)) = g_i^T g_i(1+o(1))$. Po dosazení dostaneme

$$\beta_i = \frac{y_i^T g_{i+1}}{y_i^T s_i} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}\frac{(1+o(1))}{(1+o(1))} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}(1+o(1)).$$

Lemma 20. *Nechť jsou splněny předpoklady věty 46. Nechť $\bar{x}_i \in R^n$, $i \in N$, je posloupnost získaná referenčním iteračním procesem uvedeným v poznámce 60. Pak $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \in M$ a $l \le i \le l+n$.*

Důkaz Dokážeme indukcí, že pro $l \in M$ a $1 \le i < l+n$ platí (102), (103) a

$$\bar{\beta}_i = \beta_i(1+o(1)), \tag{157}$$

takže dokazované tvrzení plyne bezprostředně z lemmatu 19. Na začátku cyklu platí $\bar{e}_l = e_l$, $\bar{g}_l = g_l$, $\bar{s}_l = s_l$, takže jako v důkazu lemmatu 19 lze psát $\bar{e}_l = e_l(1+o(1))$, $\bar{g}_l = g_l(1+o(1))$, $\bar{s}_l = g_l(1+o(1))$ a $\bar{\alpha}_l = \alpha_l(1+o(1))$. Předpokládejme, že (102), (103) a (157) platí pro $l \le i < l+n-1$ (indukční předpoklad).

(a) Jelikož podle věty 46 platí $\alpha_i \sim 1$, $s_i \sim g_i \sim e_i \sim e_l$ a $s_{i+1} \sim g_{i+1} \sim e_{i+1} \sim e_l$, můžeme stejným postupem jako v částech (a) a (b) důkazu lemmatu 19 dokázat, že platí (103).

(b) Podle (a) a indukčních předpokladů platí

$$\bar{y}_i = \bar{g}_{i+1} - \bar{g}_i = g_{i+1}(1+o(1)) - g_i(1+o(1)) = y_i + o(\|e_l\|) = y_i(1+o(1)),$$

neboť z předpokladů F4 a F5 plyne

$$y_i = \int_0^1 G(x_i + td_i)d_i \mathrm{d}t \sim d_i = \alpha_i s_i \sim e_l.$$

Můžeme tedy psát

$$\bar{\beta}_i = \frac{\bar{y}_i^T \bar{g}_{i+1}}{\bar{s}_i^T \bar{y}_i} = \frac{y_i^T g_{i+1}(1+o(1))^2}{s_i^T y_i (1+o(1))^2} = \frac{y_i^T g_{i+1}}{s_i^T y_i}(1+o(1)) = \beta_i(1+o(1)).$$

(c) Podle (b) a indukčních předpokladů platí

$$\begin{aligned} \bar{s}_{i+1} &= -\bar{g}_{i+1} + \bar{\beta}_i \bar{s}_i = -g_{i+1}(1+o(1)) + \beta_i s_i (1+o(1))^2 \\ &= -g_{i+1} + \beta_i s_i + o(\|e_l\|) = s_{i+1}(1+o(1)). \end{aligned}$$

□

Věta 47. *(n-kroková superlineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná cyklicky přerušovanou metodou sdružených gradientů s asymptoticky přesným výběrem délky kroku. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in C^2 : \mathcal{D} \to R$ vyhovující předpokladům F4 a F5. Pak platí*

$$\lim_{l \overset{M}{\to} \infty} \frac{\|x_{l+n} - x^*\|}{\|x_l - x^*\|} = 0.$$

Důkaz Tvrzení věty 47 je bezprostředním důsledkem lemmatu 20 a věty 32. □

Poznámka 75. Podle věty 9 a poznámky 15 je cyklicky přerušovaná metoda sdružených gradientů s asymptoticky přesným výběrem délky kroku R-superlineárně konvergentní.

Nyní se budeme věnovat odhadu asymptotické rychlosti konvergence metody sdružených gradientů ve vnitřních krocích každého cyklu.

Lemma 21. *Nechť jsou splněny předpoklady věty 40. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \max_{1 \leq k \leq n} P_i^2(\lambda_k), \tag{158}$$

kde $P_i(\lambda)$ je libovolný polynom stupně i takový, že $P_i(0) = 1$, a λ_k, $1 \leq k \leq n$, jsou vlastní čísla matice G seřazená vzestupně.

Důkaz (a) Dokážeme indukcí, že pro $1 \leq j \leq i$ platí $g_j \in \mathcal{K}_j$ a $s_j \in \mathcal{K}_j$, kde

$$\mathcal{K}_j = \text{span}\{g_1, Gg_1, \ldots, G^{j-1}g_1\}$$

je Krylovův podprostor stupně j generovaný maticí G a vektorem g_1. Pro $j = 1$ je to zřejmé. Nechť tedy $g_{j-1} \in \mathcal{K}_{j-1}$ a $s_{j-1} \in \mathcal{K}_{j-1}$. Protože pro každou kvadratickou funkci $x_j = x_{j-1} + \alpha_{j-1}s_{j-1}$ implikuje $g_j = g_{j-1} + \alpha_{j-1}Gs_{j-1}$ a protože podle indukčního předpokladu platí

$$g_{j-1} \in \mathcal{K}_{j-1}, \quad Gs_{j-1} \in \text{span}(Gg_1, G^2g_1, \ldots, G^{j-1}g_1) \subset \mathcal{K}_j,$$

dostaneme $g_j \in \mathcal{K}_j$. Podle (127) lze psát $s_j = -g_j + \beta_{j-1}s_{j-1}$. Protože podle indukčního předpokladu platí $s_{j-1} \in \mathcal{K}_{j-1} \subset \mathcal{K}_j$ a jak jsme právě dokázali $g_j \in \mathcal{K}_j$, dostaneme $s_j \in \mathcal{K}_j$

(b) Podle (a) platí

$$\begin{aligned} x_{i+1} - x^* &= x_1 - x^* + \sum_{j=1}^{i} \alpha_j s_j = x_1 - x^* + p_{i-1}^*(G)g_1 \\ &= x_1 - x^* + p_{i-1}^*(G)G(x_1 - x^*) = (I + Gp_{i-1}^*(G))(x_1 - x^*), \end{aligned}$$

kde $p_{i-1}^*(\lambda)$ je nějaký polynom stupně $i-1$ proměnné λ a výraz $p_{i-1}^*(G)$ dostaneme dosazením matice G za λ (matice $p_{i-1}^*(G)$ a G komutují). Označme $P_i^*(\lambda) = 1 + \lambda p_{i-1}^*(\lambda)$, takže $P_i^*(\lambda)$ je polynom stupně i proměnné λ takový, že $P_i^*(0) = 1$. Jelikož podle poznámky 40 bod $x_{i+1} = x_1 + P_i^*(G)(x_1 - x^*)$ realizuje minimum ryze konvexní kvadratické funkce Q na $\mathcal{K}_i$, platí

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &= \frac{1}{2}(x_{i+1} - x^*)^T G(x_{i+1} - x^*) = \frac{1}{2}(x_1 - x^*)^T P_i^*(G)GP_i^*(G)(x_1 - x^*) \\ &\leq \frac{1}{2}(x_1 - x^*)^T P_i(G)GP_i(G)(x_1 - x^*) \end{aligned}$$

pro libovolný polynom P_i stupně i takový, že $P_i(0) = 1$. Nechť λ_k a v_k $1 \leq k \leq n$ jsou (nezáporná) vlastní čísla a (ortonormální) vlastní vektory matice G a nechť

$$x_1 - x^* = \sum_{k=1}^{n} \gamma_k v_k.$$

Pak

$$Q(x_1) - Q(x^*) = \frac{1}{2}(x_1 - x^*)^T G(x_1 - x^*) = \frac{1}{2}\left(\sum_{k=1}^{n} \gamma_k v_k\right)^T G\left(\sum_{k=1}^{n} \gamma_k v_k\right) = \frac{1}{2}\sum_{k=1}^{n} \gamma_k^2 \lambda_k$$

a

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &\leq \frac{1}{2}(x_1 - x^*)^T P_i(G)GP_i(G)(x_1 - x^*) \\ &= \frac{1}{2}\left(\sum_{k=1}^{n} P_i(\lambda_k)\gamma_k v_k\right)^T G\left(\sum_{k=1}^{n} P_i(\lambda_k)\gamma_k v_k\right) \\ &= \frac{1}{2}\sum_{k=1}^{n} P_i^2(\lambda_k)\gamma_k^2\lambda_k \leq \frac{1}{2}\max_{1 \leq k \leq n} P_i^2(\lambda_k)\sum_{k=1}^{n} \gamma_k^2\lambda_k. \end{aligned}$$

Po vydělení dostaneme (158). □

Věta 48. *Nechť jsou splněny předpoklady věty 40. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \left(\frac{\lambda_{m+1-i} - \lambda_1}{\lambda_{m+1-i} + \lambda_1}\right)^2, \tag{159}$$

kde λ_k, $1 \leq k \leq m$, jsou různá vlastní čísla matice G seřazená vzestupně.

Důkaz Podle lemmatu 21 platí

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \max_{1 \leq k \leq n} P_i^2(\lambda_k)$$

pro libovolný polynom $P_i(\lambda)$ stupně nanejvýš i takový, že $P_i(0) = 1$. Zvolíme polynom $P_i(\lambda)$ tak, aby měl kořeny $\lambda'_{m+1-i} = (\lambda_1 + \lambda_{m+1-i})/2$ a λ_{m+1-j}, $1 \leq j \leq i-1$. Tento polynom stupně i má i kladných reálných kořenů, takže jeho kořeny a stacionární body se střídají. Jelikož $P_i(0) = 1$, $P_i(\lambda'_{m+1-i}) = 0$ a $\lambda'_{m+1-i} > 0$, je tento polynom klesající a konvexní až do svého prvního minima a pak roste až do následujícího kořenu $\lambda_{m+2-i} > \lambda_{m+1-i}$. Lze tedy usoudit, že $P_i(\lambda)$ leží v intervalu $0 \leq \lambda \leq \lambda'_{m+1-i}$ pod a v intervalu $\lambda'_{m+1-i} \leq \lambda \leq \lambda_{m+2-i}$ nad přímkou spojující body $[0,1]^T$ a $[\lambda'_{m+1-i}, 0]^T$. Z toho důvodu pro $\lambda_1 \leq \lambda \leq \lambda_{m+1-i}$ platí

$$|P_i(\lambda)| \leq \left|1 - \frac{\lambda}{\lambda'_{m+1-i}}\right| = \left|1 - \frac{2\lambda}{\lambda_1 + \lambda_{m+1-i}}\right|. \tag{160}$$

Výraz na pravé straně (160) nabývá v intervalu $\lambda_1 \leq \lambda \leq \lambda_{m+1-i}$ maxima $|\lambda_{m+1-i} - \lambda_1|/|\lambda_{m+1-i} + \lambda_1|$ (pro $\lambda = \lambda_1$ a $\lambda = \lambda_{m+1-i}$). Platí tedy

$$\begin{aligned}
\max_{1 \leq k \leq m} P_i^2(\lambda_k) &= \max_{1 \leq k \leq m+1-i} P_i^2(\lambda_k) \leq \max_{\lambda_1 \leq \lambda \leq \lambda_{m+1-i}} P_i^2(\lambda) \\
&\leq \max_{\lambda_1 \leq \lambda \leq \lambda_{m+1-i}} \left(1 - \frac{2\lambda}{\lambda_1 + \lambda_{m+1-i}}\right)^2 = \left(\frac{\lambda_{m+1-i} - \lambda_1}{\lambda_{m+1-i} + \lambda_1}\right)^2,
\end{aligned}$$

což spolu s (158) dává (159). □

Důsledek 3. *Metoda sdružených gradientů s přesným výběrem délky kroku nalezne minimum ryze konvexní kvadratické funkce po nejvýše m krocích, kde m je počet různých vlastních čísel matice G.*

Důkaz Položíme-li v (159) $i = m$, dostaneme $Q(x_{m+1}) - Q(x^*) \leq 0$. jelikož $Q(x_{m+1}) - Q(x^*) \geq 0$, musí platit $Q(x_{m+1}) = Q(x^*)$, což pro ryze konvexní kvadratickou funkci znamená, že $x_{m+1} = x^*$. □

Věta 49. *Nechť jsou splněny předpoklady věty 40. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq 4\left(\frac{\sqrt{\kappa(G)} - 1}{\sqrt{\kappa(G)} + 1}\right)^{2i}. \tag{161}$$

Důkaz Podle lemmatu 21 platí

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \max_{1 \leq k \leq n} (P_i(\lambda_k))^2$$

pro libovolný polynom $P_i(\lambda)$ stupně nanejvýš i takový, že $P_i(0) = 1$. Zvolíme polynom $P_i(\lambda)$ tak, aby minimalizoval hodnotu

$$\max_{\lambda_1 \leq \lambda \leq \lambda_n} |P_i(\lambda)|.$$

Tuto vlastnost má Čebyševův polynom transformovaný na interval $\lambda_1 \leq \lambda \leq \lambda_n$ a normovaný tak, aby nabýval hodnoty 1 pro $\lambda = 0$, tedy polynom

$$P_i(\lambda) = \frac{T_i\left(\frac{\lambda_n+\lambda_1-2\lambda}{\lambda_n-\lambda_1}\right)}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)},$$

kde $T_i(\xi)$ je klasický Čebyševův polynom, pro který platí $|T_i(\xi)| \leq 1$, pokud $|\xi| \leq 1$, a

$$T_i(\xi) = \frac{1}{2}((\xi + \sqrt{\xi^2-1})^i + (\xi - \sqrt{\xi^2-1})^i),$$

pokud $|\xi| \geq 1$. Jelikož pro $\lambda_1 \leq \lambda \leq \lambda_2$ platí $|(\lambda_n + \lambda_1 - 2\lambda)/(\lambda_n - \lambda_1)| \leq 1$, můžeme psát

$$\max_{\lambda_1 \leq \lambda \leq \lambda_n} |P_i(\lambda)| \leq \frac{1}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)}.$$

Zbývá tedy vyčíslit hodnotu na pravé straně poslední nerovnosti. Označme $\xi = (\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)$. Zřejmě $|\xi| \geq 1$, takže

$$\begin{aligned} T_i(\xi) &= \frac{1}{2}((\xi + \sqrt{\xi^2-1})^i + (\xi - \sqrt{\xi^2-1})^i) \geq \frac{1}{2}(\xi + \sqrt{\xi^2-1})^i \\ &= \frac{1}{2}\frac{1}{2^i}(\sqrt{\xi+1} + \sqrt{\xi-1})^{2i}, \end{aligned}$$

neboť

$$(\sqrt{\xi+1} + \sqrt{\xi-1})^2 = 2(\xi + \sqrt{\xi^2-1}).$$

Dosadíme-li $\xi = (\lambda_n + \lambda_1)/(\lambda_n - \lambda_1)$, dostaneme

$$\begin{aligned} T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right) &\geq \frac{1}{2}\left(\sqrt{\frac{\lambda_n}{\lambda_n-\lambda_1}} + \sqrt{\frac{\lambda_1}{\lambda_n-\lambda_1}}\right)^{2i} = \frac{1}{2}\left(\frac{(\sqrt{\lambda_n}+\sqrt{\lambda_1})^2}{\lambda_n-\lambda_1}\right)^i \\ &= \frac{1}{2}\left(\frac{\sqrt{\lambda_n}+\sqrt{\lambda_1}}{\sqrt{\lambda_n}-\sqrt{\lambda_1}}\right)^i. \end{aligned}$$

Platí tedy

$$\begin{aligned} \frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} &\leq \left(\max_{1 \leq k \leq n} |P_i(\lambda_k)|\right)^2 \leq \left(\frac{1}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)}\right)^2 \\ &\leq 4\left(\frac{\sqrt{\lambda_n}-\sqrt{\lambda_1}}{\sqrt{\lambda_n}+\sqrt{\lambda_1}}\right)^{2i} = 4\left(\frac{\sqrt{\kappa(G)}-1}{\sqrt{\kappa(G)}+1}\right)^{2i}. \end{aligned}$$

□

Poznámka 76. Použijeme-li odhad (161) spolu s (42) a (43), dostaneme

$$\frac{\|x_{i+1} - x^*\|}{\|x_1 - x^*\|} \leq 2\sqrt{\kappa(G)}\left(\frac{\sqrt{\kappa(G)}-1}{\sqrt{\kappa(G)}+1}\right)^i$$

pro $1 \leq i \leq n$.

Poznámka 77. Větu 49 lze snadno zobecnit tak, aby platila pro předpodmíněnou metodu sdružených gradientů. Podle poznámky 66 stačí použít $\kappa(H^{1/2}GH^{1/2})$ místo $\kappa(G)$. Pokud $H \approx G^{-1}$, může být $\kappa(H^{1/2}GH^{1/2})$ mnohem menší než $\kappa(G)$, a konvergence se velmi urychlí.

Věta 50. *(Asymptotický odhad) Nechť jsou splněny předpoklady věty 46. Pak pro $l \in M$ a $l \leq i < l+n$ platí*

$$\frac{\|x_i - x^*\|}{\|x_l - x^*\|} \leq 2\sqrt{\kappa(G^*)} \left(\frac{\sqrt{\kappa(G^*)} - 1}{\sqrt{\kappa(G^*)} + 1} \right)^{i-l} + o(1),$$

takže posloupnost x_i, $l \leq i < l+n$, konverguje k bodu $x^* \in R^n$ (alespoň) lineárně s asymptotickou rychlostí

$$q = \frac{\sqrt{\kappa(G^*)} - 1}{\sqrt{\kappa(G^*)} + 1}.$$

Důkaz Zvolme $l \in M$ tak, aby pro $i \geq l$ docházelo k přerušování iteračního procesu vždy po n krocích. Nechť $M = 2\sqrt{\kappa(G^*)}$ a q je kvocient uvedený ve větě 50. Pak podle poznámky 76 pro $l \leq i \leq l+n$ platí

$$\|\bar{x}_i - x^*\| \leq Mq^{i-l}\|\bar{x}_l - x^*\| = Mq^{i-l}\|x_l - x^*\|.$$

Použijeme-li lemma **??**, můžeme pro $l \leq i \leq l+n$ psát

$$\|x_i - \bar{x}_i\| = o(\|x_l - x^*\|) = \|x_l - x^*\| o(1).$$

Platí tedy

$$\frac{\|x_i - x^*\|}{\|x_l - x^*\|} \leq \frac{\|\bar{x}_i - x^*\|}{\|x_l - x^*\|} + \frac{\|x_i - \bar{x}_i\|}{\|x_l - x^*\|} \leq Mq^{i-l} + o(1).$$

□

Poznámka 78. Věta 50 se týká pouze vnitřních iterací každého cyklu. Celkově je cyklicky přerušovaná metoda sdružených gradientů s asymptoticky přesným výběrem délky kroku R-superlineárně konvergentní (poznámka 75).

Poznámka 79. Odhad $(\sqrt{\kappa} - 1)/(\sqrt{\kappa} + 1)$ je mnohem příznivější než odhad $(\kappa - 1)/(\kappa + 1)$ platný pro metodu největšího spádu jak ukazuje tato tabulka, ve které je uveden počet iterací potřebný k dosažení požadované přesnosti ε.

Problém	SD	CG
$\kappa = 10^2$, $\varepsilon = 10^{-4}$	460	46
$\kappa = 10^4$, $\varepsilon = 10^{-6}$	69077	690
$\kappa = 10^6$, $\varepsilon = 10^{-8}$	9210340	9210

3.4 Spádové metody sdružených gradientů

Abychom zlepšili účinnost metod sdružených gradientů můžeme vztah (127) různě upravovat. Obvykle se to provádí tak, že se přidávají výrazy úměrné $s_i^T g_{i+1}$, které v případě přesného výběru délky kroku vymizí a tvrzení věty 40 zůstane zachováno. Snažíme se přitom, aby byla splněna podmínka (34).

Definice 34. *Metodu sdružených gredientů, která generuje směrové vektory $s_i \in R^n$, $i \in N$, splňující podmínku (34), neboli*

$$-s_i^T g_i \geq \underline{s}\|g_i\|^2, \qquad i \in N, \tag{162}$$

kde $\underline{s} > 0$, nazveme spádovou metodou sdružených gradientů. Je-li navíc splněna podmínka $\|s_i\| \leq \overline{s}\|g_i\|$ (takže podle poznámky 27 platí (S1b) s $\varepsilon_0 = \underline{s}/\overline{s}$), řekneme že jde o stejnoměrně spádovou metodu sdružených gredientů.

Jednou z možností jak splnit podmínku (162) je nahradit vektor s_i v (127) jeho průmětem do ortogonálního doplňku podprostoru generovaného gradientem g_{i+1}, tedy položit $s_1 = -g_1$ a

$$s_{i+1} = -g_{i+1} + \beta_i P_{i+1} s_i, \quad P_{i+1} = \left(I - \frac{g_{i+1} g_{i+1}^T}{g_{i+1}^T g_{i+1}} \right), \quad i \in N. \tag{163}$$

Pak platí

$$g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1}, \tag{164}$$

neboť $g_{i+1}^T P_{i+1} = 0$. Dostaneme tak metodu popsanou v [173].

Věta 51. *Uvažujme metodu sdružených gradientů danou předpisem*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -\left(1 + \beta_i \frac{g_{i+1}^T s_i}{g_{i+1}^T g_{i+1}} \right) g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{165}$$

kde β_i je některá z hodnot (133)–(135). Pak pro tuto metodu platí tvrzení věty 40 a je splněna rovnost (164).

Důkaz Provádíme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i = 0$. Vztah (165) tedy přejde na (127) a vlastnost kvadratického ukončení zůstane zachována. Vztah (165) je pouze jiným vyjádřením vztahu (163), takže platí (164). □

Dosadíme-li hodnotu β_i^{CD} do (165), dostaneme $s_{i+1} = -\vartheta_i^{CD} g_{i+1} + \beta_i^{CD} s_i$, kde $\vartheta_i^{CD} = y_i^T s_i / |g_i^T s_i|$ a $\beta_i^{CD} = g_{i+1}^T g_{i+1} / |g_i^T s_i|$. Podobným způsobem lze modifikovat i metodu FR, když ve jmenovateli obou vztahů nahradíme výraz $|g_i^T s_i|$ skalárním součinem $g_i^T g_i$. Tyto modifikace, uvedené v [174], dovolují značně oslabit podmínky pro globální konvergenci metod FR a CD.

Věta 52. *Uvažujme modifikované metody DY, FR, CD dané předpisem*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -\vartheta_i g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{166}$$

kde hodnoty β_i^{DY}, β_i^{FR}, β_i^{CD} jsou určeny podle (134) a

$$\vartheta_i^{DY} = \frac{y_i^T s_i}{y_i^T s_i} = 1, \quad \vartheta_i^{FR} = \frac{y_i^T s_i}{g_i^T g_i}, \quad \vartheta_i^{CD} = \frac{y_i^T s_i}{|g_i^T s_i|}. \tag{167}$$

Pak pro tyto modifikované metody platí tvrzení věty 40. Splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1 a F3 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, jsou tyto metody globálně konvergentní (v případě modifikované metody DY stačí použít slabou Wolfeho podmínku).

Důkaz Provádíme-li přesný výběr délky kroku, platí $y_i^T s_i = -g_i^T s_i = g_i^T g_i$ (poznámka 65), což spolu s (167) dává $\vartheta_i^{DY} = \vartheta_i^{FR} = \vartheta_i^{CD} = 1$. Vztah (166) tedy přejde na (127) a vlastnost kvadratického ukončení zůstane zachována. Nyní dokážeme globální konvergenci.

(a) Jelikož $\vartheta_i^{DY} = 1$, metoda DY se použitím (166) nezmění, takže globální konvergence plyne z věty 42.

(b) Pro modifikovanou metodu FR platí

$$g_{i+1}^T s_{i+1} = -y_i^T s_i \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} + \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} g_{i+1}^T s_i = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} g_i^T s_i < 0.$$

Jelikož $g_1^T s_1 = -g_1^T g_1$, postupným dosazováním do předchozího vztahu (indukcí) dostaneme rovnost (164). Modifikovaná metoda FR je tedy totožná s modifikovanou metodou CD a pro obě tyto metody je splněna rovnost (164).

(c) Uvažujme modifikovanou metodu CD. Jelikož je splněna rovnost (164), jsou směrové vektory s_i, $i \in N$, spádové a platí (162) s $\underline{s} = 1$, což podle poznámky 27 implikuje nerovnost (35). Protože při výběru délky

kroku používáme zobecněnou Wolfeho podmínku (kde bez újmy na obecnosti předpokládáme, že $\varepsilon_3 \geq \varepsilon_2$), platí

$$0 < y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i \leq \varepsilon_3 |g_i^T s_i| - g_i^T s_i = (1+\varepsilon_3)|g_i^T s_i|, \tag{168}$$

neboli $\vartheta_i \leq 1 + \varepsilon_3$. Použijeme-li tento odhad spolu se vztahy (164), (166), (167), můžeme psát

$$\begin{aligned}
\|s_{i+1}\|^2 &= \left(-\vartheta_i g_{i+1} + \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right)^T \left(-\vartheta_i g_{i+1} + \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right) \\
&= \vartheta_i^2 \|g_{i+1}\|^2 - 2\vartheta_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} g_{i+1}^T s_i + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2 \\
&\leq (1+\varepsilon_3)^2 \|g_{i+1}\|^2 + 2\varepsilon_2 (1+\varepsilon_3) \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2,
\end{aligned}$$

neboli

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{(1+\varepsilon_3)(1+2\varepsilon_2+\varepsilon_3)}{\|g_{i+1}\|^2} + \frac{\|s_i\|^2}{\|g_i\|^4}.$$

Předpokládejme, že neplatí

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Pak existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$, takže z předchozí nerovnosti plyne

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{(1+\varepsilon_3)(1+2\varepsilon_2+\varepsilon_3)}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \leq \frac{(1+\varepsilon_3)(1+2\varepsilon_2+\varepsilon_3)}{\underline{\varepsilon}^2}(i+1)$$

(předpokládáme bez újmy na obecnosti, že $\underline{\varepsilon}^2 \|s_1\|^2 / \|g_1\|^4 \leq (1+\varepsilon_3)(1+2\varepsilon_2+\varepsilon_3)$). Můžeme tedy psát

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} \geq \frac{\underline{\varepsilon}^2}{(1+\varepsilon_3)(1+2\varepsilon_2+\varepsilon_3)} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

což je ve sporu s nerovností (35) uvedenou v poznámce 27. □

Poznámka 80. Z věty 52 plyne, že modifikace (166) dovoluje značně oslabit podmínky pro globální konvergenci metod FR a CD. Stačí, vybíráme-li délku kroku pomocí zobecněné Wolfeho podmínky, kde $\varepsilon_3 \geq 0$ je libovolně velké, ale konečné číslo. Tato podmínka se příliš neliší od slabé Wolfeho podmínky, kde $\varepsilon_3 = \infty$.

Vztah (166) lze také použít ke zlepšení konjugovanosti směrových vektorů v metodách PR a LS.

Věta 53. *Uvažujme modifikované metody HS, PR, LS dané předpisem*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -\vartheta_i g_{i+1} + \beta_i s_i \quad pro \quad i \in N,$$

kde hodnoty β_i^{HS}, β_i^{PR}, β_i^{LS} jsou určeny podle (133) a

$$\vartheta_i^{HS} = \frac{y_i^T s_i}{y_i^T s_i} = 1, \quad \vartheta_i^{PR} = \frac{y_i^T s_i}{g_i^T g_i}, \quad \vartheta_i^{LS} = \frac{y_i^T s_i}{|g_i^T s_i|}. \tag{169}$$

Pak pro tyto modifikované metody platí tvrzení věty 40 a navíc

$$y_i^T s_{i+1} = 0 \quad pro \quad i \in N. \tag{170}$$

Důkaz Tak jako v důkazu věty 52 platí $\vartheta_i^{HS} = \vartheta_i^{PR} = \vartheta_i^{LS} = 1$, používáme-li přesný výběr déky kroku. Vztah (166) tedy přejde na (127) a vlastnost kvadratického ukončení zůstane zachována. Metoda HS, pro kterou platí (170), se nezmění. V případě metod PR a LS dostaneme

$$\begin{aligned} y_i^T s_{i+1} &= -\frac{y_i^T s_i}{g_i^T g_i} y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{g_i^T g_i} y_i^T s_i = 0, \\ y_i^T s_{i+1} &= -\frac{y_i^T s_i}{|g_i^T s_i|} y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{|g_i^T s_i|} y_i^T s_i = 0. \end{aligned}$$

□

Použitím vzorců (166) a (169) nelze zajistit spádovost směrových vektorů modifikovaných metod HS, PR, LS. To umožňuje obecný vztah (165), získaný použitím symetrické projekční matice. Můžeme však použít i nesymetrickou projekční matici, tedy položit $s_1 = -g_1$ a

$$s_{i+1} = -g_{i+1} + \beta_i \tilde{P}_{i+1} s_i, \quad \tilde{P}_{i+1} = \left(I - \frac{p_i g_{i+1}^T}{g_{i+1}^T p_i} \right), \quad i \in N, \tag{171}$$

kde p_i je libovolný vektor takový, že $g_{i+1}^T p_i \neq 0$. Pak platí (164), neboť $g_{i+1}^T \tilde{P}_{i+1} = 0$.

Věta 54. *Uvažujme metodu sdružených gradientů danou předpisem*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -g_{i+1} + \beta_i s_i - \zeta_i p_i \quad pro \quad i \in N, \tag{172}$$

kde β_i je některá z hodnot (133)–(135), $g_{i+1}^T p_i \neq 0$ a

$$\zeta_i = \beta_i \frac{g_{i+1}^T s_i}{g_{i+1}^T p_i}. \tag{173}$$

Pak pro tuto metodu platí tvrzení věty 40 a je splněna rovnost (164).

Důkaz Provádíme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i = 0$, takže $\zeta_i = 0$ podle (173). Vztah (172) tedy přejde na (127) a vlastnost kvadratického ukončení zůstane zachována. Vztah (172) s parametrem (173) je pouze jiným vyjádřením vztahu (171), takže platí (164). □

Zbývá ukázat, jak se volí vektor p_i. Volby $p_i = d_i$ a $p_i = g_i$ jsou nevhodné, neboť v případě přesného výběru déky kroku platí $g_{i+1}^T d_i = 0$ a podle věty 40 se snažíme o to, aby platilo $g_{i+1}^T g_i = 0$. Volba $p_i = g_{i+1}$ byla použita v (163). Pokud používáme parametry (133) a (134), je výhodné volit $p_i = y_i$ [175]. Pro metody HS, PR, LS pak dosazením (133) do (173) dostaneme

$$\zeta_i^{HS} = \frac{g_{i+1}^T s_i}{y_i^T s_i}, \quad \zeta_i^{PR} = \frac{g_{i+1}^T s_i}{g_i^T g_i}, \quad \zeta_i^{LS} = \frac{g_{i+1}^T s_i}{|g_i^T s_i|}. \tag{174}$$

Použijeme-li parametry (135) a položíme-li $p_i = y_i - d_i$, dostaneme stejné metody, jako když použijeme parametry (133) a položíme $p_i = y_i$.

Další spádové metody sdružených gradientů lze získat použitím následujícího lemmatu.

Lemma 22. *Nechť $s_+ = -\vartheta g_+ + \beta s$, kde $0 < \underline{\vartheta} \leq \vartheta \leq \overline{\vartheta}$ a*

$$\beta = g_+^T z - \frac{\lambda}{\vartheta} z^T z g_+^T s, \tag{175}$$

kde $z \in R^n$ je libovolný nenulový vektor a $1/4 < \underline{\lambda} \leq \lambda \leq \overline{\lambda}$. Pak platí

$$-g_+^T s_+ \geq \underline{s} \|g_+\|^2, \qquad \underline{s} = \underline{\vartheta} \left(1 - \frac{1}{4\underline{\lambda}} \right) > 0. \tag{176}$$

Důkaz Podle předpokladu platí

$$-g_+^T s_+ = \vartheta g_+^T g_+ - \beta g_+^T s = \vartheta g_+^T g_+ - g_+^T z g_+^T s + \frac{\lambda}{\vartheta} z^T z (g_+^T s)^2. \tag{177}$$

Dosadíme-li do vztahu

$$|u^T v| \le \|u\|\|v\| \le \frac{1}{2}\left(\|u\|^2 + \|v\|^2\right) \tag{178}$$

(první část plyne ze Schwarzovy nerovnosti a druhá z nerovnosti (19)) vektory

$$u = \sqrt{\frac{\vartheta}{2\lambda}} g_+, \qquad v = \sqrt{\frac{2\lambda}{\vartheta}} (g_+^T s) z,$$

dostaneme

$$|g_+^T s \, g_+^T z| \le \frac{1}{2}\left(\frac{\vartheta}{2\lambda} g_+^T g_+ + \frac{2\lambda}{\vartheta} z^T z (g_+^T s)^2\right),$$

což po dosazení do (177) dává

$$-g_+^T s_+ \ge \vartheta g_+^T g_+ - \frac{\vartheta}{4\lambda} g_+^T g_+ - \frac{\lambda}{\vartheta} z^T z (g_+^T s)^2 + \frac{\lambda}{\vartheta} z^T z (g_+^T s)^2 = \vartheta\left(1 - \frac{1}{4\lambda}\right)\|g_+\|^2 \ge \underline{s}\|g_+\|^2.$$

Jelikož $\underline{\vartheta} > 0$ a $\underline{\lambda} > 1/4$, platí $\underline{s} = \underline{\vartheta}(1 - 1/(4\underline{\lambda})) > 0$. □

Poznámka 81. Nahradíme-li ve vzorci $s_+ = -\vartheta g_+ + \beta s$ hodnotu (175) hodnotou $\beta = -(\lambda/\vartheta) z^T z g_+^T s$, dostaneme

$$-g_+^T s_+ = \vartheta g_+^T g_+ + \frac{\lambda}{\vartheta} z^T z (g_+^T s)^2 \ge \underline{\vartheta} g_+^T g_+,$$

takže odpovídající metoda sdružených gradientů je spádová bez ohledu na velikost parametru $\lambda > 0$. Z tohoto důvodu je někdy výhodnější používat místo (175) hodnotu $\beta = \max(0, g_+^T z) - (\lambda/\vartheta) z^T z g_+^T s$.

Poznámka 82. Provádíme-li přesný výběr délky kroku (takže $g_+^T s = 0$), druhý člen v (175) odpadne. Proto je výhodné volit vektor z tak, aby první člen odpovídal některé základní metodě sdružených gradientů (pak platí tvrzení věty 40). Nechť $\vartheta = 1$. Dosadíme-li do (175) po řadě $z = y/y^T s$, $z = y/g^T g$, $z = y/|g^T s|$, dostaneme spádové modifikace metod (133), pro které

$$\beta^{HSD} = \beta^{HS} - \lambda\frac{y^T y g_+^T s}{(y^T s)^2}, \qquad \beta^{PRD} = \beta^{PR} - \lambda\frac{y^T y g_+^T s}{(g^T g)^2}, \qquad \beta^{LSD} = \beta^{LS} - \lambda\frac{y^T y g_+^T s}{(g^T s)^2}. \tag{179}$$

Dosadíme-li do (175) po řadě $z = g_+/y^T s$, $z = g_+/g^T g$, $z = g_+/|g^T s|$, dostaneme spádové modifikace metod (134), pro které

$$\beta^{DYD} = \beta^{DY} - \lambda\frac{g_+^T g_+ g_+^T s}{(y^T s)^2}, \qquad \beta^{FRD} = \beta^{FR} - \lambda\frac{g_+^T g_+ g_+^T s}{(g^T g)^2}, \qquad \beta^{CDD} = \beta^{CD} - \lambda\frac{g_+^T g_+ g_+^T s}{(g^T s)^2}. \tag{180}$$

Dosadíme-li do (175) po řadě $z = (y-d)/y^T s$, $z = (y-d)/g^T g$, $z = (y-d)/|g^T s|$, dostaneme spádové modifikace metod (135), pro které

$$\beta^{HPD} = \beta^{HP} - \lambda\frac{p^T p g_+^T s}{(y^T s)^2}, \qquad \beta^{PPD} = \beta^{PP} - \lambda\frac{p^T p g_+^T s}{(g^T g)^2}, \qquad \beta^{LPD} = \beta^{LP} - \lambda\frac{p^T p g_+^T s}{(g^T s)^2}, \tag{181}$$

kde $p = y - d$. Volíme-li parametr ϑ podle (167), dostaneme podobným způsobem

$$\beta^{HSD} = \beta^{HS} - \lambda\frac{y^T y g_+^T s}{(y^T s)^2}, \qquad \beta^{PRD} = \beta^{PR} - \lambda\frac{y^T y g_+^T s}{y^T s g^T g}, \qquad \beta^{LSD} = \beta^{LS} - \lambda\frac{y^T y g_+^T s}{y^T s |g^T s|}, \tag{182}$$

$$\beta^{DYD} = \beta^{DY} - \lambda\frac{g_+^T g_+ g_+^T s}{(y^T s)^2}, \qquad \beta^{FRD} = \beta^{FR} - \lambda\frac{g_+^T g_+ g_+^T s}{y^T s g^T g}, \qquad \beta^{CDD} = \beta^{CD} - \lambda\frac{g_+^T g_+ g_+^T s}{y^T s |g^T s|}, \tag{183}$$

$$\beta^{HPD} = \beta^{HP} - \lambda\frac{p^T p g_+^T s}{(y^T s)^2}, \qquad \beta^{PPD} = \beta^{PP} - \lambda\frac{p^T p g_+^T s}{y^T s g^T g}, \qquad \beta^{LPD} = \beta^{LP} - \lambda\frac{p^T p g_+^T s}{y^T s |g^T s|}. \tag{184}$$

Splňují-li parametry $\vartheta > 0$ a $\lambda > 0$ předpoklady lemmatu 22, jsou všechny uvedené metody spádové (definice 34).

Poznámka 83. Ve vzorcích (179)–(183) se vyskytuje parametr $\lambda > 1/4$. Ukazuje se, že je vhodné volit jeho hodnotu v intervalu $1/2 \leq \lambda \leq 2$ (raději v jeho dolní části, například $\lambda = 1/2$). Položíme-li $\lambda = 2$, odpovídá první vzorec v (179) metodě Hagera a Zhanga [78].

$$\beta_i^{HZ} = \frac{y_i^T g_{i+1}}{y_i^T s_i} - 2\frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}. \tag{185}$$

Základní metody sdružených gradientů lze také kombinovat tak, že volíme

$$\beta_i = \frac{\lambda_i^1 g_{i+1}^T y_i + \lambda_i^2 g_{i+1}^T g_{i+1}}{\mu_i^1 y_i^T s_i + \mu_i^2 g_i^T g_i - \mu_i^3 g_i^T s_i} = \frac{g_{i+1}^T (g_{i+1} - \lambda_i^1 g_i)}{\mu_i^1 y_i^T s_i + \mu_i^2 g_i^T g_i - \mu_i^3 g_i^T s_i} = \frac{g_{i+1}^T \tilde{y}_i}{\mu_i^1 y_i^T s_i + \mu_i^2 g_i^T g_i - \mu_i^3 g_i^T s_i}, \tag{186}$$

kde λ_i^1, λ_i^2, μ_i^1, μ_i^2, μ_i^3 jsou nezáporná čísla taková, že $\lambda_i^1 + \lambda_i^2 = 1$, $\mu_i^1 + \mu_i^2 + \mu_i^3 = 1$, a kde $\tilde{y}_i = g_{i+1} - \lambda_i^1 g_i$. Metody tohoto typu nejsou citlivé na výběr parametrů μ_i^1, μ_i^2, μ_i^3. Proto se omezíme pouze na jmenovatele, které se vyskytují ve vzorcích (133)–(134). Pak lze výsledné metody chápat jako konvexní kombinace odpovídajících si metod z (133)–(134). Položíme-li $\lambda_i^1 = 1$, dostaneme metody (133). Položíme-li $\lambda_i^1 = 0$, dostaneme metody (134).

Poznámka 84. Jednou z možností je volba $\lambda_i^1 = \min(1, \|g_{i+1}\|/\|g_i\|)$, která vede k modifikacím

$$\beta_i^{HSW} = \frac{g_{i+1}^T \tilde{y}_i}{y_i^T s_i}, \quad \beta_i^{PRW} = \frac{g_{i+1}^T \tilde{y}_i}{g_i^T g_i}, \quad \beta_i^{LSW} = \frac{g_{i+1}^T \tilde{y}_i}{|g_i^T s_i|}, \tag{187}$$

kde $\tilde{y}_i = g_{i+1} - \min(1, \|g_{i+1}\|/\|g_i\|) g_i$. Také se používá volba $\lambda_i^1 = \|g_{i+1}\|/\|g_i\|$ (pro kterou obecně neplatí $\lambda_i^1 \leq 1$), která vede k modifikacím

$$\begin{aligned} \beta_i^{HSW} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{y_i^T s_i}, \\ \beta_i^{PRW} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{g_i^T g_i}, \\ \beta_i^{LSW} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{|g_i^T s_i|} \end{aligned} \tag{188}$$

(W - Wei, Yao a Liu [167]).

Věta 55. *Hodnoty určené vzorci (188) vyhovují nerovnostem* $0 \leq \beta_i^{HSW} \leq 2\beta_i^{DY}$, $0 \leq \beta_i^{PRW} \leq 2\beta_i^{FR}$, $0 \leq \beta_i^{LSW} \leq 2\beta_i^{CD}$. *Předpokládejme, že je splněna silná Wolfeho podmínka. Jestliže* $\varepsilon_2 < 1/2$, *je metoda LSW spádová (platí (162) s* $\underline{s} > 0$*). Jestliže* $\varepsilon_2 < 1/3$, *je metoda HSW spádová. Jestliže* $\varepsilon_2 < 1/4$, *je metoda PRW spádová.*

Důkaz (a) Jelikož

$$\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i = \|g_{i+1}\|^2 \left(1 - \frac{g_{i+1}^T g_i}{\|g_{i+1}\|\|g_i\|}\right)$$

a podle Schwarzovy nerovnosti platí $|g_{i+1}^T g_i| \le \|g_{i+1}\|\|g_i\|$, můžeme psát

$$0 \le \|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i \le 2\|g_{i+1}\|^2, \tag{189}$$

odkud plynou nerovnosti $0 \le \beta_i^{HSW} \le 2\beta_i^{DY}$, $0 \le \beta_i^{PRW} \le 2\beta_i^{FR}$, $0 \le \beta_i^{LSW} \le 2\beta_i^{CD}$.

(b) Použijeme-li (127), (134) a (a) (pro metody LSW a CD) spolu s (S3a), dostaneme

$$g_{i+1}^T s_{i+1} \le -\|g_{i+1}\|^2 + 2\frac{\|g_{i+1}\|^2}{|g_i^T s_i|}|g_{i+1}^T s_i| \le -(1-2\varepsilon_2)\|g_{i+1}\|^2,$$

a jelikož $\varepsilon_2 < 1/2$, platí (162) s $\underline{s} = (1-2\varepsilon_2) > 0$.

(c) Použijeme-li (127), (134) a (a) (pro metody HSW a DY) spolu s nerovností

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i \ge (\varepsilon_2 - 1) g_i^T s_i = (1-\varepsilon_2)|g_i^T s_i|$$

(která plyne z (S3a)), dostaneme

$$g_{i+1}^T s_{i+1} \le -\|g_{i+1}\|^2 + 2\frac{\|g_{i+1}\|^2}{y_i^T s_i}|g_{i+1}^T s_i| \le -\|g_{i+1}\|^2 + \frac{2\|g_{i+1}\|^2|g_{i+1}^T s_i|}{(1-\varepsilon_2)|g_i^T s_i|} \le -\left(1 - \frac{2\varepsilon_2}{1-\varepsilon_2}\right)\|g_{i+1}\|^2,$$

a jelikož $\varepsilon_2 < 1/3$, platí (162) s $\underline{s} = (1 - 2\varepsilon_2/(1-\varepsilon_2)) = (1-3\varepsilon_2)/(1-\varepsilon_2) > 0$

(d) Použijeme-li (134) a (a) (pro metody PRW a FR), dostaneme

$$|\beta_i^{PRW}| \le 2\beta_i^{FR} = \frac{\tilde{\varepsilon}_2}{\varepsilon_2}\frac{\|g_{i+1}\|^2}{\|g_i\|^2}.$$

kde $0 < \varepsilon_2 < 2\varepsilon_2 = \tilde{\varepsilon}_2 < 1/2$ (neboť $\varepsilon_2 < 1/4$), takže podle poznámky 70 platí nerovnost (142), neboli $s_{i+1}^T g_{i+1} \le -\underline{s}\|g_{i+1}\|$, kde $\underline{s} = (1-2\tilde{\varepsilon}_2)/(1-\tilde{\varepsilon}_2) = (1-4\varepsilon_2)/(1-2\varepsilon_2) > 0$. □

Poznámka 85. Jelikož $\min(1, \|g_{i+1}\|/\|g_i\|) \le \|g_{i+1}\|/\|g_i\|$, platí věta 56 i pro metody s parametry (187).

Konvergenční vlastnosti metod HSW, PRW, LSW lze zlepšit úpravou jmenovatelů v (188). Dostaneme tak metody HSH, PRH, LSH, studované v [85], jejichž parametry jsou určeny vztahy

$$\begin{aligned}
\beta_i^{HSH} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{y_i^T s_i + \mu_i \max(0, g_{i+1}^T s_i)}, \\
\beta_i^{PRH} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{g_i^T g_i + \mu_i \max(0, g_{i+1}^T s_i)}, \\
\beta_i^{LSH} &= \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{|g_i^T s_i| + \mu_i \max(0, g_{i+1}^T s_i)},
\end{aligned} \tag{190}$$

kde $\mu_i > 0$ je volný parametr (H - Huang a Lin [85]). Ve jmenovatelích můžeme výraz $\max(0, g_{i+1}^T s_i)$ nahradit absolutní hodnotou $|g_{i+1}^T s_i|$. Teoretické vlastnosti metody se tím nezmění [85]. Numerické testy ukazují, že původní vzorce (190) jsou výhodnější.

Věta 56. *Hodnoty určené vzorci (190) vyhovují nerovnostem* $0 \le \beta_i^{HSH} \le 2\beta_i^{DY}$, $0 \le \beta_i^{PRH} \le 2\beta_i^{FR}$, $0 \le \beta_i^{LSH} \le 2\beta_i^{CD}$. *Jestliže* $\mu_i \ge \underline{\mu} > 2$, $i \in N$, *platí (162) s* $\underline{s} = (1 - 2/\underline{\mu}) > 0$.

Důkaz (a) Nerovnosti pro parametry β_i^{HSH}, β_i^{PRH}, β_i^{LSH}, plynou z věty 56, neboť porovnáním (190) s (188) dostaneme $0 \le \beta_i^{HSH} \le \beta_i^{HSW}$, $0 \le \beta_i^{PRH} \le \beta_i^{PRW}$, $0 \le \beta_i^{LSH} \le \beta_i^{LSW}$.

(b) Označme $\beta_i \ge 0$ libovolnou z hodnot (190). Pak platí

$$g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1} + \beta_i\, g_{i+1}^T s_i.$$

Pokud $g_{i+1}^T s_i \le 0$, je druhý člen v této rovnosti, záporný, takže $g_{i+1}^T s_{i+1} \le -g_{i+1}^T g_{i+1} \le -(1-2/\underline{\mu})\, g_{i+1}^T g_{i+1}$. V opačném případě použitím (189) a (190) dostaneme

$$\beta_i \le \frac{\|g_{i+1}\|^2 - \frac{\|g_{i+1}\|}{\|g_i\|} g_{i+1}^T g_i}{\mu_i g_{i+1}^T s_i} \le \frac{2\|g_{i+1}\|^2}{\mu_i g_{i+1}^T s_i},$$

takže

$$g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1} + \beta_i g_{i+1}^T s_i \le -\left(1 - \frac{2}{\mu_i}\right) g_{i+1}^T g_{i+1} \le -\left(1 - \frac{2}{\underline{\mu}}\right) g_{i+1}^T g_{i+1}.$$

□

Jak již bylo zmíněno, můžeme do vzorce (127) přidat členy, které vymizí, pokud $s_i^T g_{i+1} = 0$. Jednou z možností je použít vztah $s_{i+1} = -H_{i+1} g_{i+1}$, kde H_{i+1} je matice, která vznikne z jednotkové matice pomocí aktulizace BFGS (vzorec (288) uvedený v oddílu 4.1). V tomto případě platí

$$H_{i+1} = \gamma_i \left(I + \left(\frac{y_i^T y_i}{y_i^T d_i} + \frac{\rho_i}{\gamma_i} \right) \frac{1}{y_i^T d_i} d_i d_i^T - \frac{1}{y_i^T d_i} (y_i d_i^T + d_i y_i^T) \right), \tag{191}$$

kde $d_i = x_{i+1} - x_i = \alpha_i s_i$ a $y_i = g_{i+1} - g_i$, takže

$$s_{i+1} = -H_{i+1} g_{i+1} = -\gamma_i \left(g_{i+1} + \left(\frac{y_i^T y_i}{y_i^T d_i} + \frac{\rho_i}{\gamma_i} \right) \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i - \frac{d_i^T g_{i+1}}{y_i^T d_i} y_i - \frac{y_i^T g_{i+1}}{y_i^T d_i} d_i \right). \tag{192}$$

Pokud $s_i^T g_{i+1} = 0$, dostaneme

$$s_{i+1} = -\gamma_i \left(g_{i+1} - \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i \right),$$

což je směrový vektor škálované metody Hestenese a Stiefela (poznámka 67). Matice H_{i+1} je pozitivně definitní, takže směrový vektor s_{i+1} je spádový i když $s_i^T g_{i+1} \ne 0$. Tato myšlenka tvoří podklad pro metody s proměnnou metrikou s omezenou pamětí a bude dále rozvíjena v oddílu 9.1.

Předpokládejme nyní, že $\gamma_i = 1$ a $\rho_i = 1$. Vynecháme-li ve vzorci (192) všechny členy obsahující výraz $d_i^T g_{i+1}$, dostaneme metodu HS. Můžeme však postupovat také tak, že vynecháme pouze některé členy. Vynecháme-li v (192) člen úměrný y_i a první člen úměrný d_i, dostaneme Perryho modifikaci metody HS s parametrem

$$\beta_i^{HP} = \frac{(y_i - d_i)^T g_{i+1}}{y_i^T s_i}.$$

Tuto metodu můžeme zobecnit tím, že využijeme i ostatní jmenovatele. Dostaneme tak metody HP, PP, LP s parametry uvedenými v (135).

Matice (191) splňuje kvazinewtonovskou podmínku $H_{i+1} y_i = \rho_i d_i$, takže platí

$$-y_i^T s_{i+1} = y_i^T H_{i+1} g_{i+1} = \rho_i d_i^T g_{i+1},$$

zatímco pro metodu HS je splněna podmínka $-y_i^T s_{i+1} = 0$. Nabízí se tedy myšlenka, upravit metodu HS co nejjednodušším způsobem tak, aby platilo $-y_i^T s_{i+1} = \rho_i d_i^T g_{i+1}$, kde $\rho_i > 0$. Použijeme-li vztah (127), dostaneme $-y_i^T s_{i+1} = y_i^T g_{i+1} - \beta_i y_i^T s_i$, takže $-y_i^T s_{i+1} = \rho_i d_i^T g_{i+1}$ platí pro

$$\beta_i^{DL} = \frac{y_i^T g_{i+1} - \rho_i d_i^T g_{i+1}}{y_i^T s_i} = \beta_i^{HS} - \rho_i \frac{d_i^T g_{i+1}}{y_i^T s_i} \tag{193}$$

(DL – Dai a Liao [31]). Poznamenejme, že pokud položíme $\rho_i = \lambda y_i^T y_i / y_i^T d_i$ dostaneme metodu HSD uvedenou v (179).

Poznámka 86. Vzorec (193) se velmi podobá prvnímu vzorci v (179) (oba vzorce dávají stejnou hodnotu, pokud $\rho = \lambda y^T y / y^T d$). Proto se nabízí zobecnění vycházející ze zbylých vzorců v (179). Dostaneme tak metody

$$\beta^{HSL} = \beta^{HS} - \rho \frac{g_+^T d}{y^T s}, \qquad \beta^{PRL} = \beta^{PR} - \rho \frac{g_+^T d}{g^T g}, \qquad \beta^{LSL} = \beta^{LS} - \rho \frac{g_+^T d}{|g^T s|} \tag{194}$$

(zřejmě $\beta^{HSL} = \beta^{DL}$). Dosadíme-li do β^{HSL}, β^{PRL}, β^{LSL} hodnoty $\rho = \lambda y^T y / y^T d$, $\rho = \lambda y^T y / g^T g$, $\rho = \lambda y^T y / |g^T s|$, dostaneme metody HSD, PRD, LSD. Metody HSL, PRL, LSL jsou tedy spádové, pokud $\rho \geq \underline{\lambda} y^T y / y^T d$, $\rho \geq \underline{\lambda} y^T y / g^T g$, $\rho \geq \underline{\lambda} y^T y / |g^T s|$, kde $\underline{\lambda} > 1/4$.

Poznámka 87. Parametr $\rho_i > 0$ lze vybírat různým způsobem. Tak jako v oddílu 4.4 můžeme použít různé modely minimalizované funkce (vzorce (398), (399), (401)) nebo inverzní škálování, kdy

$$\rho_i = \lambda_i \frac{y_i^T y_i}{y_i^T d_i}, \qquad \rho_i = \lambda_i \frac{y_i^T d_i}{d_i^T d_i}, \qquad \rho_i = \lambda_i \sqrt{\frac{y_i^T y_i}{d_i^T d_i}}, \tag{195}$$

Dai a Liao používají konstantní hodnotu $\rho_i = 0.1$. Numerické testy ukazují, že používáme-li konstantní hodnotu, je výhodnější pokládat $\rho_i = 1$, což odpovídá standardní kvazinewtonovské podmínce $H_{i+1} y_i = d_i$.

Poznámka 88. Pokud $\gamma_i = 1$ a $\rho_i = 1$ můžeme vzorec (192) zapsat ve tvaru

$$\begin{aligned} s_{i+1} &= -g_{i+1} + \left(\frac{y_i^T g_{i+1}}{y_i^T d_i} - \frac{y_i^T y_i d_i^T g_{i+1}}{(y_i^T d_i)^2} \right) d_i - \frac{d_i^T g_{i+1}}{y_i^T d_i} (d_i - y_i) \\ &= -g_{i+1} + \beta_i^{HSD} s_i - \zeta_i^{HS} (d_i - y_i). \end{aligned} \tag{196}$$

Dostáváme tak metodu, která se od metody HSD s $\lambda = 1$ liší pouze přidáním dalšího členu. Tento člen se od posledního členu v (172) (kde $\beta_i = \beta_i^{HS}$ a $\zeta_i = \zeta_i^{HS}$) liší tím, že vektor y_i je nahražen vektorem $d_i - y_i$. Metoda, která používá směrové vektory (196) je ekvivalentní jednokrokové metodě BFGS s omezenou pamětí a je tedy globálně konvergentní (věta 170). Vzorec (196) můžeme zobecnit tím že využijeme i ostatní jmenovatele. Dostaneme tak metody tvaru

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \beta_i s_i - \zeta_i (d_i - y_i) \quad \text{pro} \quad i \in N, \tag{197}$$

kde za β_i a ζ_i dosazujeme postupně β_i^{HSD}, β_i^{PRD}, β_i^{LSD} z (179) a ζ_i^{HS}, ζ_i^{PR}, ζ_i^{LS} z (174).

V dalším výkladu budeme předpokládat, že $\gamma_i = 1$ a obecně $\rho_i \neq 1$ (případ $\rho_i = 1$. nevylučujeme) Za těchto předpokladů lze vzorec (192) zapsat ve tvaru

$$s_{i+1} = -g_{i+1} + \left(\frac{y_i^T g_{i+1}}{y_i^T d_i} - \left(\frac{y_i^T y_i}{y_i^T d_i} + \rho_i \right) \frac{d_i^T g_{i+1}}{y_i^T d_i} \right) d_i + \frac{d_i^T g_{i+1}}{y_i^T d_i} y_i. \tag{198}$$

Motivováni tímto postupem, budeme vyšetřovat metodu, která používá vzorce

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \beta_i^{KD} s_i + \zeta_i^{KD} y_i, \tag{199}$$

kde

$$\beta_i^{KD} = \frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \rho_i \right) \frac{s_i^T g_{i+1}}{y_i^T s_i}, \qquad \zeta_i^{KD} = \mu_i \frac{s_i^T g_{i+1}}{y_i^T s_i} \tag{200}$$

(KD - Kou a Dai [87]), přičemž $0 < \underline{\rho} \leq \rho_i \leq \overline{\rho}$ a $0 \leq \mu_i \leq 1$, $i \in N$ (v (198) platí $\mu_i = 1$, $i \in N$).

Lemma 23. *Nechť $s_+ = -g_+ + \beta s + \zeta z$, přičemž*

$$\beta = g_+^T z - (\lambda + 1)\, z^T z\, g_+^T s, \qquad \zeta = \mu\, g_+^T s, \tag{201}$$

kde $z \in R^n$ je libovolný nenulový vektor a $0 \leq \underline{\lambda} \leq \lambda \leq \overline{\lambda}$, $0 \leq \mu \leq \overline{\mu} \leq 1$. Pak, pokud $\underline{\lambda} > 0$ nebo $\overline{\mu} < 1$, platí

$$-g_+^T s_+ \geq \underline{s} \|g_+\|^2, \qquad \underline{s} = 1 - \frac{(1 + \overline{\mu})^2}{4(1 + \underline{\lambda})} > 0. \tag{202}$$

Důkaz Podle předpokladu platí

$$\begin{aligned} -g_+^T s_+ &= g_+^T g_+ - \beta\, g_+^T s - \zeta g_+^T z = g_+^T g_+ - g_+^T z\, g_+^T s + (\lambda+1)\, z^T z\, (g_+^T s)^2 - \mu\, g_+^T s\, g_+^T z \\ &= g_+^T g_+ + (\lambda+1)\, z^T z\, (g_+^T s)^2 - (1+\mu)\, g_+^T s\, g_+^T z. \end{aligned} \tag{203}$$

Dosadíme-li do vztahu (178) vektory

$$u = \frac{1+\mu}{\sqrt{2(1+\lambda)}}\, g_+, \qquad v = \sqrt{2(1+\lambda)}\, g_+^T s\, z,$$

dostaneme

$$|(1+\mu) g_+^T s\, g_+^T z| \le \frac{1}{2}\left(\frac{(1+\mu)^2}{2(1+\lambda)} g_+^T g_+ + 2(1+\lambda) z^T z (g_+^T s)^2\right),$$

což po dosazení do (203) dává

$$\begin{aligned} -g_+^T s_+ &\ge g_+^T g_+ + (1+\lambda)\, z^T z\, (g_+^T s)^2 - \frac{(1+\mu)^2}{4(1+\lambda)} g_+^T g_+ - (1+\lambda)\, z^T z\, (g_+^T s)^2 \\ &= \left(1 - \frac{(1+\mu)^2}{4(1+\lambda)}\right) \|g_+\|^2 \ge \underline{s} \|g_+\|^2. \end{aligned}$$

Pokud $\underline{\lambda} > 0$ nebo $\overline{\mu} < 1$, platí $\underline{s} = 1 - (1+\overline{\mu})^2/(4+4\underline{\lambda}) > 0$. □

Věta 57. *Uvažujme metodu sdružených gradientů danou předpisem (199)–(200), kde* $0 \le \underline{\rho} \le \rho_i \le \underline{\rho}$ *a* $0 \le \mu_i \le \overline{\mu} \le 1$. *Pak pro tuto metodu platí tvrzení věty 40 a je-li* $\overline{\mu} < 1$ *je tato metoda spádová (platí (162)). Splňuje-li funkce* $F : \mathcal{D} \to R$ *předpoklad F4 je tato metoda spádová i tehdy, když* $\underline{\rho} > 0$ *a* $\overline{\mu} = 1$.

Důkaz Provádíme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i = 0$, takže $\beta_i^{KD} = \beta_i^{HS}$ a $\zeta_i^{KD} = 0$. Metoda (199)–(200) je tedy ekvivalentní metodě HS a vlastnost kvadratického ukončení zůstane zachována. Položíme-li

$$\rho_i = \lambda_i \frac{y_i^T y_i}{y_i^T d_i}, \qquad \alpha_i \rho_i = \lambda_i \frac{y_i^T y_i}{y_i^T s_i}, \qquad z_i = \frac{y_i}{y_i^T s_i},$$

lze (200) zapsat ve tvaru (201) a pokud $\overline{\mu} < 1$, je podle lemmatu 23 splněna nerovnost (202) s $\underline{s} > 0$. Splňuje-li funkce $F : \mathcal{D} \to R$ předpoklad F4, můžeme užitím (209) psát

$$\underline{\lambda} = \underline{\rho}\, \frac{y_i^T d_i}{y_i^T y_i} \ge \frac{\underline{\rho}}{\overline{\overline{G}}} > 0,$$

takže podle lemmatu 23 platí (202) s $\underline{s} > 0$. □

Rekurentní vztah (199) obsahuje tři členy. Proto je účelné hledat dvoučlenný vztah, který by co nejlépe aproximoval (199). Nechť $\beta_i = \beta_i^{KD}$ a $\zeta_i = \zeta_i^{KD}$. Položme $\beta_i^{DK} = \beta_i + \delta_i$ a hledejme δ_i tak, aby vektor $(\beta_i + \delta_i) s_i$ byl so nejblíže k $\beta_i s_i + \zeta_i y_i$. To nastane tehdy, je-li výraz $\|\delta_i s_i - \zeta_i y_i\|^2$ minimální. Jelikož

$$\|\delta_i s_i - \zeta_i y_i\|^2 = \delta_i^2 \|s_i\|^2 - 2\delta_i \zeta_i s_i^T y_i + \zeta_i^2 \|y_i\|^2,$$

získáme derivováním rovnici $2\delta_i \|s_i\|^2 - 2\zeta_i s_i^T y_i = 0$, takže podle (200), kde $\mu_i = 1$, platí

$$\delta_i = \zeta_i \frac{s_i^T y_i}{s_i^T s_i} = \zeta_i^{KD} \frac{s_i^T y_i}{s_i^T s_i} = \frac{s_i^T g_{i+1}}{s_i^T s_i}.$$

Dostaneme tak rekurentní vztah

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \beta_i^{DK} s_i, \tag{204}$$

kde

$$\beta_i^{DK} = \frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \rho_i\right) \frac{s_i^T g_{i+1}}{y_i^T s_i} + \frac{s_i^T g_{i+1}}{s_i^T s_i} \tag{205}$$

(DK - Dai a Kou [30]), přičemž $0 < \underline{\rho} \le \rho_i \le \overline{\rho}$.

Věta 58. *Uvažujme metodu sdružených gradientů danou předpisem (204)–(205), kde $\rho_i > 0$ je některá z hodnot (195). Pak pro tuto metodu platí tvrzení věty 40 a pokud $\lambda_i \geq 1$, je tato metoda spádová (používáme-li první hodnotu z (195), stačí aby byla splněna nerovnost $\lambda_i \geq \underline{\lambda}$, kde $\underline{\lambda} > 1/4$).*

Důkaz Provádíme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i = 0$, takže $\beta_i^{DK} = \beta_i^{HS}$. Metoda (204)–(205) je tedy ekvivalentní metodě HS a vlastnost kvadratického ukončení zůstane zachována. Označme $\rho_i^{(1)}$, $\rho_i^{(2)}$, $\rho_i^{(3)}$ hodnoty uvedené v (195).

(a) Jestliže $\rho_i = \rho_i^{(1)}$, pak podle (204)–(205) a lemmatu 22, platí

$$\begin{aligned}
-g_{i+1}s_{i+1} &= g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \rho_i^{(1)}\right)\frac{s_i^T g_{i+1}}{y_i^T s_i} + \frac{s_i^T g_{i+1}}{s_i^T s_i}\right) g_{i+1}^T s_i \\
&= g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \lambda_i \frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}\right) s_i^T g_{i+1}^T + \frac{s_i^T s_i y_i^T y_i - (y_i^T s_i)^2}{y_i^T s_i s_i^T s_i}\frac{(s_i^T g_{i+1}^T)^2}{y_i^T s_i} \\
&\geq g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}\right) s_i^T g_{i+1} \geq \left(1 - \frac{1}{4}\right) g_{i+1}^T g_{i+1}.
\end{aligned}$$

(b) Jestliže $\rho_i = \rho_i^{(2)}$, pak podle (204)–(205) a lemmatu 22, platí

$$\begin{aligned}
-g_{i+1}s_{i+1} &= g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \rho_i^{(2)}\right)\frac{s_i^T g_{i+1}}{y_i^T s_i} + \frac{s_i^T g_{i+1}}{s_i^T s_i}\right) g_{i+1}^T s_i \\
&= g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}\right) s_i^T g_{i+1}^T + (\lambda_i - 1)\frac{(s_i^T g_{i+1})^2}{s_i^T s_i} \\
&\geq g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}\right) s_i^T g_{i+1} \geq \left(1 - \frac{1}{4}\right) g_{i+1}^T g_{i+1}.
\end{aligned}$$

(c) Jelikož $\rho_i^{(3)} = \sqrt{\rho_i^{(1)}\rho_i^{(2)}}$, jsou splněny nerovnosti $\min(\rho_i^{(1)}, \rho_i^{(2)}) \leq \rho_i^{(3)} \leq \max(\rho_i^{(1)}, \rho_i^{(2)})$, takže pro $\rho_i = \rho_i^{(3)}$ podle (a) a (b) platí

$$\begin{aligned}
-g_{i+1}s_{i+1} &= g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \rho_i^{(3)}\right)\frac{s_i^T g_{i+1}}{s_i^T s_i} + \frac{s_i^T g_{i+1}}{s_i^T s_i}\right) g_{i+1}^T s_i \\
&\geq g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \left(\frac{y_i^T y_i}{y_i^T s_i} + \alpha_i \min(\rho_i^{(1)}, \rho_i^{(2)})\right)\frac{s_i^T g_{i+1}}{y_i^T s_i} + \frac{d_i^T g_{i+1}}{d_i^T s_i}\right) g_{i+1}^T s_i \\
&\geq g_{i+1}^T g_{i+1} - \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - \frac{y_i^T y_i s_i^T g_{i+1}}{(y_i^T s_i)^2}\right) s_i^T g_{i+1} \geq \left(1 - \frac{1}{4}\right) g_{i+1}^T g_{i+1}.
\end{aligned}$$

□

3.5 Globální konvergence spádových metod sdružených gradientů

Při vyšetřování globální konvergence spádových metod sdružených gradientů budeme předpokládat, že směrové vektory lze vyjádřit ve tvaru

$$s_1 = -g_1 \quad a \quad s_{i+1} = s_{i+1}^{(1)} + \beta_i^{(2)} s_i \quad pro \quad i \in N. \tag{206}$$

Nejprve budeme dokazovat globální konvergenci v případě že minimalizovaná funkce je stejnoměrně silně konvexní (platí předpoklad F5).

Lemma 24. *Uvažujme spádovou metodu sdružených gradientů (definice 34), pro kterou platí (206), kde*

$$\|s_{i+1}^{(1)}\| \leq C_1 \|g_{i+1}\| \qquad |\beta_i^{(2)}| \leq C_2 \frac{\|g_{i+1}\|}{\|s_i\|}, \qquad i \in N \tag{207}$$

(konstanty $C_1 > 0$ a $C_2 > 0$ nezávisejí na indexu $i \in N$). Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1 a F3 a vybíráme-li délku kroku pomocí slabé Wolfeho podmínky, je tato metoda stejnoměrně spádová a tudíž globálně konvergentní.

Důkaz Použijeme-li vztah (206) a nerovnosti (207), dostaneme

$$\|s_{i+1}\| \le \|s_{i+1}^{(1)}\| + |\beta_i^{(2)}|\|s_i\| \le C_1\|g_{i+1}\| + C_2\frac{\|g_{i+1}\|}{\|s_i\|}\|s_i\| = \overline{s}\|g_{i+1}\|,$$

kde $\overline{s} = C_1 + C_2$. Protože je splněna podmínka (162), je podle poznámky 27 uvažovaná metoda stejnoměrně spádová (platí (S1b) s $\varepsilon_0 = \underline{s}/\overline{s}$), takže podle poznámky 29 dostaneme $\|g_i\| \to 0$. □

Lemma 24 použijeme k vyšetřování globální konvergence spádových metod sdružených gradientů odvozených ze základních metod definovaných vztahy (133), (134), (135). Metody DY, FR, CD hůře zachovávají sdruženost směrových vektorů a ortogonalitu gradientů. Proto budeme hodnoty (134) používat pouze tehdy, když

$$|g_{i+1}^T g_i| \le \eta g_{i+1}^T g_{i+1}, \tag{208}$$

kde $0 < \eta < 1$ (v opačném případě položíme $\beta_i = 0$). V tomto případě budeme psát (134)+(208) místo (134). Pokud použijeme zobecněnou Wolfeho podmínku, budeme bez újmy na obecnosti předpokládat, že $\varepsilon_3 \ge \varepsilon_2$.

Lemma 25. *Uvažujme spádovou metodu sdružených gradientů používající při výběru délky kroku zobecněnou Wolfeho podmínku takovou, že platí (164). Označme β_i některou z hodnot (133), (134)+(208), (135). Pak splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5, existuje konstanta $C_2 > 0$ taková, že platí $|\beta_i| \le C_2\|g_{i+1}\|/\|s_i\|$. V případě hodnot β_i^{HS}, β_i^{DY}, β_i^{HP} stačí použít slabou Wolfeho podmínku. Podmínku (164) potřebujeme pouze tehdy, používáme-li hodnoty β_i^{PR}, β_i^{FR}, β_i^{PP}. V ostatních případech stačí je-li splněna podmínka (162).*

Důkaz Jelikož $y_i = \tilde{G}_i d_i$, kde $\tilde{G}_i$ je matice určená vzorcem (136), můžeme s použitím předpokladů F4 a F5 psát

$$\|y_i\| = \|\tilde{G}_i d_i\| \le \overline{G}\|d_i\|, \qquad y_i^T d_i = d_i^T \tilde{G}_i d_i \ge \underline{G}\|d_i\|^2. \tag{209}$$

(a) Použijeme-li Schwarzovu nerovnost a první nerovnost v (209), dostaneme

$$\begin{aligned} |y_i^T g_{i+1}| &\le \|y_i\|\|g_{i+1}\| \le \overline{G}\|d_i\|\|g_{i+1}\|, \\ |(y_i - d_i)^T g_{i+1}| &\le \|y_i\|\|g_{i+1}\| + \|d_i\|\|g_{i+1}\| \le (\overline{G}+1)\|d_i\|\|g_{i+1}\|. \end{aligned}$$

Platí-li (208), můžeme psát $|y_i^T g_{i+1}| = |(g_{i+1} - g_i)^T g_{i+1}| \ge g_{i+1}^T g_{i+1} - |g_i^T g_{i+1}| \ge (1-\eta) g_{i+1}^T g_{i+1}$, takže

$$g_{i+1}^T g_{i+1} \le \frac{1}{1-\eta}|y_i^T g_{i+1}| \le \frac{\overline{G}}{1-\eta}\|d_i\|\|g_{i+1}\|. \tag{210}$$

(b) Podle druhé nerovnosti v (209) platí

$$y_i^T s_i = \frac{1}{\alpha_i} y_i^T d_i \ge \frac{1}{\alpha_i}\underline{G}\|d_i\|^2 = \underline{G}\|d_i\|\|s_i\|.$$

Vzhledem k tomu, že je splněna nerovnost (S3a), kde $\varepsilon_3 \ge \varepsilon_2$, můžeme tak jako v (168) psát

$$(1+\varepsilon_3)|g_i^T s_i| \ge |g_{i+1}^T s_i| + |g_i^T s_i| \ge y_i^T s_i$$

a je-li splněna podmínka (164) platí $g_i^T g_i = |g_i^T s_i|$.

(c) Použijeme-li nerovnosti uvedené v (a) a (b), vidíme, že pro libovolnou hodnotu β_i, určenou podle vzorců (133), (134)+(208), (135), platí

$$|\beta_i| \le C_2\frac{\|g_{i+1}\|}{\|s_i\|}, \qquad C_2 = \frac{(1+\varepsilon_3)(\overline{G}+1)}{(1-\eta)\underline{G}}. \tag{211}$$

V případě metod HS, DY, HP odpadne faktor $1+\varepsilon_3$ (stačí slabá Wolfeho podmínka). V případě parametrů (133) a (135) odpadne faktor $1/(1-\eta)$ (nepoužíváme nerovnost (208)). Podmínka (164) implikuje rovnost $g_i^T g_i = |g_i^T s_i|$. Používáme-li ve vzorcích (133)–(135) jmenovatele $y_i^T d_i$ nebo $|g_i^T s_i|$, stačí když je splněna podmínka (162). □

Věta 59. *Uvažujme metodu sdružených gradientů danou předpisem (165), kde β_i je některá z hodnot (133), (134)+(208), (135). Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda stejnoměrně spádová a tedy globálně konvergentní. Pokud $\beta_i = \beta_i^{HS}$, $\beta_i = \beta_i^{DY}$, $\beta_i = \beta_i^{HP}$, stačí použít slabou Wolfeho podmínku.*

Důkaz Položme

$$s_{i+1}^{(1)} = -\left(1 + \beta_i \frac{g_{i+1}^T s_i}{g_{i+1}^T g_{i+1}}\right) g_{i+1}, \qquad \beta_i^{(2)} = \beta_i,$$

kde β_i je některá z hodnot (133), (134)+(208), (135). Podle lemmatu 25 platí $|\beta_i| \le C_2 \|g_{i+1}\|/\|s_i\|$, takže

$$\|s_{i+1}^{(1)}\| \le \left(1 + C_2 \frac{\|g_{i+1}\|}{\|s_i\|} \frac{\|g_{i+1}\|\|s_i\|}{\|g_{i+1}\|^2}\right) \|g_{i+1}\| = C_1 \|g_{i+1}\|,$$

kde $C_1 = 1 + C_2$. Jsou tedy splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní. □

Věta 60. *Uvažujme metodu sdružených gradientů danou předpisem (172), kde β_i je některá z hodnot (133), (134)+(208) a ζ_i je hodnota určená vztahem (173) s $p_i = y_i$. Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda stejnoměrně spádová a tedy globálně konvergentní. Pokud $\beta_i = \beta_i^{HS}$, $\beta_i = \beta_i^{DY}$, $\beta_i = \beta_i^{HP}$, stačí použít slabou Wolfeha podmínku.*

Důkaz (a) Položme

$$s_{i+1}^{(1)} = -g_{i+1} - \zeta_i y_i, \qquad \beta_i^{(2)} = \beta_i,$$

kde β_i je některá z hodnot (133) a ζ_i je hodnota určená podle vzorce (173) s $p_i = y_i$, takže platí (174). Ukážeme, že hodnoty (174) splňují nerovnost $|\zeta_i| \le C\|g_{i+1}\|/\|d_i\|$. Použijeme-li nerovnost $|g_{i+1}^T s_i| \le \|g_{i+1}\|\|s_i\|$ spolu s nerovnostmi uvedenými v části (b) důkazu lemmatu 25, vidíme, že pro libovolnou hodnotu ζ_i, určenou podle vzorců (174), platí

$$|\zeta_i| \le C \frac{\|g_{i+1}\|}{\|d_i\|}, \qquad C = \frac{1+\varepsilon_3}{\underline{G}}.$$

(b) Nechť β_i je některá z hodnot (134)+(208) a ζ_i je hodnota určená podle vzorce (173) s $p_i = y_i$. Podle (210) platí $g_{i+1}^T g_{i+1}/|g_{i+1}^T y_i| \le 1/(1-\eta)$, takže s použitím (134), (173) s $p_i = y_i$ a (174) dostaneme

$$|\zeta_i^{DY}| \le \frac{|\zeta_i^{HS}|}{1-\eta}, \qquad |\zeta_i^{FR}| \le \frac{|\zeta_i^{PR}|}{1-\eta}, \qquad |\zeta_i^{CD}| \le \frac{|\zeta_i^{LS}|}{1-\eta}, \tag{212}$$

což podle (a) dává

$$|\zeta_i| \le C \frac{\|g_{i+1}\|}{\|d_i\|}, \qquad C = \frac{1+\varepsilon_3}{(1-\eta)\underline{G}}.$$

(c) Použijeme-li (a), (b) a první nerovnost v (209), můžeme psát

$$\|s_{i+1}^{(1)}\| \le \|g_{i+1}\| + |\zeta_i|\|y_i\| \le \|g_{i+1}\| + C \frac{\|g_{i+1}\|}{\|d_i\|}\|y_i\| \le (1 + C\overline{G})\|g_{i+1}\| = C_1\|g_{i+1}\|,$$

kde $C_1 = 1 + C\overline{G}$. Jelikož podle lemmatu 25 platí (211), jsou splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní. □

Lemma 26. *Uvažujme spádovou metodu sdružených gradientů používající při výběru délky kroku zobecněnou Wolfeho podmínku. Označme β_i některou z hodnot (179), (180)+(208), (181), (182), (183)+(208), (184) (kromě hodnot β_i^{PRD}, β_i^{FRD}, β_i^{PPD}), kde $0 \le \lambda_i \le \overline{\lambda}$. Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5, existuje konstanta $C > 0$ taková, že platí $|\beta_i| \le C\|g_{i+1}\|/\|s_i\|$. V případě hodnot β_i^{HSD}, β_i^{DYD}, β_i^{HPD} stačí použít slabou Wolfeho podmínku.*

Důkaz (a) Použijeme-li druhou nerovnost v (209), můžeme psát

$$\begin{aligned} y_i^T y_i = \|y_i\|^2 &\le \overline{G}^2\|d_i\|^2, \\ p_i^T p_i = \|y_i - d_i\|^2 &\le (\overline{G}+1)^2\|d_i\|^2. \end{aligned}$$

Platí-li (208), můžeme podle (210) psát $\|g_{i+1}\| \le \overline{G}\|d_i\|/(1-\eta)$, takže

$$g_{i+1}^T g_{i+1} \le \frac{\overline{G}^2}{(1-\eta)^2}\|d_i\|^2$$

(b) Použijeme-li nerovnosti uvedené v části (b) důkazu lemmatu 25, dostaneme

$$(y_i^T s_i)^2 \ge \underline{G}^2\|d_i\|^2\|s_i\|^2,$$

$$y_i^T s_i |g_i^T s_i| \ge \frac{1}{1+\varepsilon_3}(y_i^T s_i)^2 \ge \frac{1}{1+\varepsilon_3}\underline{G}^2\|d_i\|^2\|s_i\|^2.$$

$$(g_i^T s_i)^2 \ge \frac{1}{(1+\varepsilon_3)^2}(y_i^T s_i)^2 \ge \frac{1}{(1+\varepsilon_3)^2}\underline{G}^2\|d_i\|^2\|s_i\|^2$$

(c) Uvažované hodnoty parametru β_i obsahují dva členy. První člen nabývá hodnot (133), (134)+(208), (135) a jeho absolutní hodnota splňuje podle lemmatu 25 nerovnost (211). Použijeme-li (a), (b) a nerovnost $|g_{i+1}^T s_i| \le \|g_{i+1}\|\|s_i\|$, můžeme podobným způsobem omezit druhý člen výrazem $\overline{\lambda}C_2^2\|g_{i+1}\|/\|s_i\|$, kde C_2 je konstanta použitá ve vzorci (211). Sečteme-li nerovnosti pro oba členy, dostaneme

$$|\beta_i| \le C\frac{\|g_{i+1}\|}{\|s_i\|}, \qquad C = C_2 + \overline{\lambda}C_2^2, \qquad C_2 = \frac{(1+\varepsilon_3)(\overline{G}+1)}{(1-\eta)\underline{G}}. \tag{213}$$

V případě metod HSD, DYD, HPD odpadne faktor $1+\varepsilon_3$ (stačí slabá Wolfeho podmínka). V případě parametrů (179), (181), (182), (184) odpadne faktor $1/(1-\eta)$ (nepoužíváme nerovnost (208)). □

V lemmatu 26 neuvažujeme hodnoty β_i^{PRD}, β_i^{FRD} a β_i^{PPD}. Je to proto, že absolutní hodnotu podílu $|g_i^T s_i|/g_i^T g_i$ nelze shora ohraničit, neboť nemáme k disposici nerovnost opačnou k (162) (například rovnost (164)).

Věta 61. *Uvažujme metodu sdružených gradientů danou předpisem (166), kde:*

- *(a) β_i je některá z hodnot (179), (180)+(208), (181) (kromě hodnot β_i^{PRD}, β_i^{FRD}, β_i^{PPD}), kde $1/4 < \underline{\lambda} \le \lambda \le \overline{\lambda}$ a $\vartheta_i = 1$.*
- *(b) β_i je některá z hodnot (182), (183)+(208), (184) (kromě hodnot β_i^{PRD}, β_i^{FRD}, β_i^{PPD}), kde $1/4 < \underline{\lambda} \le \lambda \le \overline{\lambda}$ a ϑ_i je odpovídající hodnota určená podle (169).*

Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda stejnoměrně spádová a tedy globálně konvergentní. Pokud $\beta_i = \beta_i^{HSD}$, $\beta_i = \beta_i^{DYD}$, $\beta_i = \beta_i^{HPD}$, stačí použít slabou Wolfeho podmínku.

Důkaz Položme

$$s_{i+1}^{(1)} = -\vartheta_i g_{i+1}, \qquad \beta_i^{(2)} = \beta_i,$$

kde β_i je některá z hodnot uvedených v dokazovaném tvrzení. Podle lemmatu 22 je splněna nerovnost (176), takže uvažovaná metode je spádová.

(a) Jelikož $\vartheta_i = 1$, můžeme psát $\|s_{i+1}^{(1)}\| = \|g_{i+1}\|$ a podle lemmatu 26 platí $|\beta_i^{(2)}| \leq C\|g_{i+1}\|/\|s_i\|$. Jsou tedy splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní.

(b) Zřejmě $\vartheta_i^{HS} = 1$. Stačí tedy vyšetřit případ, kdy $\vartheta_i = \vartheta_i^{LS}$. Je-li splněna zobecněná Wolfeho podmínka, můžeme tak jako v (168) psát

$$(1-\varepsilon_2)|g_i^T s_i| \leq |g_i^T s_i| - |g_{i+1}^T s_i| \leq y_i^T s_i \leq |g_i^T s_i| + |g_{i+1}^T s_i| \leq (1+\varepsilon_3)|g_i^T s_i|$$

(předpokládáme, že $\varepsilon_3 \geq \varepsilon_2$) a pokud $\vartheta_i = y_i^T s_i/|g_i^T s_i|$, platí $0 < (1-\varepsilon_2) \leq \vartheta_i \leq (1+\varepsilon_3)$, takže $\|s_{i+1}^{(1)}\| = \vartheta_i\|g_{i+1}\| \leq (1+\varepsilon_3)\|g_{i+1}\|$. Jelikož platí i (213), jsou splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní. □

Věta 62. *Uvažujme metodu sdružených gradientů danou předpisem (199)–(200), kde* $0 < \underline{\rho} \leq \rho_i \leq \overline{\rho}$ *a* $0 \leq \mu_i \leq \overline{\mu} \leq 1$. *Pak, splňuje-li funkce* $F : \mathcal{D} \to R$ *předpoklady F1, F4, F5 a používáme-li při výběru délky kroku slabou Wolfeho podmínku, je tato metoda stejnoměrně spádová a tedy globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -g_{i+1} - \zeta_i^{KD} y_i, \qquad \beta_i^{(2)} = \beta_i^{KD},$$

kde β_i^{KD} a ζ_i^{KD} jsou hodnoty určené podle vzorce (200). Druhý vzorec v (200) se liší od prvního vzorce v (174) pouze tím, že parametr ζ_i je vynásoben číslem $\mu_i \leq 1$. Existuje tedy konstanta $C_1 > 0$ taková, že $\|s_{i+1}^{(1)}\| \leq C_1\|g_{i+1}\|$. Položme $\alpha_i\rho_i = \lambda_i y_i^T y_i/y_i^T s_i$. Pak první vzorec v (200) se od prvního vzorce v (179) liší pouze tím, že hodnota $\lambda_i > 0$ je nahražena hodnotou $\lambda_i + 1 > 1$. Podle lemmatu 26 tedy existuje konstanta $C > 0$ taková, že $|\beta_i^{(2)}| \leq C\|g_{i+1}\|/\|s_i\|$. Jsou tedy splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní. □

Věta 63. *Uvažujme metodu sdružených gradientů danou předpisem (204)–(205), kde* $\rho_i > 0$ *je některá z hodnot (195) s* $\lambda_i \geq 1$. *Pak, splňuje-li funkce* $F : \mathcal{D} \to R$ *předpoklady F1, F4, F5 a používáme-li při výběru délky kroku slabou Wolfeho podmínku, je tato metoda stejnoměrně spádová a tedy globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -g_{i+1}, \qquad \beta_i^{(2)} = \beta_i^{DK},$$

kde β_i^{DK} je hodnota určená podle vzorce (205). Tato hodnota se od hodnoty β_i^{KD} liší tím, že obsahuje nový člen, pro který platí

$$\left|\frac{s_i^T g_{i+1}}{s_i^T s_i}\right| \leq \frac{\|s_i\|\|g_{i+1}\|}{\|s_i\|^2} = \frac{\|g_{i+1}\|}{\|s_i\|}.$$

Použijeme-li odhad $|\beta_i^{KD}| \leq C\|g_{i+1}\|/\|s_i\|$, získaný v důkazy věty 62, dostaneme

$$|\beta_i^{(2)}| \leq |\beta_i^{KD}| + \frac{\|g_{i+1}\|}{\|s_i\|} \leq (C+1)\frac{\|g_{i+1}\|}{\|s_i\|}.$$

Jelikož $\|s_{i+1}^{(1)}\| = \|g_{i+1}\|$, jsou splněny předpoklady lemmatu 24, takže uvažovaná metoda je stejnoměrně spádová a tudíž globálně konvergentní. □

Poznámka 89. Věty, které jsme zatím dokázali vyžadují, aby byl splněn předpoklad F5 (existence konstanty $\underline{G} > 0$), takže je lze použít pouze pro konvexní funkce. Z důkazu těchto vět je zřejmé, že předpoklad F5 slouží pouze k tomu, aby byla splněna druhá nerovnost v (209). Jednou z možností jak tento předpoklad obejít, je zvolit malé číslo $\underline{\tau} > 0$ a položit $\beta_i = 0$, pokud $y_i^T d_i < \underline{\tau}\|d_i\|^2$. Další možností, použitelnou v případě, kdy vzorec pro β_i obsahuje ve jmenovateli výraz $y_i^T s_i$, je nahradit vektor $y_i = g_{i+1} - g_i$ vektorem $\tilde{y}_i = y_i + \tau_i d_i$, kde $\tau_i = \max(0, \underline{\tau} - y_i^T d_i/d_i^T d_i)$. Je-li splněna slabá Wolfeho podmínka, platí $y_i^T d_i > 0$, takže $\underline{\tau} - y_i^T d_i/d_i^T d_i \leq \tau_i \leq \underline{\tau}$ a

$$\tilde{y}_i^T d_i = y_i^T d_i + \tau_i\|d_i\|^2 \geq \frac{y_i^T d_i}{d_i^T d_i}\|d_i\|^2 + \left(\underline{\tau} - \frac{y_i^T d_i}{d_i^T d_i}\right)\|d_i\|^2 = \underline{\tau}\|d_i\|^2.$$

Jelikož $\tilde{y}_i = (\tilde{G}_i + \tau_i I)d_i$, můžeme psát

$$\|\tilde{y}_i\| \leq \|\tilde{G}_i + \tau_i I\|\|d_i\| \leq (\overline{G} + \underline{\tau})\|d_i\|.$$

Vektory $\tilde{y}_i = y_i + \tau_i d_i$, $i \in N$, tedy splňují nerovnosti (209) a použijeme-li je v metodách HS, HSD, HSL, HP, HPD, HPL, KD a DK, platí všechna dokázaná tvrzení i pro nekonvexní funkce (v důkazech místo $\underline{G}$ a $\overline{G}$ používáme $\underline{\tau}$ a $\overline{G} + \underline{\tau}$). Místo konstanty $\underline{\tau}$ můžeme použít proměnnou hodnotu $\underline{\tau}_i = \overline{\tau}\min(1, \|g_i\|)$, kde $\overline{\tau} > 0$. Protože důkaz globální konvergence provádíme sporem a předpokládáme, že $\|g_i\| > \underline{\varepsilon}$, platí v tomto případě $\underline{\tau}_i \geq \underline{\tau}$, kde $\underline{\tau} = \overline{\tau}\min(1, \underline{\varepsilon})$. Poznamenejme, že použitím vektoru $\tilde{y}_i$ sice zaručíme globální konvergenci některých metod sdružených gradientů i pro nekonvexní funkce, porušíme tím však předpoklady věty 40, čímž příjdeme o vlastnost kvadratického ukončení. Nicméně, zvolíme-li číslo $\underline{\tau}$ dostatečně malé, změna vektoru y_i se téměř nikdy neprojeví. Tato praktická zkušenost ukazuje, že globální konvergence bývá porušena pouze výjimečně a lze ji zaručit i jednodušším způsobem než změnou vektoru y_i, například podle poznámky 32.

Nyní opustíme předpoklad F5 a budeme se snažit upravit spádové metody sdružených gradientů tak, aby byly globálně konvergentní za slabších předpokladů F1, F2, F3. Pro tento účel je klíčové následující tvrzení, jehož důkaz je modifikací důkazů podobných tvrzení uvedených v [63] a [78].

Lemma 27. *Uvažujme spádovou metodu sdružených gradientů se směrovými vektory (206) a výběrem délky kroku splňujícím slabou Wolfeho podmínku. Předpokládejme že funkce $F : \mathcal{D} \to R$ splňuje podmínky F1, F2 (kde $\overline{F} \geq F(x_i)$, $i \in N$), F3 a že $g_i \geq \underline{\varepsilon}$, $i \in N$, kde číslo $\underline{\varepsilon} > 0$ nezávisí na indexu $i \in N$. Pak existují-li konstanty $C_1 > 0$, $C_2 > 0$ takové, že*

$$\|s_{i+1}^{(1)}\| \leq C_1, \qquad 0 < \beta_i^{(2)} \leq C_2\|d_i\|, \qquad i \in N, \tag{214}$$

platí

$$\sum_{i=1}^{\infty} \frac{1}{\|s_i\|^2} = \infty. \tag{215}$$

Důkaz (a) Ukážeme, že platí

$$\sum_{i=1}^{\infty} \|u_{i+1} - u_i\|^2 < \infty, \qquad u_i = \frac{s_i}{\|s_i\|} = \frac{d_i}{\|d_i\|}, \qquad i \in N. \tag{216}$$

Označme

$$w_{i+1} = \frac{s_{i+1}^{(1)}}{\|s_{i+1}\|}, \qquad \delta_i = \frac{\|s_i\|}{\|s_{i+1}\|}\beta_i^{(2)} \geq 0. \tag{217}$$

Pak použitím (206) dostaneme

$$u_{i+1} = \frac{s_{i+1}}{\|s_{i+1}\|} = w_{i+1} + \delta_i u_i. \tag{218}$$

Jelikož $\|u_{i+1}\| = \|u_i\| = 1$, můžeme psát

$$\|u_{i+1} - \delta_i u_i\|^2 = 1 - 2\delta_i u_{i+1}^T u_i + \delta_i^2, \qquad \|\delta_i u_{i+1} - u_i\| = \delta_i^2 - 2\delta_i u_{i+1}^T u_i + 1,$$

takže podle (218) platí

$$\|w_{i+1}\| = \|u_{i+1} - \delta_i u_i\| = \|\delta_i u_{i+1} - u_i\|,$$

což spolu s (217) a nerovností $\|s_{i+1}^{(1)}\| \leq C_1$, dává

$$\begin{aligned} \|u_{i+1} - u_i\| &\leq (1 + \delta_i)\|u_{i+1} - u_i\| = \|(1 + \delta_i)u_{i+1} - (1 + \delta_i)u_i\| \\ &\leq \|u_{i+1} - \delta_i u_i\| + \|\delta_i u_{i+1} - u_i\| = 2\|w_{i+1}\| \leq 2C_1 \frac{1}{\|s_{i+1}\|} \end{aligned}$$

a použijeme-li nerovnost (35), která platí pro libovolnou spádovou metodu sdružených gradientů s výběrem délky kroku splňujícím slabou Wolfeho podmínku, dostaneme

$$\sum_{i=1}^{\infty}\|u_{i+1}-u_i\|^2 \le 4C_1^2\sum_{i=1}^{\infty}\frac{1}{\|s_{i+1}\|^2} \le 4C_1^2\sum_{i=1}^{\infty}\frac{1}{\|s_i\|^2} \le \frac{4C_1^2}{\underline{\varepsilon}^4}\sum_{i=1}^{\infty}\frac{\|g_i\|^4}{\|s_i\|^2} < \infty$$

(b) Označme D diametr množiny $\mathcal{D}_F(\overline{F})$, která je podle předpokladu F2 omezená. Ukážeme, že ke každému číslu $m \in N$ existuje index $\underline{i} \in N$ takový, že

$$\sum_{j=i}^{i+m-1}\|d_j\| \le 2D, \tag{219}$$

pokud $i \ge \underline{i}$. Nechť $m \in N$. Zřejmě

$$x_{i+m}-x_i = \sum_{j=i}^{i+m-1}(x_{j+1}-x_j) = \sum_{j=i}^{i+m-1} d_j = \sum_{j=i}^{i+m-1}\|d_j\|u_j = \sum_{j=i}^{i+m-1}\|d_j\|u_i + \sum_{j=i}^{i+m-1}\|d_j\|(u_j-u_i),$$

takže

$$\sum_{j=i}^{i+m-1}\|d_j\| = \sum_{j=i}^{i+m-1}\|d_j\|\|u_i\| \le \|x_{i+m}-x_i\| + \sum_{j=i}^{i+m-1}\|d_j\|\|u_j-u_i\| \le D + \sum_{j=i}^{i+m-1}\|d_j\|\|u_j-u_i\|. \tag{220}$$

Podle (a) existuje index $\underline{i} \in N$ takový, že

$$\sum_{i=\underline{i}}^{\infty}\|u_{i+1}-u_i\|^2 \le \frac{1}{4m}, \tag{221}$$

takže pro $i \ge \underline{i}$ a $i \le j < i+m$ použitím nerovnosti (20) dostaneme

$$\|u_j-u_i\| \le \sum_{k=i}^{j-1}\|u_{k+1}-u_k\| \le \sqrt{j-i}\left(\sum_{k=i}^{j-1}\|u_{k+1}-u_k\|^2\right)^{1/2} \le \sqrt{m}\left(\frac{1}{4m}\right)^{1/2} = \frac{1}{2},$$

což po dosazení do (220) dává (219).

(c) Předpokládejme nyní, že $m \ge 2\sqrt{2}C_2D$ a $\underline{i}$ je index, pro který platí (221). Použijeme-li (206), (214) a nerovnost (20), můžeme pro $i \in N$ psát

$$\|s_{i+1}\|^2 \le (\|s_{i+1}^{(1)}\| + \beta_i^{(2)}\|s_i\|)^2 \le 2\|s_{i+1}^{(1)}\|^2 + 2(\beta_i^{(2)})^2\|s_i\|^2 \le 2C_1^2 + 2C_2^2\|d_i\|^2\|s_i\|^2 = c_0 + c_i\|s_i\|^2,$$

kde $c_0 = 2C_1^2$ a $c_i = 2C_2^2\|d_i\|^2$, takže pro $l > \underline{i}$ indukcí dostaneme

$$\|s_l\|^2 \le c_0 + c_{l-1}\|s_{l-1}\|^2 \le c_0 + c_{l-1}(c_0 + c_{l-2}\|s_{l-2}\|^2) \le \cdots \le c_0\left(\sum_{i=\underline{i}}^{l}\prod_{j=i}^{l-1}c_j\right) + \|s_{\underline{i}}\|^2\prod_{j=\underline{i}}^{l-1}c_j. \tag{222}$$

Použijeme-li nerovnost (19), vztah (219) a nerovnost $m \ge 2\sqrt{2}C_2D$, můžeme pro $i \ge \underline{i}$ psát

$$\begin{aligned}\prod_{j=i}^{i+m-1}c_j &= \prod_{j=i}^{i+m-1}2C_2^2\|d_j\|^2 = \left(\prod_{j=i}^{i+m-1}\sqrt{2}C_2\|d_j\|\right)^2 \\ &\le \left(\frac{1}{m}\sum_{j=i}^{i+m-1}\sqrt{2}C_2\|d_j\|\right)^{2m} \le \left(\frac{2\sqrt{2}C_2D}{m}\right)^{2m} \le 1.\end{aligned}$$

Nechť $\underline{i}+(k'-1)m \le i < \underline{i}+k'm$, kde $1 \le k' \le k$. Pak opakovaným použitím předchozí nerovnosti pro $l = \underline{i}+k\,m$ dostaneme

$$\prod_{j=i}^{l-1} c_j = \prod_{j=i}^{\underline{i}+k'm-1} c_j \prod_{j=\underline{i}+k'm}^{\underline{i}+k\,m-1} c_j \le \prod_{j=i}^{\underline{i}+k'm-1} c_j = \prod_{j=i}^{\underline{i}+k'm-1} 2C_2^2\|d_j\|^2 \le \prod_{j=i}^{\underline{i}+k'm-1} 2C_2^2 D^2 \le (2C_2^2D^2)^m$$

(předpokládáme bez újmy na obecnosti, že $2C_2^2D^2 \ge 1$), což po dosazení do (222) dává

$$\|s_l\|^2 \le 2C_1^2(l-\underline{i}+1)(2C_2^2D^2)^m + (2C_2^2D^2)^m\|s_{\underline{i}}\|^2 \triangleq K_2(l-\underline{i}+1)+K_1 \le (K_1+K_2)l.$$

Můžeme tedy psát

$$\sum_{i=1}^{\infty}\frac{1}{\|s_i\|^2} \ge \sum_{k=1}^{\infty}\frac{1}{\|s_{\underline{i}+k\,m}\|^2} \ge \sum_{k=1}^{\infty}\frac{1}{(K_1+K_2)(\underline{i}+k\,m)} \ge \frac{1}{(K_1+K_2)(\underline{i}+m)}\sum_{k=1}^{\infty}\frac{1}{k} = \infty.$$

□

Poznámka 90. Lemma 27 je velmi užitečným prostředkem pro dokazování globální konvergence spádových metod sdružených gradientů. Důkaz se provádí sporem. Ukáže se, že z předpokladu $\|g_i\| \ge \underline{\varepsilon}$, $i \in N$, plynou nerovnosti (214). Pak podle lemmatu 27 platí (215), což pokud $\|g_i\| \ge \underline{\varepsilon}$, $i \in N$, je ve sporu s nerovností (35).

Lemma 28. *Uvažujme spádovou metodu sdružených gradientů používající při výběru délky kroku slabou Wolfeho podmínku, předpokládejme že funkce $F : \mathcal{D} \to R$ splňuje podmínky F1, F2, F3 a označme β_i některou z hodnot (133), (134)+(208)), (135). Pak, pokud $g_i \ge \underline{\varepsilon}$, $i \in N$, existuje konstanta $C_2 > 0$ taková, že $|\beta_i| \le C_2\|d_i\|$, $i \in N$.*

Důkaz Podle předpokladů F2, F3 existují čísla $\overline{g} > 0$ a $\overline{G} > 0$ taková že, $\|g_{i+1}\| \le \overline{g}$, $\|g_i\| \le \overline{g}$ a $\|y_i\| = \|g(x_i+d_i)-g(x_i)\| \le \overline{G}\|d_i\|$.

(a) Použijeme-li uvedené nerovnosti a nerovnost (210), dostaneme

$$\begin{aligned} |y_i^T g_{i+1}| &\le \|y_i\|\|g_{i+1}\| \le \overline{g}\overline{G}\|d_i\| \\ g_{i+1}^T g_{i+1} &\le \frac{1}{1-\eta}|y_i^T g_{i+1}| \le \frac{\overline{g}\overline{G}}{1-\eta}\|d_i\| \\ |(y_i-d_i)^T g_{i+1}| &\le \|y_i\|\|g_{i+1}\| + \|d_i\|\|g_{i+1}\| \le \overline{g}(\overline{G}+1)\|d_i\|. \end{aligned}$$

(b) Podle předpokladu platí $\|g_i\| \ge \underline{\varepsilon}$, takže

$$g_i^T g_i \ge \underline{\varepsilon}^2$$

a použijeme-li (162), dostaneme

$$|g_i^T s_i| = -g_i^T s_i \ge \underline{s}\, g_i^T g_i \ge \underline{s}\,\underline{\varepsilon}^2.$$

Jelikož používáme slabou Wolfeho podmínku, můžeme psát

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i \ge (\varepsilon_2-1)g_i^T s_i = (1-\varepsilon_2)|g_i^T s_i| \ge (1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2.$$

(c) Použijeme-li nerovnosti uvedené v (a) a (b), vidíme, že pro libovolnou hodnotu β_i, určenou podle vzorců (133), (134)+(208), (135), platí

$$|\beta_i| \le C_2\|d_i\|, \qquad C_2 = \frac{\overline{g}(\overline{G}+1)}{(1-\eta)(1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2}. \tag{223}$$

V případě parametrů (133) a (135) odpadne faktor $1/(1-\eta)$. □

Právě dokázaná tvrzení nevyžadují, aby byl splněn předpoklad F5. Jistou komplikací je však požadavek nezápornosti čísla $\beta_i^{(2)}$. Nechť

$$\beta_i = \beta_i^+ + \beta_i^-, \qquad \beta_i^+ = \max(0, \beta_i), \qquad \beta_i^- = \min(0, \beta_i).$$

Pak lze položit $\beta_i^{(2)} = \beta_i^+$ a člen $\beta_i^- s_i$ zahrnout do vektoru $s_{i+1}^{(1)}$. Jelikož chceme, aby platilo $\|s_{i+1}^{(1)}\| \leq C_1$, je třeba, aby existovala konstanta $C > 0$ taková, že $|\beta_i^-| \leq C/\|s_i\|$ (pokud $\|g_i\| \geq \underline{\varepsilon}$). To lze zajistit tak, že místo β_i použijeme hodnotu

$$\beta_i' = \max(\underline{\beta_i}, \beta_i), \qquad 0 \leq -\underline{\beta_i} \leq \frac{C}{\|s_i\|}, \tag{224}$$

Nejjednodušším způsobem je zvolit $\underline{\beta_i} = 0$, takže (podobně jako v (146)) $\beta_i' = \beta_i^+ = \max(0, \beta_i)$. V práci [78] se používá hodnota

$$\underline{\beta_i} = -\frac{1}{\|s_i\| \min(\gamma, \|g_i\|)}. \tag{225}$$

kde γ je vhodná konstanta (pak $C = 1/\min(\gamma, \underline{\varepsilon})$). V dalším výkladu budeme pro jednoduchost předpokládat, že $\underline{\beta_i} = 0$, takže $\beta_i' = \beta_i^+ = \max(0, \beta_i)$. Tento předpoklad, vhodný pro praktické výpočty, nikterak nesnižuje obecnost následujících úvah, neboť volba (224) neporuší platnost předpokladů lemmatu 27.

Věta 64. *Uvažujme metodu sdružených gradientů danou předpisem (165), kde $\beta_i = \beta_i^+$ je některá z hodnot (134)+(208), (146), (147). Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F2, F3 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -\left(1 + \beta_i^+ \frac{g_{i+1}^T s_i}{g_{i+1}^T g_{i+1}}\right) g_{i+1}, \qquad \beta_i^{(2)} = \beta_i^+,$$

kde β_i^+ je některá z hodnot (134)+(208), (146), (147) a předpokládejme, že $g_i \geq \underline{\varepsilon}$, $i \in N$. Použijeme-li vztahy (S3a), (164) a předpoklady F2, F3, dostaneme

$$|g_{i+1}^T s_i| \leq \varepsilon_3 |g_i^T s_i| = \varepsilon_3 g_i^T g_i \leq \varepsilon_3 \overline{g}^2 \tag{226}$$

Podle lemmatu 28 platí $|\beta_i^+| \leq |\beta_i| \leq C_2 \|d_i\|$, což spolu s (226) dává

$$\|s_{i+1}^{(1)}\| \leq \left(1 + |\beta_i^+| \frac{|g_{i+1}^T s_i|}{g_{i+1}^T g_{i+1}}\right) \|g_{i+1}\| \leq \left(1 + C_2 D \frac{\varepsilon_3 \overline{g}^2}{\underline{\varepsilon}^2}\right) \overline{g} \triangleq C_1.$$

Jsou tedy splněny předpoklady lemmatu 24, takže platí (215), což je podle poznámky 90 ve sporu s nerovností (35). □

Věta 65. *Uvažujme metodu sdružených gradientů danou předpisem (172), kde $\beta_i = \beta_i^+$ je některá z hodnot (134)+(208), (146) a ζ_i je hodnota určená vztahem (173) s $p_i = y_i$, přičemž $\zeta_i = 0$, pokud $\beta_i^+ = 0$. Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F2, F3 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -g_{i+1} - \zeta_i y_i, \qquad \beta_i^{(2)} = \beta_i^+$$

a předpokládejme, že $g_i \geq \underline{\varepsilon}$, $i \in N$.

(a) Nechť β_i^+ je některá z hodnot (146) a ζ_i je hodnota určená podle vzorce (173) s $p_i = y_i$, přičemž $\zeta_i = 0$, pokud $\beta_i^+ = 0$. Pak, pokud $\zeta_i \neq 0$, platí (174), Nerovnost (226) spolu s nerovnostmi uvedenými v části (b) důkazu lemmatu 28 implikují, že pro libovolnou hodnotu ζ_i, určenou podle vzorců (174), platí

$$|\zeta_i| \leq \frac{\varepsilon_3 \overline{g}^2}{(1 - \varepsilon_2) \underline{s}\, \underline{\varepsilon}^2}. \tag{227}$$

(b) Nechť β_i^+ je některá z hodnot (134)+(208), a ζ_i je hodnota určená podle vzorce (173) s $p_i = y_i$, přičemž $\zeta_i = 0$, pokud $\beta_i^+ = 0$. Pak, pokud $\zeta_i \neq 0$, použitím nerovností (212) a (227) dostaneme

$$|\zeta_i| \leq C, \qquad C = \frac{\varepsilon_3 \overline{g}^2}{(1-\eta)(1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2}.$$

(c) Použijeme-li (a) a (b), můžeme psát

$$\|s_{i+1}^{(1)}\| \leq \|g_{i+1}\| + |\zeta_i|\|y_i\| \leq \|g_{i+1}\| + C\|g_{i+1} - g_i\| \leq (1+2C)\overline{g} \stackrel{\Delta}{=} C_1.$$

Jelikož podle lemmatu 28 platí $|\beta_i^+| \leq |\beta_i| \leq C_2\|d_i\|$, jsou splněny předpoklady lemmatu 24, takže platí (215), což je podle poznámky 90 ve sporu s nerovností (35). □

Lemma 29. *Uvažujme spádovou metodu sdružených gradientů používající při výběru délky kroku slabou Wolfeho podmínku, předpokládejme že funkce $F : \mathcal{D} \to R$ splňuje podmínky F1, F2, F3 a označme β_i některou z hodnot (179), (180)+(208), (181), (182), (183)+(208), (184) (kromě hodnot β_i^{PRD}, β_i^{FRD}, β_i^{PPD}), kde $0 \leq \lambda_i \leq \overline{\lambda}$. Pak, pokud $g_i \geq \varepsilon$, $i \in N$, existuje konstanta $C > 0$ taková, že $|\beta_i| \leq C\|d_i\|$, $i \in N$.*

Důkaz (a) Použijeme-li nerovnosti $\|g_{i+1}\| \leq \overline{g}$, $\|g_i\| \leq \overline{g}$, $\|y_i\| = \|g(x_i + d_i) - g(x_i)\| \leq \overline{G}\|d_i\|$, $\|d_i\| \leq D$ a nerovnost (210), dostaneme

$$\|y_i\|^2 = \|g_{i+1} - g_i\|\|y_i\| \leq 2\overline{g}\overline{G}\|d_i\|,$$

$$g_{i+1}^T g_{i+1} \leq \frac{\overline{g}\overline{G}}{1-\eta}\|d_i\|,$$

$$\|y_i - d_i\|^2 = \|g_{i+1} - g_i - d_i\|\|y_i - d_i\| \leq (2\overline{g} + D)(\overline{G}+1)\|d_i\|$$

(b) Použijeme-li (S3a), můžeme psát

$$g_{i+1}^T s_i \geq \varepsilon_2 g_i^T s_i = -\varepsilon_2 y_i^T s_i + \varepsilon_2 g_{i+1}^T s_i,$$

neboli $(1-\varepsilon_2) g_{i+1}^T s_i \geq -\varepsilon_2 y_i^T s_i$, což po úpravě dává

$$g_{i+1} s_i \geq -\frac{\varepsilon_2}{1-\varepsilon_2} y_i^T s_i.$$

Jelikož $g_i^T s_i < 0$, můžeme psát $g_{i+1}^T s_i = y_i^T s_i + g_i^T s_i \leq y_i^T s_i$, což spolu s předchozí nerovností dává

$$\left|\frac{g_{i+1}^T s_i}{y_i^T s_i}\right| \leq \max\left(\frac{\varepsilon_2}{1-\varepsilon_2}, 1\right).$$

Podle (S3a) navíc platí $|g_{i+1}^T s_i / g_i^T s_i| \leq \varepsilon_3$.

(c) Uvažované hodnoty parametru β_i obsahují dva členy. První člen nabývá hodnot (133), (134)+(208), (135) a jeho absolutní hodnota splňuje podle lemmatu 25 nerovnost (223). Druhý člen je λ násobkem součinu dvou zlomků. První zlomek lze podle (a) a podle části (b) důkazu lemmatu 28 omezit výrazem $C_a\|d_i\|$, kde

$$C_a = \frac{(2\overline{g} + D)(\overline{G}+1)}{(1-\eta)(1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2}.$$

Druhý zlomek lze podle (b) omezit konstantou

$$C_b = \max\left(\frac{\varepsilon_2}{1-\varepsilon_2}, 1, \varepsilon_3\right).$$

Platí tedy

$$|\beta_i| \leq C\|d_i\|, \qquad C = C_2 + \overline{\lambda} C_a C_b, \qquad C_2 = \frac{\overline{g}(\overline{G}+1)}{(1-\eta)(1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2}. \tag{228}$$

□

V lemmatu 29 neuvažujeme hodnoty β_i^{PRD}, β_i^{FRD} a β_i^{PPD}. Je to proto, že absolutní hodnotu podílu $|g_i^T s_i|/g_i^T g_i$ nelze shora ohraničit, neboť nemáme k disposici nerovnost opačnou k (162) (například rovnost (164)).

Věta 66. *Uvažujme metodu sdružených gradientů danou předpisem* $s_{i+1} = \vartheta_i g_{i+1} + \beta_i^+ s_i$, $\beta_i^+ = \max(0, \beta_i)$, *kde:*

(a) β_i *je některá z hodnot (179), (180)+(208), (181) (kromě hodnot* β_i^{PRD}, β_i^{FRD}, β_i^{PPD}*), kde* $1/4 < \underline{\lambda} \le \lambda \le \overline{\lambda}$ *a* $\vartheta_i = 1$.

(b) β_i *je některá z hodnot (182), (183)+(208), (184) (kromě hodnot* β_i^{PRD}, β_i^{FRD}, β_i^{PPD}*), kde* $1/4 < \underline{\lambda} \le \lambda \le \overline{\lambda}$ *a* ϑ_i *je odpovídající hodnota určená podle (169), přičemž* $\vartheta_i = 1$, *pokud* $\beta_i^+ = 0$.

Pak, splňuje-li funkce $F : \mathcal{D} \to R$ *předpoklady F1, F2, F3 a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku, je tato metoda globálně konvergentní. Pokud* $\beta_i = \beta_i^{HSD}$, $\beta_i = \beta_i^{DYD}$, $\beta_i = \beta_i^{HPD}$, *stačí použít slabou Wolfeho podmínku.*

Důkaz Položme

$$s_{i+1}^{(1)} = -\vartheta_i g_{i+1}, \qquad \beta_i^{(2)} = \beta_i^+,$$

kde $\beta_i^+ = \max(0, \beta_i)$ a β_i je některá z hodnot uvedených v dokazovaném tvrzení, a předpokládejme, že $g_i \ge \underline{\varepsilon}$, $i \in N$. Podle lemmatu 22 je splněna nerovnost (176), takže uvažovaná metode je spádová.

(a) Jelikož $\vartheta_i = 1$, můžeme psát $\|s_{i+1}^{(1)}\| = \|g_{i+1}\|$ a podle lemmatu 29 platí $|\beta_i^{(2)}| \le C\|d_i\|$. Jsou tedy splněny předpoklady lemmatu 27, takže uvažovaná metoda je globálně konvergentní.

(b) Zřejmě $\vartheta_i^{HS} = 1$. Stačí tedy vyšetřit případ, kdy $\vartheta_i = \vartheta_i^{LS}$. Použijeme-li stejný postup jako v části (b) důkazu věty 61, dostaneme $0 < (1 - \varepsilon_2) \le \vartheta_i \le (1 + \varepsilon_3)$, takže $\|s_{i+1}^{(1)}\| = \vartheta_i \|g_{i+1}\| \le (1 + \varepsilon_3)\overline{g}$. Jelikož platí i (228), jsou splněny předpoklady lemmatu 27, takže uvažovaná metoda je globálně konvergentní. □

Věta 67. *Uvažujme metodu sdružených gradientů danou předpisem*

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \max(0, \beta_i^{KD})\, s_i - \zeta_i^{KD} y_i,$$

kde $0 < \underline{\rho} \le \rho_i \le \underline{\rho}$ *a* $0 \le \mu_i \le \overline{\mu} \le 1$. *Pak, splňuje-li funkce* $F : \mathcal{D} \to R$ *předpoklady F1, F2, F3 a používáme-li při výběru délky kroku slabou Wolfeho podmínku, je tato metoda globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -g_{i+1} - \zeta_i^{KD} y_i, \qquad \beta_i^{(2)} = \max(0, \beta_i^{KD}),$$

kde β_i^{KD} a ζ_i^{KD} jsou hodnoty určené podle vzorce (200), a předpokládejme, že $g_i \ge \underline{\varepsilon}$, $i \in N$. Druhý vzorec v (200) se liší od prvního vzorce v (174) pouze tím, že parametr ζ_i je vynásoben číslem $\mu_i \le 1$. Podle (227) tedy existuje konstanta $C_1 > 0$ taková, že $\|s_{i+1}^{(1)}\| \le C_1$. Položme $\alpha_i \rho_i = \lambda_i y_i^T y_i / y_i^T s_i$. Pak první vzorec v (200) se od prvního vzorce v (179) liší pouze tím, že hodnota $\lambda_i > 0$ je nahražena hodnotou $\lambda_i + 1 > 1$. Podle lemmatu 29 tedy existuje konstanta $C > 0$ taková, že $|\beta_i^{(2)}| \le C\|d_i\|$. Jsou tedy splněny předpoklady lemmatu 27, takže uvažovaná metoda je globálně konvergentní. □

Věta 68. *Uvažujme metodu sdružených gradientů danou předpisem*

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \max(0, \beta_i^{DK})\, d_i,$$

kde $\rho_i > 0$ *je některá z hodnot (195) s* $\lambda_i \ge 1$. *Pak, splňuje-li funkce* $F : \mathcal{D} \to R$ *předpoklady F1, F4, F5 a používáme-li při výběru délky kroku slabou Wolfeho podmínku, je tato metoda globálně konvergentní.*

Důkaz Položme

$$s_{i+1}^{(1)} = -g_{i+1}, \qquad \beta_i^{(2)} = \max(0, \beta_i^{DK}),$$

kde β_i^{DK} je hodnota určená podle vzorce (205), a předpokládejme, že $g_i \geq \underline{\varepsilon}$, $i \in N$. Hodnota β_i^{DK} se od hodnoty β_i^{KD} liší tím, že obsahuje nový člen, pro který platí

$$\left|\frac{s_i^T g_{i+1}}{s_i^T s_i}\right| = \left|\frac{d_i^T g_{i+1}}{y_i^T s_i}\right| \frac{y_i^T d_i}{d_i^T d_i} \leq \frac{\overline{g}\|d_i\|}{y_i^T s_i}\overline{G} \leq \frac{\overline{g}\overline{G}}{(1-\varepsilon_2)\underline{s}\,\underline{\varepsilon}^2}\|d_i\| \triangleq C_3\|d_i\|$$

(používame nerovnosti z části (b) důkazu lemmatu 28). Použijeme-li odhad $|\beta_i^{KD}| \leq C\|d_i\|$, získaný v důkazu věty 67, dostaneme

$$|\beta_i^{(2)}| \leq |\beta_i^{KD}| + \left|\frac{s_i^T g_{i+1}}{s_i^T s_i}\right| \leq (C + C_3)\|d_i\|.$$

Jelikož $\|s_{i+1}^{(1)}\| = \|g_{i+1}\| \leq \overline{g}$, jsou splněny předpoklady lemmatu 27, takže uvažovaná metoda je globálně konvergentní. □

3.6 Implementace metod sdružených gradientů

Existuje řada dalších modifikací základních metod sdružených gradientů, z nichž některé se jen nepatrně liší od modifikací uvedených v oddílu 3.4 a jiné, ač teoreticky podložené, nejsou efektivní pro praktické výpočty. Jedna z poměrně účinných úprav, založená na myšlenkách podobných těm, které vedou na (190), používá parametry

$$\begin{aligned}
\beta_i^{HSY} &= \frac{\|g_{i+1}\|^2 - |g_{i+1}^T g_i|}{y_i^T s_i + \mu_i \max(0, g_{i+1}^T s_i)}, \\
\beta_i^{PRY} &= \frac{\|g_{i+1}\|^2 - |g_{i+1}^T g_i|}{g_i^T g_i + \mu_i \max(0, g_{i+1}^T s_i)}, \\
\beta_i^{LSY} &= \frac{\|g_{i+1}\|^2 - |g_{i+1}^T g_i|}{|g_i^T s_i| + \mu_i \max(0, g_{i+1}^T s_i)}
\end{aligned} \tag{229}$$

(Y – Yu, Zhao, Wei [171]). Použití absolutní hodnoty v čitatelích je podstatné. Odstraníme-li absolutní hodnotu, dostaneme metody (133) s upravenými jmenovateli, které jsou méně účinné než metody (133) s původními jmenovateli.

Další metody sdružených gradientů lze získat použitím zobecněných kvazinewtonovských podmínek. Zobecněné kvazinewtonovské podmínky jsou podrobně studovány v oddílu 4.8. Zde uvedeme pouze základní myšlenky. Ukážeme, jak lze zobecnit metodu DL, která byla odvozena pomocí standardní kvazinewtonovské podmínky $H_{i+1}y_i = d_i$.

Poznámka 91. Jednou z možností je použít kvazinewtonovskou podmínku $H_{i+1}\tilde{y}_i = d_i$, kde $\tilde{y}_i = y_i + \tau_i d_i$. Pak

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \tilde{\beta}_i^{DL} s_i \quad \text{pro} \quad i \in N,$$

kde

$$\tilde{\beta}_i^{DL} = \frac{\tilde{y}_i^T g_{i+1} - d_i^T g_{i+1}}{\tilde{y}_i^T s_i} = \tilde{\beta}_i^{HS} - \frac{d_i^T g_{i+1}}{\tilde{y}_i{}^T s_i} \tag{230}$$

Abychom dostali vhodné korekce, můžeme (tak jako v poznámce 182) použít některou z hodnot

$$\tau_i = \frac{2(F_i - F_{i+1}) + d_i^T g_{i+1} + d_i^T g_i}{\|d_i\|^2}, \tag{231}$$

$$\tau_i = \frac{6(F_i - F_{i+1}) + 3(d_i^T g_{i+1} + d_i^T g_i)}{\|d_i\|^2}. \tag{232}$$

Poznamenejme, že jmenovatel $d_i^T \tilde{y}_i$ v (230) je kladný, pokud $\tau_i > -d_i^T y_i / d_i^T d_i$, takže je výhodné parametr τ_i zvětšit (například položit $\tau_i = 0$), není-li tato nerovnost splněna.

Poznámka 92. Další možností je použít vícekrokovou kvazinewtonovskou podmínku $H_{i+1}\hat{y}_i = \hat{d}_i$, kde $\hat{y}_i = y_i + \lambda_i y_{i-1}$, $\hat{d}_i = d_i + \lambda_i d_{i-1}$ a $\lambda_i = 1/(\tau_i(\tau_i + 2))$ (poznámka 183). Pak

$$s_1 = -g_1 \quad \text{a} \quad s_{i+1} = -g_{i+1} + \hat{\beta}_i^{DL} s_i \quad \text{pro} \quad i \in N,$$

kde

$$\hat{\beta}_i^{DL} = \frac{\hat{y}_i^T g_{i+1} - \hat{d}_i^T g_{i+1}}{\hat{y}_i^T s_i} = \hat{\beta}_i^{HS} - \frac{\hat{d}_i^T g_{i+1}}{\hat{y}_i{}^T s_i}. \tag{233}$$

Parametr τ_i lze volit například podle vzorce $\tau_i = \|d_{i-1}\|/\|d_i\|$.

Poznámka 93. Účinnost metod sdružených gradientů lze zvýšit vhodným přerušováním iteračního procesu. Přerušování se provádí tak, že se po výpočtu směrového vektoru testuje splnění předepsané podmínky. Není-li tato podmínka splněna, nahradí se vypočtený směrový vektor záporně vzatým gradientem (což odpovídá volbě $\beta_i = 0$). Velmi vhodné je použít podmínku stejnoměrné spádovosti (S1b) a iterační proces přerušit, pokud neplatí

$$-g_{i+1}^T s_{i+1} \geq \varepsilon_0 \|g_{i+1}\| \|s_{i+1}\|, \tag{234}$$

kde $\varepsilon_0 > 0$ je nějaké malé číslo (například $\varepsilon_0 = 10^{-8}$). V poznámce 73 je uvedeno, že takto upravená metoda sdružených gradientů je globálně konvergentní, aniž by k přerušení docházelo příliš často. Používáme-li parametry (134), (180), (183), je výhodné testovat ortogonalitu gradientů. V tomto případě se iterační proces přeruší, pokud neplatí

$$|g_i^T g_{i+1}| \leq \eta_2 \|g_{i+1}\| \|g_i\|. \tag{235}$$

kde hodnota η_2 závisí na zvolené Wolfeho podmínce. Také je možné testovat sdruženost směrových vektorů. V tomto případě se iterační proces přeruší, pokud neplatí

$$|y_i^T s_{i+1}| \leq \eta_1 \|s_{i+1}\| \|y_i\|, \tag{236}$$

kde hodnota η_1 závisí na zvolené Wolfeho podmínce. Je-li počet proměnných dostatečně velký, vyplatí se v případě parametrů (134), (180), (183) iterační proces přerušovat vždy po n krocích, počítaných od posledního přerušení (pak jsou splněny předpoklady věty 32).

Poznámka 94. Metody sdružených gradientů jsou velmi citlivé na výběr délky kroku. Použití slabé Wolfeho podmínky se standardními parametry $\varepsilon_1 = 0.0001$, $\varepsilon_2 = 0.9$ a $\varepsilon_3 = \infty$ je neefektivní. Mnohem výhodnější je používat silnou Wolfeho podmínku s parametry $\varepsilon_1 = 0.0001$ a $\varepsilon_3 = \varepsilon_2 = 0.1$. Kromě toho velmi záleží na volbě počáteční délky kroku α_i^1. Je výhodné pokládat

$$\alpha_i^1 = \min(1,\, 2(F_i - F_{i-1})/s_i^T g_i,\, 2(\underline{F} - F_i)/s_i^T g_i),$$

kde $\underline{F}$ je dolní odhad pro minimální hodnotu funkce F. Další hodnoty α_i^j, $j > 1$, se určují pomocí kubické extrapolace nebo interpolace. Metody sdružených gradientů negenerují tak kvalitní směrové vektory jako Newtonova metoda nebo metody s proměnnou metrikou a zaokrouhlovací chyby způsobují, že je někdy obtížné získat řešení s požadovanou přesností. Proto je třeba výběr délky kroku upravit použitím složitějších zastavovacích kriterií než jsou Wolfeho podmínky. Velmi efektivní proceduru pro výběr délky kroku, která je součástí programu CG–DESCENT, vyvinuli Hager a Zhang [79]. Tato procedura byla použita při testování metod sdružených gradientú v oddílu 3.7.

Algoritmus metody sdružených gradientů lze popsat zhruba takto:

Algoritmus 5. Data $\varepsilon_0 = 10^{-8}$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-1}$, $\eta_1 = 0.05$, $\eta_2 = 0.5$, $\underline{\varepsilon} > 0$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$, vypočteme $F_1 := F(x_1)$, $g_1 := g(x_1)$ a položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \leq \underline{\varepsilon}$ ukončíme výpočet. V opačném případě určíme směrový vektor pomocí zvolené metody sdružených gradientů a rozhodneme o přerušení iteračního procesu podle pokynů uvedených v poznámce 93. Rozhodneme-li se pro škálování, určíme škálovací koeficient $\gamma_i > 0$ podle poznámky 67 (obvykle se škálování neprovádí, takže $\gamma_i = 1$). Pokud $\gamma_i \neq 1$ vynásobíme směrový vektor s_i číslem γ_i.

Krok 3 Určíme délku kroku α_i tak aby byla splněna silná Wolfeho podmínka. Položíme $x_{i+1} := x_i + \alpha_i s_i$, vypočteme $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$.

Krok 4 Zvětšíme i o 1 a přejdeme na krok 2.

3.7 Numerické porovnání metod sdružených gradientů

V tomto oddílu uvedeme výsledky numerických testů, jejichž cílem je ukázat účinnost jednotlivých metod sdružených gradientů. Vybrali jsme pouze modifikace, které se jevily nejúčinnější. Ostatní metody byly také testovány, ale výsledky těchto testů nejsou uvedeny ve výsledných tabulkách.

K testování metod sdružených gradientů bylo použito 73 testovacích úloh s 10000 proměnnými (TEST12 z oddílu 1.5), které jsou uvedeny v [3] a lze je stáhnout z `camo.ici.ro/neculai/ansoft.htm`). Výsledky testů jsou prezentovány v první tabulce, která obsahuje celkové počty iterací NIT, funkčních hodnot NFV, gradientů NFG, jakož i celkový čas výpočtu. K výběru délky kroku byla použita procedura převzatá z programu CG-DESCENT.

K označení jednotlivých metod sdružených gradientů byly zvoleny řetězce znaků, kde M značí typ metody (za M se dosazuje HS, PR, LS, DY, FR, CD, HP, PP, LP) a kde význam ostatních znaků je uveden v následujícím seznamu

M	-	Základní metody (127), (133), (134), (135).
MS	-	Spádové metody používající vzorec (165).
M I	-	Modifikovanné metody používající vzorce (166), (167), (169).
MT	-	Spádové metody používající vzorce (172), (173), (174).
MD	-	Spádové metody používající vzorce (127), (179), (180), (181).
MD I	-	Modifikované spádové metody používající vzorce (166), (167), (169),(182), (183), (184).
ML	-	Metody typu Dai-Liao používající vzorce (127), (194).
MM	-	Modifikovanné metody používající vzorce (127), (190)

Znaménko + značí, že se hodnota β_i nahražuje číslem $\max(0, \beta_i)$.

Testovací úlohy z [3] jsou sice rozsáhlé, ale méně obtížné. Proto jsme použili další sadu 73 testovacích funkcí o 1000 proměnných (TEST25 z oddílu 1.5), které jsou uvedeny v [106] (9 úloh ze sbírky TEST25 bylo vynecháno, protože je některá z testovaných metod nevyřešila). Výsledky nových testů jsou uvedeny v druhé tabulce, jejíž obsah má podobný význam jako obsah předchozí tabulky. V druhé tabulce nejsou uvedeny metody DY, FR, CD a jejich modifikace, neboť nedokázaly nalézt řešení všech 73 problémů.

Metoda	Metody typu HS NIT - NFV - NFG	čas	Metody typu PR NIT - NFV - NFG	čas	Metody typu LS NIT - NFV - NFG	čas
M	73500 - 146562 - 82875	45.5	97522 - 153458 - 94111	52.5	90844 - 182707 - 98993	59.2
M+	64776 - 130153 - 70449	42.2	99012 - 199048 - 105904	52.2	109072 - 217871 - 122362	59.1
MS+	64267 - 127877 - 69952	39.4	81135 - 162484 - 85922	46.9	98472 - 197386 - 104029	54.6
M I+	64776 - 130153 - 70449	42.2	59242 - 118194 - 64077	37.5	92908 - 185231 - 97328	49.8
MT+	54197 - 109310 - 60435	38.2	78887 - 154099 - 98045	48.4	60850 - 122040 - 65886	38.4
MD+	63923 - 128143 - 69497	42.0	93105 - 187343 - 99545	51.3	70265 - 140260 - 74808	41.7
MD I+	63923 - 128143 - 69497	42.0	71623 - 140667 - 91332	46.0	83138 - 161521 - 101357	49.0
ML+	65197 - 130629 - 71040	42.2	76842 - 148778 - 97018	46.3	77947 - 156497 - 83805	46.3
MM	62598 - 124900 - 67430	39.0	63481 - 126236 - 68753	38.6	63597 - 126651 - 68208	37.7
Metoda	Metody typu DY NIT - NFV - NFG	čas	Metody typu FR NIT - NFV - NFG	čas	Metody typu CD NIT - NFV - NFG	čas
M	72624 - 145100 - 78735	47.1	81152 - 162513 - 85939	48.0	87805 - 176088 - 92754	63.7
MS	85372 - 161985 - 105303	57.9	84886 - 170639 - 89805	68.1	69839 - 140434 - 74992	42.2
M I	72624 - 145100 - 78735	47.1	70155 - 141153 - 75368	49.1	83105 - 166870 - 88196	49.6
MT	85249 - 169741 - 95273	51.1	84001 - 175873 - 97099	63.8	88634 - 184105 - 102816	76.1
MD+	84267 - 170722 - 90918	52.5	82341 - 164020 - 88737	61.3	75449 - 151144 - 80078	46.6
MD I+	84267 - 170722 - 90918	52.5	81187 - 164149 - 87045	66.3	80027 - 161857 - 86512	55.1
Metoda	Metody typu HP NIT - NFV - NFG	čas	Metody typu PP NIT - NFV - NFG	čas	Metody typu LP NIT - NFV - NFG	čas
M	94217 - 189553 - 105896	99.9	98579 - 195634 - 112126	52.3	89764 - 168900 - 115584	55.3
M+	75175 - 150631 - 81148	46.9	65729 - 132372 - 72109	40.6	85626 - 164338 - 106336	48.9
MS+	63356 - 126299 - 67971	39.6	65561 - 131168 - 70887	41.6	84874 - 170016 - 90408	50.0
M I+	75175 - 150631 - 81148	47.0	66181 - 133055 - 76592	43.6	68377 - 136899 - 73861	43.8
MT+	53245 - 107591 - 59446	39.3	77128 - 155885 - 84388	54.2	73569 - 147557 - 78856	44.7
MD+	67298 - 134304 - 72828	43.8	68450 - 138780 - 76567	44.3	68501 - 138216 - 75175	43.7
MD I+	67298 - 134304 - 72828	43.8	71206 - 143152 - 80100	47.7	69167 - 138243 - 75678	44.9

Tabulka 1: TEST12 – 73 úloh

Metoda	Metody typu HS NIT - NFV - NFG	čas	Metody typu PR NIT - NFV - NFG	čas	Metody typu LS NIT - NFV - NFG	čas
M	180415 - 356820 - 189289	44.3	192559 - 378024 - 205888	48.9	183373 - 358893 - 198523	47.6
M+	177404 - 350275 - 187121	44.3	193256 - 380830 - 205681	47.6	168082 - 332005 - 178585	37.8
MS+	172989 - 341748 - 181636	44.2	176074 - 347721 - 185784	44.5	178542 - 350962 - 189409	46.0
M I+	177404 - 350275 - 187121	44.4	188245 - 369774 - 202312	48.3	177415 - 350384 - 187819	45.9
MT+	169096 - 333431 - 176767	38.7	166449 - 328628 - 175692	35.0	171611 - 335298 - 184615	38.6
MD+	184723 - 363463 - 195516	47.8	196565 - 386410 - 209068	49.1	179928 - 355002 - 189560	44.9
MD I+	184723 - 363463 - 195516	47.8	180829 - 356770 - 190564	45.4	182770 - 359600 - 194898	47.4
ML+	180151 - 356348 - 189308	44.5	169269 - 335000 - 179524	37.2	177867 - 351356 - 189243	45.0
MM	172571 - 338998 - 183992	45.3	187137 - 369122 - 201579	47.4	186313 - 367725 - 198195	45.5
Metoda	Metody typu HP NIT - NFV - NFG	čas	Metody typu PP NIT - NFV - NFG	čas	Metody typu LP NIT - NFV - NFG	čas
M	177550 - 350250 - 188436	45.2	181377 - 357178 - 194646	46.9	190279 - 374119 - 203658	47.5
M+	170130 - 336539 - 178231	43.5	170629 - 336368 - 182559	37.7	198182 - 387620 - 213258	50.0
MS+	181742 - 358446 - 191016	45.5	181792 - 358098 - 191930	47.5	179386 - 353043 - 190938	45.8
M I+	170130 - 336539 - 178231	43.5	171369 - 338888 - 180437	44.1	177134 - 350693 - 184418	44.0
MT+	169960 - 336307 - 178765	39.5	166228 - 328024 - 176246	35.4	171784 - 335591 - 185189	38.4
MD+	187718 - 370074 - 198195	47.3	184662 - 367014 - 195265	45.8	187866 - 371548 - 197882	46.0
MD I+	187718 - 370074 - 198195	47.3	188003 - 369635 - 202390	48.8	190851 - 374267 - 203765	50.3

Tabulka 2: TEST25 – 73 úloh

Z údajů uvedených v těchto tabulkách lze vyvodit několik závěrů:

- Metody typu HS se zdají být poněkud účinnější než metody typu PR a LS. Metody typu DY, FR a CD se svou výkonností příliš neliší, což platí i pro metody typu HP, PP a LP.
- Metody typu HS se zdají být poněkud účinnější než metody HP. Metody PR a PP se svou výkonností příliš neliší, což platí i pro metody typu LS a LP.
- Metody typu DY, FR a CD dávají horší výsledky než ostatní metody. Testy uvedené v první tabulce vyžadovaly častější přerušování iteračního procesu, zejména kvůli ztrátě konjugovanosti. Tyto metody nevyřešily tři probémy ze sbírky použité k sestavení druhé tabulky (proto nejsou v této tabulce obsaženy).
- Modifikace M+ dává většinou lepší výsledky než základní metoda (s výjimkou PR a LS). Ještě lepší výsledky dává modifikace MS+ (opět s výjimkou PR a LS) a zejména modifikace MT+, která se zdá být nejefektivnější.
- Modifikace MS a MT nejsou vhodné pro metody DY, FR, a CD. Pro tyto metody se vyplácí používat modifikaci MI.
- Pro velmi rozsáhlé úlohy jsou metody sdružených gradientů poměrně spolehlivé a účinné.

3.8 Předpodmíněná metoda sdružených gradientů pro řešení soustav lineárních rovnic

Podle vět 40 a 49 je metoda sdružených gradientů zvláště vhodná k hledání minima ryze konvexní kvadratické funkce nebo, což je totéž, pro řešení soustavy lineárních rovnic se symetrickou pozitivně definitní maticí. Nyní budeme uvažovat kvadratickou funkci

$$Q(s) = g^T s + \frac{1}{2} s^T B s, \tag{237}$$

která se používá k určení směrového vektoru v metodách spádových směrů (i v metodách s lokálně omezeným krokem popsaných v páté kapitole). V poznámce 66 jsme ukázali, že metodu sdružených gradientů lze předpodmínit tak, že se místo kvadratické funkce $Q(s)$ minimalizuje kvadratická funkce

$$\tilde{Q}(\tilde{s}) = \tilde{g}^T \tilde{s} + \frac{1}{2} \tilde{s}^T \tilde{B} \tilde{s}, \tag{238}$$

kde $\tilde{s} = C^{1/2} s$, $\tilde{g} = C^{-1/2} g$ a $\tilde{B} = C^{-1/2} B C^{-1/2}$ (v poznámce 66 bylo použito označení $H = C^{-1}$). Matice C se vybírá tak, aby soustava lineárních rovnic $\tilde{B}\tilde{s} = -\tilde{g}$, definující minimum kvadratické funkce $\tilde{Q}(\tilde{s})$, byla co nejlépe podmíněná. Pokud $C \approx B$, platí $\tilde{B} \approx I$ a $\kappa(\tilde{B}) \approx 1$, což podle věty 49 zaručuje rychlou konvergenci metody.

Definice 35. *Nechť $B \in R^{n\times n}$, $C \in R^{n\times n}$ jsou symetrické pozitivně definitní matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \qquad g_1 = g, \qquad p_1 = -C^{-1} g$$

a

$$\begin{aligned} q_i &= B p_i, & \alpha_i &= -p_i^T g_i / p_i^T q_i, \\ s_{i+1} &= s_i + \alpha_i p_i, & g_{i+1} &= g_i + \alpha_i q_i, \\ \beta_i = g_{i+1}^T C^{-1} g_{i+1} / g_i^T C^{-1} g_i, & & p_{i+1} &= -C^{-1} g_{i+1} + \beta_i p_i \end{aligned}$$

pro $1 \le i \le n$, nazveme předpodmíněnou metodou sdružených gradientů (s předpodmiňovačem C) pro řešení soustavy lineárních rovnic $Bs = -g$.

Poznámka 95. Některé vlastnosti nepředpodmíněné metody sdružených gradientů jsou ukázány v důkazu věty 40 (vztahy (129)–(131)). Analogické vztahy pro předpodmíněnou metodu sdružených gradientů

dostaneme formálně tak, že místo s, g, p, q a B dosazujeme $\tilde{s} = C^{1/2}s$, $\tilde{g} = C^{-1/2}g$, $\tilde{p} = C^{1/2}p$, $\tilde{q} = C^{-1/2}q$ a $\tilde{B} = C^{-1/2}BC^{-1/2}$. Platí

$$p_j^T g_i = 0, \tag{239}$$

$$g_j^T C^{-1} g_i = 0, \tag{240}$$

$$p_j^T B p_i = 0 \tag{241}$$

pro $1 \leq j < i \leq m$, kde m je index takový, že $p_i^T B p_i > 0$ pro $1 \leq i \leq m$. Tento postup budeme používat i nadále. Nejprve zformulujeme a dokážeme tvrzení pro $C = I$ a pak jako důsledek uvedeme tvrzení pro $C \neq I$.

Poznámka 96. Z definice 35 plyne, že

$$g_i = Bs_i + g, \quad \text{kde} \quad s_i = \sum_{j=1}^{i-1} \alpha_j p_j, \tag{242}$$

což spolu s (239) dává

$$s_i^T(Bs_i + g) = s_i^T g_i = \sum_{j=1}^{i-1} \alpha_j p_j^T g_i = 0. \tag{243}$$

Poznámka 97. Použijeme-li vztah (239), dostaneme $g_i^T C^{-1} g_i = (\beta_i p_{i-1} - p_i) = -p_i^T g_i$, takže

$$\alpha_i = \frac{g_i^T C^{-1} g_i}{p_i^T B p_i}. \tag{244}$$

Tuto hodnotu budeme používat při vyšetřování vlastností předpodmíněných metod sdružených gradientů. Pro praktické použití je však výhodnější hodnota uvedená v defimici 35, která je méně citlivá vzhledem k zaokrouhlovacím chybám. Hodnota α_i realizuje přesný výběr délky kroku, neboť z efinice 35 a vztahu (239) plyne

$$p_i^T g_{i+1} = p_i^T g_i + \alpha_i p_i^T q_i = p_i^T g_i - \frac{p_i^T g_i}{p_i^T q_i} p_i^T q_i = 0.$$

Jelikož podle efinice 35 a vztahu (239) platí

$$-p_i^T g_i = -p_i^T \left(g + \sum_{j=1}^{i-1} \alpha_j B p_j \right) = -p_i^T g,$$

můžeme psát

$$\alpha_i = -\frac{p_i^T g_i}{p_i^T B p_i} = -\frac{p_i^T g}{p_i^T B p_i}. \tag{245}$$

Nejprve dokážeme monotonnost změny některých důležitých skalárních veličin (vzorec (249) je uveden v práci [32]). Budeme přitom používat označení $s_i(\alpha) = s_i + \alpha p_i$ pro $0 \leq \alpha \leq \alpha_i$, takže $s_{i+1} = s_i(\alpha_i)$.

Věta 69. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s $C = I$ na kvadratickou funkci $Q(s)$ a platí-li $g_j^T g_j > 0$, $p_j^T B p_j > 0$ pro $1 \leq j \leq i$, můžeme pro $0 < \alpha < \alpha_i$ psát*

$$Q(s_{i+1}) < Q(s_i(\alpha)) < Q(s_i), \tag{246}$$

$$g^T s_{i+1} < g^T s_i(\alpha) < g^T s_i, \tag{247}$$

$$\|s_{i+1}\| > \|s_i(\alpha)\| > \|s_i\|, \tag{248}$$

$$\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} > \frac{g^T s_i(\alpha)}{\|g\|\|s_i(\alpha)\|} > \frac{g^T s_i}{\|g\|\|s_i\|}. \tag{249}$$

Důkaz (a) Platí

$$\begin{aligned} Q(s_i(\alpha)) &= g^T(s_i+\alpha p_i)+\frac{1}{2}(s_i+\alpha p_i)^T B(s_i+\alpha p_i) \\ &= Q(s_i)+\alpha(g+Bs_i)^T p_i+\frac{1}{2}\alpha^2 p_i^T B p_i \\ &= Q(s_i)-\alpha g_i^T g_i+\frac{1}{2}\alpha^2 p_i^T B p_i, \end{aligned}$$

neboť $g+Bs_i=g_i$ podle (242) a $g_i^T p_i=-g_i^T g_i+\beta_{i-1}g_i^T p_{i-1}=-g_i^T g_i$ podle (239). Jelikož $p_i^T B p_i>0$, je kvadratická funkce $Q(s_i(\alpha))$ ryze konvexní. Její derivace $Q'(s_i(\alpha))=-g_i^T g_i+\alpha p_i^T B p_i$ je záporná pro $0\le\alpha<\alpha_i$ a nulová pro $\alpha=\alpha_i$, takže funkce $Q(s_i(\alpha))$ klesá pro $0\le\alpha<\alpha_i$ a nabývá svého minima pro $\alpha=\alpha_i$. Pro $\alpha=\alpha_i$ podle (244) s $C=I$ platí

$$\begin{aligned} Q(s_{i+1}) &= Q(s_i)-\alpha_i g_i^T g_i+\frac{1}{2}\alpha_i^2 p_i^T B p_i = Q(s_i)-\frac{(g_i^T g_i)^2}{p_i^T B p_i}+\frac{1}{2}\frac{(g_i^T g_i)^2}{p_i^T B p_i} \\ &= Q(s_i)-\frac{1}{2}\frac{(g_i^T g_i)^2}{p_i^T B p_i}. \end{aligned}$$

(b) Jelikož z (239)–(240) plyne

$$g_j^T p_i=-g_j^T g_i+\beta_{i-1}g_j^T p_{i-1}=\left(\prod_{k=j}^{i-1}\beta_k\right)g_j^T p_j=-\frac{g_i^T g_i}{g_j^T g_j}g_j^T(g_j-\beta_{j-1}p_{j-1})=-g_i^T g_i$$

pro $1\le j<i$ a jelikož $g=g_1$, můžeme psát

$$g^T s_i(\alpha)=g^T s_i+\alpha g^T p_i=g^T s_i-\alpha g_i^T g_i.$$

Tato lineární funkce klesá pro $\alpha\ge 0$.

(c) Použijeme-li vztah (242), dostaneme

$$\begin{aligned} s_i(\alpha)^T s_i(\alpha) &= (s_i+\alpha p_i)^T(s_i+\alpha p_i)=s_i^T s_i+\alpha^2 p_i^T p_i+2\alpha s_i^T p_i \\ &= s_i^T s_i+\alpha^2 p_i^T p_i+2\alpha\sum_{j=1}^{i-1}\alpha_j p_j^T p_i \\ &= s_i^T s_i+\alpha^2 p_i^T p_i+2\alpha g_i^T g_i\sum_{j=1}^{i-1}\frac{p_j^T p_j}{p_j^T B p_j}, \end{aligned}$$

neboť pro $1\le j<i$ platí $\alpha_j=g_j^T g_j/p_j^T B p_j$ a

$$p_j^T p_i=p_j^T(-g_i+\beta_{i-1}p_{i-1})=\beta_{i-1}p_j^T p_{i-1}=\left(\prod_{k=j}^{i-1}\beta_k\right)p_j^T p_j=\frac{g_i^T g_i}{g_j^T g_j}p_j^T p_j.$$

Tato kvadratická funkce roste pro $\alpha\ge 0$.

(d) Pro $1\le j\le i$ můžeme psát

$$\frac{p_j}{\|g_j\|^2}=-\frac{g_j}{\|g_j\|^2}+\frac{p_{j-1}}{\|g_{j-1}\|^2}=-\sum_{k=1}^{j}\frac{g_k}{\|g_k\|^2},$$

takže

$$\begin{aligned}
-s_i(\alpha) &= -\sum_{j=1}^{i-1}\alpha_j p_j - \alpha p_i = \sum_{j=1}^{i-1}\alpha_j\|g_j\|^2\left(\sum_{k=1}^{j}\frac{g_k}{\|g_k\|^2}\right)+\alpha\|g_i\|^2\left(\sum_{k=1}^{i}\frac{g_k}{\|g_k\|^2}\right)\\
&= \alpha_1\|g_1\|^2\left(\frac{g_1}{\|g_1\|^2}\right)+\alpha_2\|g_2\|^2\left(\frac{g_1}{\|g_1\|^2}+\frac{g_2}{\|g_2\|^2}\right)+\cdots+\alpha\|g_i\|^2\left(\frac{g_1}{\|g_1\|^2}+\frac{g_2}{\|g_2\|^2}+\cdots+\frac{g_i}{\|g_i\|^2}\right)\\
&= \sum_{j=1}^{i}\left(\sum_{k=j}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2\right)\frac{g_j}{\|g_j\|^2}.
\end{aligned}$$

Použijeme-li (240) s $C = I$, dostaneme

$$s_i^T(\alpha)s_i(\alpha) = \sum_{j=1}^{i}\left(\sum_{k=j}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2\right)^2\frac{1}{\|g_j\|^2}$$

a

$$-g^T s_i(\alpha) = \sum_{k=1}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2,$$

takže

$$\frac{s_i^T(\alpha)s_i(\alpha)}{(g^T s_i(\alpha))^2} = \sum_{j=1}^{i}\left(\frac{\sum_{k=j}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2}{\sum_{k=1}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2}\right)^2\frac{1}{\|g_j\|^2}.$$

Nyní použijeme toho, že racionální funkce $\varphi(t) = (a+t)/(b+t)$ je pro $a < b$ rostoucí (můžeme se o tom přesvědčit derivováním). Pro $0 < \alpha < \alpha_i$ tedy platí

$$\frac{\sum_{k=j}^{i}\alpha_k\|g_k\|^2}{\sum_{k=1}^{i}\alpha_k\|g_k\|^2} > \frac{\sum_{k=j}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2}{\sum_{k=1}^{i-1}\alpha_k\|g_k\|^2+\alpha\|g_i\|^2} > \frac{\sum_{k=j}^{i-1}\alpha_k\|g_k\|^2}{\sum_{k=1}^{i-1}\alpha_k\|g_k\|^2},$$

což po dosazení dává

$$\frac{s_{i+1}^T s_{i+1}}{(g^T s_{i+1})^2} > \frac{s_i(\alpha)^T s_i(\alpha)}{(g^T s_i(\alpha))^2} > \frac{s_i^T s_i}{(g^T s_i)^2}$$

Bezprostředním použitím této nerovnosti dostaneme (249). □

Důsledek 4. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s pozitivně definitní maticí C na kvadratickou funkci $Q(s)$ a platí-li $g_j^T C^{-1} g_j > 0$, $p_j^T B p_j > 0$ pro $1 \le j \le i$, můžeme pro $0 < \alpha < \alpha_i$ psát*

$$Q(s_{i+1}) < Q(s_i(\alpha)) < Q(s_i),$$

$$g^T s_{i+1} < g^T s_i(\alpha) < g^T s_i,$$

$$\|s_{i+1}\|_C > \|s_i(\alpha)\|_C > \|s_i\|_C, \tag{250}$$

$$\frac{g^T s_{i+1}}{\|g\|_D\|s_{i+1}\|_C} > \frac{g^T s_i(\alpha)}{\|g\|_D\|s_i(\alpha)\|_C} > \frac{g^T s_i}{\|g\|_D\|s_i\|_C}, \tag{251}$$

kde $\|s\|_C^2 = s^T C s$ a $\|g\|_D^2 = g^T C^{-1} g$ (norma $\|.\|_D$ je duální k normě $\|.\|_C$).

Důkaz Stačí použít substituce uvedené v poznámce 95. □

Věta 70. *Jsou-li splněny předpoklady věty 69, platí*

$$-Q(s_{i+1}) \geq \frac{1}{2}\frac{\|g\|^2}{\|B\|}, \quad -g^T s_{i+1} \geq \frac{\|g\|^2}{\|B\|}, \quad \|s_{i+1}\| \geq \frac{\|g\|}{\|B\|}. \tag{252}$$

Je-li navíc matice B pozitivně definitní, platí

$$\|s_{i+1}\| \leq \frac{1}{\underline{\lambda}(B)}\|g\|, \quad -\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{\sqrt{\kappa(B)}}. \tag{253}$$

Důkaz (a) Protože

$$s_2 = s_1 + \alpha_1 p_1 = \frac{g_1^T g_1}{p_1^T B p_1} p_1 = -\frac{g^T g}{g^T B g} g,$$

platí

$$-Q(s_2) = \frac{(g^T g)^2}{g^T B g} - \frac{1}{2}\frac{(g^T g)^2 g^T B g}{(g^T B g)^2} = \frac{1}{2}\frac{(g^T g)^2}{g^T B g} \geq \frac{1}{2}\frac{\|g\|^2}{\|B\|},$$

$$-g^T s_2 = \frac{(g^T g)^2}{g^T B g} \geq \frac{\|g\|^2}{\|B\|}, \qquad \|s_2\| = \frac{g^T g}{g^T B g}\|g\| \geq \frac{\|g\|}{\|B\|},$$

takže podle (246)–(248) dostaneme (252).

(b) Je-li matice B pozitivně definitní, platí $p_j^T B p_j > 0$, kdykoliv $\|g_j\| > 0$, neboť z (239) plyne

$$p_j^T p_j = (-g_j + \beta_{j-1} p_{j-1})^T (-g_j + \beta_{j-1} p_{j-1}) = g_j^T g_j + \beta_{j-1}^2 p_{j-1}^T p_{j-1} \geq g_j^T g_j,$$

neboli $\|p_j\| \geq \|g_j\|$. Podle věty 40 existuje index $m \leq n$ takový, že $\|g_j\| > 0$ pro $1 \leq j \leq m$ a $g_{m+1} = 0$. Podle (248)–(249) můžeme pro $i \leq m$ psát

$$\|s_{i+1}\| \leq \|s_{m+1}\|, \qquad \frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \leq \frac{g^T s_{m+1}}{\|g\|\|s_{m+1}\|}. \tag{254}$$

Jelikož $g_{m+1} = g + B s_{m+1} = 0$, je vektor s_{m+1} řešením soustavy rovnic $g + Bs = 0$, takže lze psát $\|g\| = \|B s_{m+1}\| \geq \underline{\lambda}(B)\|s_{m+1}\|$, a podle věty 10 platí

$$-\frac{g^T s_{m+1}}{\|g\|\|s_{m+1}\|} \geq \frac{1}{\sqrt{\kappa(B)}}.$$

Po dosazení těchto vztahů do (254) dostaneme (253). □

Důsledek 5. *Jsou-li splněny předpoklady důsledku 4, platí*

$$-Q(s_{i+1}) \geq \frac{1}{2}\frac{\|g\|^2}{\kappa(C)\|B\|}, \quad -g^T s_{i+1} \geq \frac{\|g\|^2}{\kappa(C)\|B\|}, \quad \|s_{i+1}\| \geq \frac{\|g\|}{\kappa(C)\|B\|}. \tag{255}$$

Je-li navíc matice B pozitivně definitní, platí

$$\|s_{i+1}\| \leq \frac{\sqrt{\kappa(C)}}{\underline{\lambda}(B)}\|g\|, \quad -\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{\kappa(C)\sqrt{\kappa(B)}}. \tag{256}$$

Důkaz (a) Podobně jako v důkazu věty 70 dostaneme

$$-Q(s_2) = -\tilde{Q}(\tilde{s}_2) = \frac{1}{2}\frac{(\tilde{g}^T\tilde{g})^2}{\tilde{g}^T\tilde{B}\tilde{g}} \geq \frac{1}{2}\frac{\|\tilde{g}\|^2}{\|\tilde{B}\|} \geq \frac{1}{2}\frac{g^T C^{-1} g}{\|C^{-1}\|\|B\|} \geq \frac{1}{2}\frac{\|g\|^2}{\kappa(C)\|B\|},$$

$$-g^T s_2 = -\tilde{g}^T \tilde{s}_2 = \frac{(\tilde{g}^T \tilde{g})^2}{\tilde{g}^T \tilde{B} \tilde{g}} \geq \frac{\|g\|^2}{\kappa(C)\|B\|},$$

$$\|s_2\| \geq \frac{1}{\sqrt{\|C\|}} \|\tilde{s}_2\| \geq \frac{1}{\sqrt{\|C\|}} \frac{\|\tilde{g}\|}{\|\tilde{B}\|} \geq \frac{1}{\kappa(C)} \frac{\|g\|}{\|B\|},$$

což spolu s (246)–(248) dává (255).

(b) Jelikož pro $1 \leq i \leq m$ platí

$$\underline{\lambda}(C)\|s_{i+1}\|^2 \leq s_{i+1}^T C s_{i+1} \leq \overline{\lambda}(C)\|s_{i+1}\|^2, \qquad \frac{1}{\overline{\lambda}(C)}\|g\|^2 \leq g^T C^{-1} g \leq \frac{1}{\underline{\lambda}(C)}\|g\|^2$$

a tedy

$$\frac{1}{\kappa(C)}\|g\|^2\|s_{i+1}\|^2 \leq g^T C^{-1} g s_{i+1}^T C s_{i+1} \leq \kappa(C)\|g\|^2\|s_{i+1}\|^2,$$

můžeme podle (250)–(251) psát

$$\|s_{i+1}\|^2 \leq \frac{1}{\underline{\lambda}(C)} s_{i+1}^T C s_{i+1} \leq \frac{1}{\underline{\lambda}(C)} s_{m+1}^T C s_{m+1} \leq \kappa(C)\|s_{m+1}\|^2$$

a

$$\frac{(g^T s_{i+1})^2}{\|g\|^2\|s_{i+1}\|^2} \geq \frac{1}{\kappa(C)} \frac{(g^T s_{i+1})^2}{g^T C^{-1} g s_{i+1}^T C s_{i+1}} \geq \frac{1}{\kappa(C)} \frac{(g^T s_{m+1})^2}{g^T C^{-1} g s_{m+1}^T C s_{m+1}} \geq \frac{1}{\kappa^2(C)} \frac{(g^T s_{m+1})^2}{\|g\|^2\|s_{m+1}\|^2}.$$

Protože vektor s_{m+1} je řešením soustavy rovnic $g + Bs = 0$, můžeme pokračovat stejným způsobem jako v důkazu věty 70. □

Poznámka 98. Je-li matice C vybrána tak, že $\kappa(\tilde{B}) \leq \kappa(B)$ (což je účelem předpodmínění), můžeme nerovnost (256) nahradit nerovností

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{\sqrt{\kappa(B)\kappa(C)}},$$

neboť podle věty 70 platí

$$\begin{aligned} \frac{(g^T s_{i+1})^2}{\|g\|^2\|s_{i+1}\|^2} &\geq \frac{1}{\kappa(C)} \frac{(g^T s_{i+1})^2}{g^T C^{-1} g s_{i+1}^T C s_{i+1}} = \frac{1}{\kappa(C)} \frac{(\tilde{g}^T \tilde{s}_{i+1})^2}{\tilde{g}^T \tilde{g} \tilde{s}_{i+1}^T \tilde{s}_{i+1}} \\ &\geq \frac{1}{\kappa(C)} \frac{(\tilde{g}^T \tilde{s}_{i+1})^2}{\tilde{g}^T \tilde{g} \tilde{s}_{m+1}^T \tilde{s}_{m+1}} \geq \frac{1}{\kappa(C)\kappa(\tilde{B})} \geq \frac{1}{\kappa(C)\kappa(B)} \end{aligned}$$

Zatím jsme se zabývali případem, kdy $p_j^T B p_j > 0$, $1 \leq j \leq i$. Nyní vyšetříme případ, kdy $p_j^T B p_j > 0$, $1 \leq j \leq i-1$ a $p_i^T B p_i \leq 0$.

Věta 71. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s $C = I$ na kvadratickou funkci $Q(s)$ a platí-li $g_j^T g_j > 0$, $p_j^T B p_j > 0$ pro $1 \leq j \leq i-1$ a $g_i^T g_i > 0$, $p_i^T B p_i \leq 0$ můžeme pro $\alpha > 0$ psát*

$$Q(s_i(\alpha)) < Q(s_i), \tag{257}$$

$$g^T s_i(\alpha) < g^T s_i, \tag{258}$$

$$\|s_i(\alpha)\| > \|s_i\|, \tag{259}$$

$$\frac{g^T s_i(\alpha)}{\|g\|\|s_i(\alpha)\|} > \frac{g^T s_i}{\|g\|\|s_i\|}. \tag{260}$$

Důkaz Podobně jako v části (a) důkazu věty 69 platí

$$Q(s_i(\alpha)) = Q(s_i) - \alpha g_i^T g_i + \frac{1}{2}\alpha^2 p_i^T B p_i \leq Q(s_i) - \alpha g_i^T g_i$$

neboť $p_i^T B p_i \leq 0$. Lineární funkce na pravé straně této nerovnosti klesá pro $\alpha \geq 0$. Zbytek důkazu je totožný s částmi (b)–(d) důkazu věty 69, neboť se v těchto částech nepoužívá výraz $p_i^T B p_i$. □

Věta 71 má velký význam při vyšetřování nepřesných metod s lokálně omezeným krokem (oddíl 6.3), neboť ukazuje, že funkce $Q(s_i + \alpha p_i)$ klesá a norma $\|s_i + \alpha p_i\|$ roste se vzrůstající hodnotou parametru α i v případě, že $p_i^T B p_i < 0$ (pokud $p_j^T B p_j > 0$ pro $1 \leq j \leq i-1$). Hodnota funkce $Q(s_i + \alpha p_i)$ se tedy sníží, zvětšíme-li hodnotu parametru $\alpha > 0$ tak, aby platilo $\|s_i + \alpha p_i\| = \Delta$. Ve větě 71 je podstatné, že $\alpha > 0$. V dalším kroku metody sdružených gradientů bychom totiž dostali $\alpha_i = g_i^T C^{-1} g_i / p_i^T B p_i < 0$.

Používáme-li metodu sdružených gradientů pro výpočet směrového vektoru v metodách spádových směrů, je třeba výpočet ukončit pokud neplatí $p_i^T B p_i \geq \underline{c}\, p_i^T p_i$, kde $\underline{c}$ je vhodně zvolená malá hodnota (v opačném případě bychom mohli dostat nevhodný směrový vektor, například takový, který by nebyl spádový). Jelikož podmínka $p_i^T B p_i \geq \underline{c}\, p_i^T p_i$ nemusí být splněna pro všechny indexy $1 \leq i \leq m$, nemůžeme použít nerovnosti (256). Platí však tato věta [147].

Věta 72. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s pozitivně definitní maticí C na kvadratickou funkci $Q(s)$ a jsou-li splněny nerovnosti $p_j^T B p_j \geq \underline{c}\, p_j^T p_j$ pro $1 \leq j \leq i$, platí (255) a*

$$\|s_{i+1}\| \leq \frac{n}{\underline{c}}\|g\|, \qquad -\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{\underline{c}}{n\kappa(C)\|B\|}. \tag{261}$$

Důkaz Nerovnosti (255) plynou bezprostředně z důsledku 5, neboť k jejich odvození není třeba předpokládat pozitivní definitnost matice B. Použijeme-li (245), dostaneme

$$s_{i+1} = \sum_{j=1}^{i} \alpha_j p_j = -\sum_{j=1}^{i} \frac{p_j^T g}{p_j^T B p_j} p_j = -\sum_{j=1}^{i} \frac{p_j p_j^T}{p_j^T B p_j} g,$$

takže

$$\|s_{i+1}\| \leq \sum_{j=1}^{i} \frac{\|p_j p_j^T\|}{p_j^T B p_j}\|g\| = \sum_{j=1}^{i} \frac{p_j^T p_j}{p_j^T B p_j}\|g\| \leq \frac{n}{\underline{c}}\|g\|.$$

Spojíme-li tuto nerovnost s druhou nerovností v (255), dostaneme

$$-\frac{g^T s_{i+1}}{\|s_{i+1}\|\|g\|} \geq \frac{\|g\|^2}{\kappa(C)\|B\|}\,\frac{\underline{c}}{n\|g\|^2} = \frac{\underline{c}}{n\kappa(C)\|B\|}.$$

□

Poznámka 99. Je-li matice B pozitivně definitní, můžeme položit $\underline{c} = \underline{\lambda}(B)$, takže dostaneme

$$\|s_{i+1}\| \leq \frac{n}{\underline{\lambda}(B)}\|g\|, \qquad -\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{n\kappa(B)\kappa(C)}.$$

Pokud $n < \sqrt{\kappa(C)}$, dává první nerovnost lepší odhad než odpovídající nerovnost v (256). Druhá nerovnost však dává vždy horší odhad než odpovídající nerovnost v (256).

Algoritmus předpodmíněné metody sdružených gradientů pro výpočet směrových vektorů v metodách spádových směrů lze popsat zhruba takto:

Algoritmus 6. Data $C \succ 0$, $\underline{c} > 0$, $0 < \omega < 1$, $m \geq n$ (obvykle $m = n + 3$).

Krok 1 Položíme $s := 0$, $r := -g$, $v := C^{-1}r$, $\sigma := r^T v$, $\overline{\sigma} := \sigma$, $p := r$ a $k := 1$.

Krok 2 Položíme $\rho := \sigma$, vypočteme vektor $q := Bp$ a číslo $\tau := p^T q$. Jestliže $\tau \geq \underline{c} p^T p$, přejdeme na krok 3. Pokud $k = 1$, položíme $s := -g$. Ukončíme výpočet.

Krok 3 Položíme $\alpha := \rho/\tau$. Položíme $s := s + \alpha p$, $r := r - \alpha q$, $v := C^{-1}r$ a $\sigma := r^T v$. Jestliže $\sigma \leq \omega^2 \overline{\sigma}$ nebo $k \geq m$, ukončíme výpočet.

Krok 5 Položíme $\beta := \sigma/\rho$, $p := r + \beta p$, $k := k + 1$ a přejdeme na krok 2.

Výpočet skončí buď v kroku 2 (matice B není pozitivně definitní) nebo v kroku 3 (je nalezeno řešení s požadovanou přesností nebo byl překročen povolený počet iterací.

Poznámka 100. Platí-li $\|B\| \leq \overline{B}$ a $\kappa(C) \leq \overline{\kappa}$, kde $\overline{B}$ a $\overline{\kappa}$ jsou konstanty společné všem iteračním krokům metody spádových směrů, pak směrový vektor s, vypočtený algoritmem 6, splňuje podle věty 72 podmínku

$$-\frac{g^T s}{\|g\|\|s\|} \geq \frac{\underline{c}}{n\overline{\kappa}\overline{B}} \stackrel{\Delta}{=} \varepsilon_0,$$

takže je stejnoměrně spádový a odpovídající metoda spádových směrů je globálně konvergentní. Algoritmus 6 lze tedy použít k realizaci globálně konvergentní modifikované Newtonovy metody spádových směrů.

Metodu sdružených gradientů můžeme použít k určení spádového směru a směru se zápornou křivostí v metodách spádových párů popsaných v oddílu 2.6. V tomto případě metodu sdružených gradientů přerušíme až tehdy, když $|p_i^T B p_i| < \underline{c}\, p_i^T p_i$. Pak položíme

$$s = + \sum_{j \in I_+} \alpha_j p_j = \sum_{j \in I_+} \frac{g_j^T C^{-1} g_j}{|p_j^T B p_j|} p_j, \qquad I_+ = \{j : 1 \leq j < i,\, p_j^T B p_j > 0\}, \tag{262}$$

$$z = - \sum_{j \in I_-} \alpha_j p_j = \sum_{j \in I_-} \frac{g_j^T C^{-1} g_j}{|p_j^T B p_j|} p_j, \qquad I_- = \{j : 1 \leq j < i,\, p_j^T B p_j < 0\}, \tag{263}$$

přičemž $s = 0$, pokud $I_+ = \emptyset$, a $z = 0$, pokud $I_- = \emptyset$. Jestliže $|p_1^T B p_1| < \underline{c}\, p_1^T p_1$ (takže $I_+ \cup I_- = \emptyset$), pokládáme $s = -g$ a $z = 0$.

Lemma 30. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s pozitivně definitní maticí C na kvadratickou funkci (237) a jsou-li vektory s a z určeny podle (262) a (263), kde $I_+ \cup I_- \neq \emptyset$, platí*

$$-g^T(s + z) \geq \frac{\|g\|^2}{\kappa(C)\|B\|}, \qquad \|s + z\| \leq \frac{n}{\underline{c}}\|g\|. \tag{264}$$

Důkaz (a) Podobně jako v části (b) důkazu věty 69 dostaneme

$$-g^T(s + z) = -\sum_{j=1}^{i} |\alpha_j| g^T p_j = \sum_{j=1}^{i} |\alpha_j| g_j^T C^{-1} g_j,$$

takže jako v části (a) důkazu důsledku 5 platí

$$-g^T(s + z) \geq |\alpha_1| g_1^T C^{-1} g_1 = \frac{(g^T C^{-1} g)^2}{g^T C^{-1} B C^{-1} g} = \frac{(\tilde{g}^T \tilde{g})^2}{\tilde{g}^T \tilde{B} \tilde{g}} \geq \frac{\|g\|^2}{\kappa(C)\|B\|}$$

(b) Podobně jako v důkazu věty 72 dostaneme

$$s+z=\sum_{j=1}^{i}|\alpha_j|p_j=-\sum_{j=1}^{i}\frac{p_j^Tg}{|p_j^TBp_j|}p_j=-\sum_{j=1}^{i}\frac{p_jp_j^T}{|p_j^TBp_j|}g,$$

takže

$$\|s+z\|\le\sum_{j=1}^{i}\frac{\|p_jp_j^T\|}{|p_j^TBp_j|}\|g\|=\sum_{j=1}^{i}\frac{p_j^Tp_j}{|p_j^TBp_j|}\|g\|\le\frac{n}{\underline{c}}\|g\|.$$

□

Věta 73. *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s pozitivně definitní maticí C na kvadratickou funkci (237) a jsou-li vektory s a z určeny podle (262) a (263), kde $I_+\cup I_-\neq\emptyset$, platí*

$$-g^Ts\ge\frac{3}{4}\frac{\|g\|^2}{\kappa(C)\|B\|},\quad \|s\|\le\frac{n}{\underline{c}}\|g\|,\quad \textit{pokud}\quad I_+\neq\emptyset\quad a\quad Q(s)\le Q(z),\tag{265}$$

a

$$-g^Tz\ge\frac{1}{4}\frac{\|g\|^2}{\kappa(C)\|B\|},\quad \|z\|\le\frac{n}{\underline{c}}\|g\|,\quad \textit{pokud}\quad I_-\neq\emptyset\quad a\quad Q(z)\le Q(s).\tag{266}$$

Pokud $z\neq0$, můžeme psát

$$\frac{z^TBz}{z^Tz}\le\frac{1}{4}\frac{\underline{c}^2}{n^2\kappa(C)\|B\|^2}\underline{\lambda}(B).\tag{267}$$

Důkaz (a) Použijeme-li (262) spolu se (241) a (245), dostaneme

$$\begin{aligned}Q(s)&=g^T\left(\sum_{j\in I_+}\alpha_jp_j\right)+\frac{1}{2}\left(\sum_{j\in I_+}\alpha_jp_j\right)^TB\left(\sum_{j\in I_+}\alpha_jp_j\right)\\&=g^T\left(\sum_{j\in I_+}\alpha_jp_j\right)+\frac{1}{2}\left(\sum_{j\in I_+}\alpha_j^2p_j^TBp_j\right)=\left(\sum_{j\in I_+}\alpha_jg^Tp_j\right)-\frac{1}{2}\left(\sum_{j\in I_+}\alpha_jg^Tp_j\right)=\frac{1}{2}g^Ts\end{aligned}$$

a podobně z (263) plyne

$$Q(z)=\left(\sum_{j\in I_-}|\alpha_j|g^Tp_j\right)+\frac{1}{2}\left(\sum_{j\in I_-}|\alpha_j|g^Tp_j\right)=\frac{3}{2}g^Tz.$$

Pokud $Q(s)\le Q(z)$, platí $g^Ts\le3g^Tz$, neboli $4g^Ts\le3g^T(s+z)$, což spolu s první nerovností v (264) dává první nerovnost v (265). Pokud $Q(z)\le Q(s)$, platí $3g^Tz\le g^Ts$, neboli $4g^Tz\le g^T(s+z)$, což spolu s první nerovností v (264) dává první nerovnost v (266).

(b) Druhé nerovnosti v (265) a (266) jsou bezprostředním důsledkem druhé nerovnosti v (264) (plyne to z části (b) důkazu lemmatu 30).

(c) Použijeme-li (266), dostaneme

$$z^Tz\le\frac{n^2}{\underline{c}^2}\|g\|^2\quad\text{a}\quad z^TBz=g^Tz\le-\frac{1}{4}\frac{\|g\|^2}{\kappa(C)\|B\|}$$

(vztah $z^TBz=g^Tz$ je odvozen v části (a)). Platí tedy

$$\frac{z^TBz}{z^Tz}\le-\frac{1}{4}\frac{\underline{c}^2\|B\|}{n^2\kappa(C)\|B\|^2}$$

a jelikož $-\|B\| \leq \underline{\lambda}(B) < 0$, dostaneme (267). □

Algoritmus předpodmíněné metody sdružených gradientů pro výpočet spádových směrů a směrů se zápornou křivostí v metodách spádových párů lze popsat zhruba takto:

Algoritmus 7. Data $C \succ 0$, $\underline{c} > 0$, $0 < \omega < 1$, $m \geq n$ (obvykle $m = n + 3$).

Krok 1 Položíme $s := 0$, $z := 0$, $r := -g$, $v := C^{-1}r$, $\sigma := r^T v$, $\overline{\sigma} := \sigma$, $p := r$ a $k := 1$.

Krok 2 Položíme $\rho := \sigma$, vypočteme vektor $q := Bp$ a číslo $\tau := p^T q$. Jestliže $|\tau| \geq \underline{c}\, p^T p$, přejdeme na krok 3. Pokud $k = 1$, položíme $s := -g$. Ukončíme výpočet.

Krok 3 Položíme $\alpha := \rho/\tau$. Jestliže $\alpha > 0$, položíme $s := s + \alpha p$. Jestliže $\alpha < 0$, položíme $z := z - \alpha p$. Položíme $r := r - \alpha q$, $v := C^{-1}r$ a $\sigma := r^T v$. Jestliže $\sigma \leq \omega^2 \overline{\sigma}$ nebo $k \geq m$, ukončíme výpočet.

Krok 4 Položíme $\beta := \sigma/\rho$, $p := r + \beta p$, $k := k + 1$ a přejdeme na krok 2.

Poznámka 101. Platí-li $\|B\| \leq \overline{B}$ a $\kappa(C) \leq \overline{\kappa}$, kde $\overline{B}$ a $\overline{\kappa}$ jsou konstanty společné všem iteračním krokům metody spádových směrů, pak dvojice vektorů $(s, z) \in R^n \times R^n$, vypočtená algoritmem 7, kde $B = G(x)$, je přijatelným spádovým párem pro funkci $F \in C^2 : D \to R$ v bodě x, přičemž platí (80) a (81), kde

$$\underline{s} = \frac{3}{4\overline{\kappa}\overline{B}}, \qquad \overline{s} = \frac{n}{\underline{c}}, \qquad \varepsilon_0 = \frac{3\underline{c}}{4n\overline{\kappa}\overline{B}}$$
$$\underline{z} = \frac{1}{4\overline{\kappa}\overline{B}}, \qquad \overline{z} = \frac{n}{\underline{c}}, \qquad \delta_0 = \frac{\underline{c}^2}{4n^2\overline{\kappa}\overline{B}^2}.$$

Poznámka 102. Chceme-li použít metodu sdružených gradientů k určení spádového směrového vektoru, máme tyto tři možnosti: vektor s získaný algoritmem 6, vektor s získaný algoritmem 7 a vektor $s + z$ získaný algoritmem 7. Numerické testy ukazují, že nejvýhodnější volbou je vektor s získaný algoritmem 6. Tento směrový vektor je přinejmenším stejně dobrý jako ostatní dva, ale jeho výpočet spotřebuje méně iterací metody sdružených gradientů.

Poznámka 103. Určujeme-li spádový pár metodou sdružených gradientů, můžeme podmínky (88)–(89) nahradit podmínkami

$$d_i(\alpha) = \alpha s_i, \quad \text{pokud} \quad Q_i(z_i) > Q_i(s_i), \tag{268}$$
$$d_i(\alpha) = \alpha z_i, \quad \text{pokud} \quad Q_i(z_i) \leq Q_i(s_i). \tag{269}$$

V důkazu věty 73 je ukázáno, že $Q_i(s_i) = (1/2) g_i^T s_i$, takže $Q_i(z_i) \leq Q_i(s_i)$ implikuje

$$Q_i(z_i) \leq \frac{1}{2} g_i^T s_i \leq \frac{\underline{s}}{2} \frac{g_i^T s_i}{\|s_i\|}.$$

Tuto nerovnost můžeme použít v částech (b) a (c) důkazu věty 27 (kde položíme $\underline{c} = \underline{s}/2$).

4 Metody s proměnnou metrikou

Metody s proměnnou metrikou patří mezi nejefektivnější metody pro řešení optimalizačích úloh menšího rozměru (do 250 proměnných) s hustou Hessovou maticí. Tyto metody jsou založeny na aktualizaci matic aproximujících Hessovu matici nebo její inverzi. Aktualizace mají hodnost nanejvýš 2 a vybírají se tak, aby byla splněna standardní nebo zobecněná kvazinewtonovská podmínka, přičemž je kladen důraz na to aby metoda nalezla minimum ryze konvexní kvadratické funkce po konečném počtu kroků.

První metodu s proměnnou metrikou (metodu DFP) popsal Davidon ve své (prakticky nedostupné) práci z roku 1959, která v tehdejší době nebyla přijata k časopisecké publikaci. Teprve později, po vydání článku [56], se metody s proměnnou metrikou dostaly do popředí zájmu a v roce 1991 (32 let po napsání výzkumné zprávy) vyšel původní Davidonův text v prvním čísle časopisu SIAM Journal on Optimization [38]. Brzy se ukázalo, že existuje jednoparametrická třída metod s proměnnou metrikou (Broydenova třída [11]), která obsahuje metody účinější než je metoda DFP, a pro další zvýšení účinnosti byly metody s proměnnou metrikou různě modifikovány (škálování [129] a korekce [5]).

V oddílech 4.1–4.8 se budeme zabývat základními metodami s proměnnou metrikou, patřícími do Broydenovy třídy, a jejich modifikacemi. V oddílu 4.9 budeme vyšetřovat Davidonovu třídu metod s proměnnou metrikou.

4.1 Základní vlastnosti metod s proměnnou metrikou

Definice 36. *Řekneme, že metoda spádových směrů (definice 17) je metodou s proměnnou metrikou, jestliže*

$$s_i = -H_i g_i, \tag{270}$$

kde H_i, $i \in N$, jsou symetrické pozitivně definitní matice konstruované podle rekurentního vztahu

$$H_{i+1} = \gamma_i (H_i + U_i M_i U_i^T), \tag{271}$$

kde $U_i \in R^{n\times 2}$, $M_i \in R^{2\times 2}$ a $\gamma_i > 0$, a vyhovující podmínce

$$H_{i+1} y_i = \rho_i d_i, \tag{272}$$

kde $y_i = g_{i+1} - g_i$, $d_i = x_{i+1} - x_i = \alpha_i s_i$ a $\rho_i > 0$.

Poznámka 104. Matice H_{i+1} se získává z matice H_i aktualizací jejíž hodnost je nanejvýš 2. Nejefektivnější metody s proměnnou metrikou patří do Broydenovy třídy, která je charakterizovaná výběrem $U_i = [d_i, H_i y_i]$. Podmínka (272) se nazývá (zobecněnou) kvazinewtonovskou podmínkou (předpokládáme, že $d_i \neq 0$ a $y_i \neq 0$, neboť v opačném případě nemá podmínka (272) smysl). Původní metody s proměnnou metrikou byly navrženy s hodnotami $\gamma_i = 1$ a $\rho_i = 1$ (bez škálování a korekce). Jelikož efektivní škálování a vhodná korekce zlepšují účinnost metod s proměnnou metrikou, budeme vyšetřovat obecný případ, kdy $\gamma_i > 0$ a $\rho_i > 0$. Škálování s $\gamma_i \neq 1$ bylo poprvé použito v práci [129] a korekce s $\rho_i \neq 1$ v práci [4].

Vývoj metod s proměnnou metrikou byl motivován tím, že tyto metody, realizované jako metody spádových směrů s přesným výběrem délky kroku, najdou minimum kvadratické funkce po konečném počtu kroků.

Věta 74. *(Kvadratické ukončení) Nechť x_i $i \in N$, je posloupnost generovaná metodou s proměnnou metrikou z Broydenovy třídy s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0$, $i \in N$) aplikovaná na ryze konvexní kvadratickou funkci (99). Pak existuje index $m \leq n$ takový, že $g_{m+1} = 0$ a $x_{m+1} = x^*$.*

Důkaz Předpokládejme, že $g_i \neq 0$, $1 \leq i \leq n$ (není-li tato podmínka splněna, platí $g_{m+1} = 0$ a $x_{m+1} = x^*$ pro nějaký index $m < n$). Dokážeme indukcí, že pro libovolný index $1 \leq i \leq n$ je $s_i \neq 0$, $\alpha_i > 0$, přičemž

$$H_{i+1} y_j = \lambda_i^j d_j, \qquad 1 \leq j \leq i, \tag{273}$$

$$s_j^T g_{i+1} = 0, \qquad 1 \leq j \leq i, \tag{274}$$

$$s_j^T G s_i = 0, \qquad 1 \leq j < i, \tag{275}$$

$$s_j^T y_i = y_j^T s_i = 0, \qquad 1 \leq j < i, \tag{276}$$

kde $\lambda_i^j > 0$, $1 \le j \le i$ (viz (277)). Rovnosti (275) a (276) jsou ekvivalentní, neboť pro kvadratickou funkci (99) platí $y_i = g_{i+1} - g_i = G(x_{i+1} - x_i) = Gd_i = \alpha_i G s_i$ a $\alpha_i > 0$. Z (270) a (275) plyne, že vektory s_i, $1 \le i \le n$, jsou nenulové a vzájemně sdružené (G-ortogonální), tudíž lineárně nezávislé, takže podle (274) nutně $g_{n+1} = 0$ a $x_{n+1} = x^*$. Pro $i = 1$ je $s_1^T g_1 = -g_1^T H_1 g_1 < 0$ (matice H_1 je pozitivně definitní), takže $s_1 \neq 0$ a $\alpha_1 > 0$. Nechť pro nějaký index $1 \le i < n$ platí $s_j \neq 0$, $\alpha_j > 0$, pro $1 \le j \le i$, a $H_i y_j = \lambda_i^j d_j$, $s_j^T g_i = 0$, $s_j^T G s_i = s_j^T y_i = y_j^T s_i = 0$, pro $1 \le j < i$ (indukční předpoklad).

(a) Použijeme-li (271), dostaneme

$$H_{i+1} y_j = \gamma_i (H_i y_j + U_i^T M_i U_i^T y_j) = \gamma_i H_i y_j = \gamma_i \lambda_i^j d_j \triangleq \lambda_{i+1}^j d_j,$$

pro $1 \le j < i$, neboť podle indukčního předpokladu pro $1 \le j < i$ platí $H_i y_j = \lambda_i^j d_j$ a

$$U_i^T y_j = [d_i, H_i y_i]^T y_j = \alpha_i [s_i^T y_j, \lambda_i^j y_i^T s_j] = 0.$$

Dále podle (272) platí $H_{i+1} y_i = \rho_i d_i \triangleq \lambda_{i+1}^i d_i$, takže $H_{i+1} y_j = \lambda_{i+1}^j d_i$, $1 \le j \le i$.

(b) Zřejmě $s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = 0$ pro $1 \le j < i$ (neboť podle indukčního předpokladu pro $1 \le j < i$ platí $s_j^T g_i = 0$ a $s_j^T y_i = 0$). Z přesného výběru délky kroku plyne, že $s_i^T g_{i+1} = 0$. Platí tedy $s_j^T g_{i+1} = 0$ pro $1 \le j \le i$.

(c) Podle (270) je $g_{i+1}^T s_{i+1} = -g_{i+1}^T H_{i+1} g_{i+1} < 0$ takže $s_{i+1} \neq 0$ a $\alpha_{i+1} > 0$. Použijeme-li (270) spolu s (a), (b), dostaneme

$$y_j^T s_{i+1} = -y_j^T H_{i+1} g_{i+1} = -\lambda_{i+1}^j d_j^T g_{i+1} = -\lambda_{i+1}^j \alpha_j s_j^T g_{i+1} = 0$$

pro $1 \le j \le i$. □

Poznámka 105. Z části (a) důkazu věty 74 plyne, že pro $1 \le j < i \le n$ platí

$$\lambda_i^j = \left(\prod_{k=j}^{i-1} \gamma_k \right) \frac{\rho_j}{\gamma_j}. \tag{277}$$

Důsledek 6. *Nechť jsou splněny předpoklady věty 74. Pak pro $1 \le i < n$ tvoří vektory s_i a s_{i+1} bázi v $\mathcal{L}(U_i)$ a pro $1 \le i < j \le n$ platí*

$$H_i g_j = \left(\prod_{k=1}^{i-1} \gamma_k \right) H_1 g_j. \tag{278}$$

Důkaz (a) Nechť $1 \le i < n$ a v je libovolný vektor takový, že $v^T U_i = v^T [d_i, H_i y_i] = 0$ (takže v leží v ortogonálním doplňku podprostoru $\mathcal{L}(U_i)$). Pak $v^T s_i = v^T d_i / \alpha_i = 0$ a podle (270) a (271) platí

$$v^T s_{i+1} = -v^T H_{i+1} g_{i+1} = -\gamma_i v^T (H_i g_i + H_i y_i) = \gamma_i v^T (s_i - H_i y_i) = 0,$$

takže vektory s_i a s_{i+1} leží v $\mathcal{L}(U_i)$, a jelikož podle (275) jsou tyto vektory lineárně nezávislé, tvoří bázi v $\mathcal{L}(U_i)$.

(b) Pro $i = 1$ je rovnost (278) zřejmá. Nechť $1 < i < j \le n$. Podle (274) platí $s_k^T g_j = 0$ a $s_{k+1}^T g_j = 0$ $\forall 1 \le k < i < j$, což podle (a) dává $U_k^T g_j = 0$ $\forall 1 \le k < i < j$. Několikanásobným použitím (271) pak dostaneme

$$H_i g_j = \gamma_{i-1} H_{i-1} g_j = \cdots = \left(\prod_{k=1}^{i-1} \gamma_k \right) H_1 g_j.$$

□

Ukážeme nyní, že jsou-li splněny předpoklady věty 74, generují všechny metody s proměnnou metrikou z Broydenovy třídy stejnou posloupnost bodů x_i, $i \in N$.

Věta 75. *Nechť jsou splněny předpoklady věty 74. Pak všechny metody s proměnnou metrikou z Broydenovy třídy generují stejnou posloupnost bodů x_i, $i \in N$.*

Důkaz Protože, v případě ryze konvexní kvadratické funkce, přesný výběr délky kroku určuje posloupnost bodů x_i, $i \in N$, jednoznačně, nezávisle na velikosti směrových vektorů s_i, $i \in N$, stačí dokázat že příslušné směrové vektory jsou rovnoběžné. Uvažujme iterační procesy $x_{i+1} = x_i - \alpha_i H_i g_i$ a $\bar{x}_{i+1} = \bar{x}_i - \bar{\alpha}_i \bar{H}_i \bar{g}_i$, kde $\bar{x}_1 = x_1$ a $\bar{H}_1 = H_1$, určené dvěma metodami s proměnnou metrikou (lišícími se výběrem parametrů γ_i, ρ_i a matic M_i v (271) a (272)). Důkaz provedeme indukcí. Budeme předpokládat, že pro nějaký index $1 \leq i < n$ platí $\bar{s}_i \| s_j$ a $\mathcal{L}(\bar{U}_j) = \mathcal{L}(U_j)$ $\forall 1 \leq j \leq i$ (a tudíž také $\bar{d}_j = d_j$ a $\bar{y}_j = y_j$ $\forall 1 \leq j \leq i$, neboť přesný výběr délky kroku je jednoznačný). Platí to zcela jistě pro $i = 1$, neboť $\bar{g}_1 = g_1$ a $\bar{H}_1 = H_1$, takže $\bar{s}_1 = s_1$, $\bar{d}_1 = d_1$, $\bar{y}_1 = \bar{y}_1$ $\bar{H}_1 y_1 = H_1 y_1$ a $\bar{U}_1 = U_1$.

(a) Podle důsledku 6 leží vektor s_{i+1} v $\mathcal{L}(U_i)$ a podle (276) platí $s_{i+1}^T y_i = 0$. Jelikož y_i neleží v ortogonálním doplňku podprostoru $\mathcal{L}(U_i)$ (neboť platí $d_i^T y_i > 0$) a $\mathcal{L}(U_i)$ má dimenzi 2, je směr vektoru s_{i+1} jednoznačně určen podprostorem $\mathcal{L}(U_i)$ a vektorem y_i. Stejné úvahy platí pro vektor $\bar{s}_{i+1}$. Jelikož $\mathcal{L}(\bar{U}_i) = \mathcal{L}(U_i)$ a $\bar{y}_i = y_i$, musí platit $\bar{s}_{i+1} \| s_{i+1}$ a tedy $\bar{d}_{i+1} = d_{i+1}$ a $\bar{y}_{i+1} = y_{i+1}$.

(b) Nechť $1 \leq j < i+1$. Použijeme-li vztahy (273) a (276), dostaneme $y_j^T H_{i+1} y_{i+1} = \lambda_i^j \alpha_j s_j^T y_{i+1} = 0$ a podobně $\bar{y}_j^T \bar{H}_{i+1} \bar{y}_{i+1} = 0$, což spolu s $\bar{y}_j = y_j$ a $\bar{y}_{i+1} = y_{i+1}$ dává $y_j^T \bar{H}_{i+1} y_{i+1} = 0$. Platí tedy

$$y_j^T \left(\bar{H}_{i+1} - \lambda_i H_{i+1} \right) y_{i+1} = 0, \qquad 1 \leq j < i+1 \tag{279}$$

pro libovolné číslo $\lambda_i > 0$. Nechť $i+1 < j \leq n$. Použijeme-li (278), můžeme psát

$$H_{i+1} y_j = \left(\prod_{k=1}^{i} \gamma_k \right) H_1 y_j \triangleq \omega_i H_1 y_j.$$

Protože $\bar{H}_1 = H_1$, $\mathcal{L}(\bar{U}_j) = \mathcal{L}(U_j)$ a $\bar{y}_j = y_j$, $1 \leq j \leq i$, platí také

$$\bar{H}_{i+1} y_j = \bar{H}_{i+1} \bar{y}_j = \left(\prod_{k=1}^{i} \bar{\gamma}_k \right) \bar{H}_1 \bar{y}_j \triangleq \bar{\omega}_i \bar{H}_1 \bar{y}_j = \bar{\omega}_i H_1 y_j,$$

což spolu s předchozí rovností dává

$$y_j^T \left(\bar{H}_{i+1} - \lambda_i H_{i+1} \right) y_{i+1} = 0, \qquad i+1 < j \leq n \tag{280}$$

pro $\lambda_i = \bar{\omega}_i / \omega_i$. Vektory y_j, $1 \leq j \leq n$ jsou lineárně nezávislé a podle (276) platí $y_j^T s_{i+1} = 0$ pro $1 \leq j < i+1$ a $i+1 < j \leq n$. Porovnáme-li tyto rovnosti s rovnostmi (279) a (280), vidíme, že vektory s_{i+1} a $\left(\bar{H}_{i+1} - \lambda_i H_{i+1} \right) y_{i+1}$ jsou rovnoběžné, takže vektor $\bar{H}_{i+1} \bar{y}_{i+1} = \bar{H}_{i+1} y_{i+1}$ je lineární kombinací vektorů s_{i+1} a $H_{i+1} y_{i+1}$, což spolu s $\bar{s}_{i+1} \| s_{i+1}$ dává $\mathcal{L}(\bar{U}_{i+1}) = \mathcal{L}(U_{i+1})$. □

Důsledek 7. *Nechť jsou splněny předpoklady věty 74. Pak metody s proměnnou metrikou z Broydenovy třídy generují stejnou posloupnost bodů x_i, $i \in N$, jako metoda sdružených gradientů předpodmíněná maticí H_1.*

Důkaz Protože podle věty 75 generují všechny metody s proměnnou metrikou z Broydenovy třídy stejnou posloupnost bodů, stačí si vybrat jednu z těchto metod. Zde poněkud předběhneme výklad a vybereme metodu BFGS (vzorec (288)). Pak pro libovolný index $1 \leq i \leq n$ platí

$$H_{i+1} = \gamma_i \left(H_i + \left(\frac{y_i^T H_i y_i}{y_i^T d_i} + \frac{\rho_i}{\gamma_i} \right) \frac{1}{y_i^T d_i} d_i d_i^T - \frac{1}{y_i^T d_i} \left(H_i y_i d_i^T + d_i y_i^T H_i \right) \right).$$

Jelikož předpokládáme přesný výběr délky kroku, platí $d_i^T g_{i+1} = 0$, takže s použitím předchozího vztahu a vzorce (278) dostaneme

$$s_{i+1} = -H_{i+1} g_{i+1} = -\gamma_i \left(H_i g_{i+1} - \frac{y_i^T H_i g_{i+1}}{y_i^T d_i} d_i \right) = - \left(\prod_{k=1}^{i} \gamma_k \right) \left(H_1 g_{i+1} - \frac{y_i^T H_1 g_{i+1}}{y_i^T d_i} d_i \right),$$

takže směrový vektor s_{i+1} je rovnoběžný se směrovým vektorem metody sdružených gradientů předpodmíněné maticí H_1 (poznámka 66). Protože, v případě ryze konvexní kvadratické funkce, přesný výběr délky kroku určuje posloupnost bodů x_i, $i \in N$, jednoznačně, nezávisle na velikosti směrových vektorů s_i, $i \in N$, generuje metoda BFGS (a tudíž všechny metody s proměnnou metrikou z Broydenovy třídy) stejnou posloupnost bodů x_i, $i \in N$, jako metoda sdružených gradientů předpodmíněná maticí H_1 □

Pro ryze konvexní kvadratickou funkci je nejvýhodnější volit parametry γ_i a ρ_i tak, že $\gamma_i = 1$ a $\rho_i = 1$, $i \in N$.

Věta 76. *(Aproximace Hessovy matice). Nechť jsou splněny předpoklady věty 74 s $\gamma_i = 1$ a $\rho_i = 1$, $i \in N$. Pak platí $H_{n+1} = G^{-1}$.*

Důkaz Jestliže $\gamma_i = 1$ a $\rho_i = 1$, $i \in N$, platí podle (277) $\lambda_i^j = 1$, $1 \le j < i \le n+1$. Můžeme tedy psát

$$H_{n+1} y_j = d_j, \qquad 1 \le j \le n,$$

a jelikož vektory $d_j = \alpha_j s_j$, $1 \le j \le n$ (a tedy i $y_j = G d_j$, $1 \le j \le n$) jsou podle (275) lineárně nezávislé, musí platit $H_{n+1} = G^{-1}$. □

V předchozích větách jsme využívali toho, že minimalizovaná funkce je kvadratická. Překvapivě se dá dokázat, jak je uvedeno v práci [43], že všechny metody s proměnnou metrikou z Broydenovy třídy s přesným výběrem délky kroku, generují stejnou posloupnost bodů x_i, $i \in N$, i když minimalizovaná funkce není kvadratická.

Věta 77. *Nechť funkce $F \in C^2 : \mathrm{D} \to R$ splňuje předpoklady (F4) a (F5). Nechť $x_{i+1} = x_i - \alpha_i H_i g_i$ a $\bar{x}_{i+1} = \bar{x}_i - \bar{\alpha}_i \bar{H}_i \bar{g}_i$, jsou iterační procesy aplikované na funkci F, určené dvěma metodami s proměnnou metrikou z Broydenovy třídy s jednoznačně určeným přesným výběrem délky kroku, které vycházejí ze stejného bodu $\bar{x}_1 = x_1$ a v kterých používáme stejnou počáteční matici $\bar{H}_1 = H_1$. Nechť $s_i^T g_i \neq 0$, $\bar{s}_i^T \bar{g}_i \neq 0$ a $\bar{\gamma}_i = \gamma_i$, $\bar{\rho}_i = \rho_i$ $\forall i \in N$. Pak platí $\bar{x}_i = x_i$, $\bar{s}_i \| s_i$ a $\mathcal{L}(\bar{U}_i) = \mathcal{L}(U_i)$ $\forall i \in N$.*

Důkaz Větu dokážeme indukcí. Podle předpokladu platí $\bar{x}_1 = x_1$, $\bar{g}_1 = g_1$ a $\bar{H}_1 = H_1$, což podle (270) dává $\bar{s}_1 = s_1$. Protože délka kroku je určena jednoznačně, platí $\bar{x}_2 = x_2$ a $\bar{g}_2 = g_2$, takže $\bar{d}_1 = d_1$ a $\bar{y}_1 = y_1$. Jelikož $\bar{H}_1 = H_1$, platí $\bar{H}_1 \bar{y}_1 = H_1 y_1$, což spolu s $\bar{d}_1 = d_1$ dává $\mathcal{L}(\bar{U}_1) = \mathcal{L}(U_1)$. Můžeme tedy předpokládat, že pro nějaký index $i < n$ platí $\bar{x}_i = x_i$, $\bar{s}_i \| s_i$, $\mathcal{L}(\bar{U}_i) = \mathcal{L}(U_i)$ a $\bar{H}_i v = H_i v$ pro každý vektor v z ortogonálního doplňku podprostoru $\mathcal{L}(U_i)$. Zřejmě $s_i \neq 0$, $\alpha_i \neq 0$ a $\bar{s}_i \neq 0$, $\bar{\alpha}_i \neq 0$, neboť $\bar{s}_i^T \bar{g}_i \neq 0$ a $s_i^T g_i \neq 0$. Protože délka kroku je určena jednoznačně, platí $\bar{x}_{i+1} = x_{i+1}$ a $\bar{g}_{i+1} = g_{i+1}$, takže $\bar{d}_i = d_i$ a $\bar{y}_i = y_i$.

(a) Úplně stejně jako v části (a) důkazu důsledku 6 se ukáže, že vektor s_{i+1} leží v $\mathcal{L}(U_i)$. Navíc podle (271) a (272) platí $s_{i+1}^T y_i = 0$. Vektor y_i neleží v ortogonálním doplňku podprostoru $\mathcal{L}(U_i)$, neboť $d_i^T y_i = \alpha_i s_i^T (g_{i+1} - g_i) = -\alpha_i s_i^T g_i \neq 0$. Vektor s_{i+1} je tedy, stejně jako v části (a) důkazu věty 75, jednoznačně určen podprostorem $\mathcal{L}(U_i)$ a vektorem y_i. Stejné úvahy platí pro vektor $\bar{s}_{i+1}$. Jelikož $\mathcal{L}(\bar{U}_i) = \mathcal{L}(U_i)$ a $\bar{y}_i = y_i$, musí platit $\bar{s}_{i+1} \| s_{i+1}$ a tedy $\bar{d}_{i+1} = d_{i+1}$ a $\bar{y}_{i+1} = y_{i+1}$.

(b) Nechť v je libovolný vektor z ortogonálního doplňku podprostoru $\mathcal{L}(U_{i+1})$ (takže platí $d_{i+1}^T v = 0$ a $y_{i+1}^T H_{i+1} v = 0$). Jelikož $s_{i+1} \neq 0$ a $\alpha_{i+1} \neq 0$ (plyne to z nerovnosti $s_{i+1}^T g_{i+1} \neq 0$), platí nutně $s_{i+1}^T v = 0$. Vektor s_{i+1} leží podle (a) v podprostoru $\mathcal{L}(U_i)$ a je kolmý k vektoru y_i, takže $s_{i+1}^T v = 0$ platí pouze tehdy, jestliže $v = w + \lambda y_i$, kde w leží v ortogonálním doplňku podprostoru $\mathcal{L}(U_i)$. Použijeme-li vztahy (271) a (272), dostaneme

$$H_{i+1} v = H_{i+1} w + \lambda H_{i+1} y_i = \gamma_i H_i w + \lambda \rho_i d_i = \bar{\gamma}_i \bar{H}_i w + \lambda \bar{\rho}_i \bar{d}_i = \bar{H}_{i+1} w + \lambda \bar{H}_{i+1} \bar{y}_i = \bar{H}_{i+1} v, \qquad (281)$$

neboť $\bar{\gamma}_i = \gamma_i$, $\bar{\rho}_i = \rho_i$, $\bar{d}_i = d_i$, $\bar{y}_i = y_i$ a $\bar{H}_i w = H_i w$ pro libovolný vektor w z ortogonálního doplňku podprostoru $\mathcal{L}(U_i)$ (indukční předpoklad). Dokázali jsme tedy, že $\bar{H}_{i+1} v = H_{i+1} v$ pro libovolný vektor v z ortogonálního doplňku podprostoru $\mathcal{L}(U_{i+1})$.

(c) Nechť v je libovolný vektor z ortogonálního doplňku podprostoru $\mathcal{L}(U_{i+1})$. Jelikož podle (a) platí $\bar{d}_{i+1} = d_{i+1}$ a $\bar{y}_{i+1} = y_{i+1}$ a z (b) plyne $\bar{H}_{i+1} v = H_{i+1} v$, můžeme psát $\bar{d}_{i+1} v = 0$ a $\bar{y}_{i+1} \bar{H}_{i+1} v = 0$, takže v

leží v ortogonálním doplňku podprostoru $\mathcal{L}(\bar{U}_{i+1})$. Je tedy splněna inkluze $\mathcal{L}(\bar{U}_{i+1}) \subset \mathcal{L}(U_{i+1})$ a protože použité úvahy nezávisí na pořadí použitých metod, platí $\mathcal{L}(\bar{U}_{i+1}) = \mathcal{L}(U_{i+1})$. □

Nyní se budeme zabývat vyšetřováním aktualizací tvaru (271). Pro zjednodušení budeme index i vynechávat a index $i+1$ nahradíme symbolem $+$. Nejprve uvedeme několik pomocných tvrzení týkajících se těchto aktualizací.

Lemma 31. *Nechť $U \in R^{n\times m}$, $V \in R^{n\times m}$. Pak:*

(a) Matice UV^T má stejná nenulová vlastní čísla jako matice V^TU.
(b) Platí $\operatorname{Tr} UV^T = \operatorname{Tr} V^TU$.
(c) Matice $I + UV^T$ má stejná nejednotková vlastní čísla jako matice $I + V^TU$.
(d) Platí $\det(I + UV^T) = \det(I + V^TU)$.
(e) Je-li matice $I + UV^T$ regulární, platí $(I + UV^T)^{-1} = I - U(I + V^TU)^{-1}V^T$.

Důkaz (a) Nechť $UV^Tx = \lambda x$, $x \neq 0$ a $\lambda \neq 0$. Pak nutně $V^Tx \neq 0$ a můžeme psát $V^TUV^Tx = \lambda V^Tx$, neboli $V^TUy = \lambda y$, kde $y = V^Tx \neq 0$. Nechť naopak $V^TUy = \lambda y$, $y \neq 0$ a $\lambda \neq 0$. Pak nutně $Uy \neq 0$ a můžeme psát $UV^TUy = \lambda Uy$, neboli $UV^Tx = \lambda x$, kde $x = Uy \neq 0$.

(b) Stopa matice je rovna součtu jejích vlastních čísel. Tvrzení (b) tedy plyne z (a).

(c) Zřejmě $(I + UV^T)x = \lambda x$ právě tehdy, když $UV^Tx = (\lambda - 1)x$, a $(I + V^TU)y = \lambda y$ právě tehdy, když $V^TUy = (\lambda - 1)y$. Tvrzení (c) tedy plyne z (a).

(d) Determinant matice je roven součinu jejích vlastních čísel. Tvrzení (d) tedy plyne z (c).

(e) Je-li matice $I + UV^T$ regulární, je podle (c) i matice $(I + V^TU)$ regulární a platí

$$(I + UV^T)(I - U(I + V^TU)^{-1}V^T) = I + UV^T - U(I + V^TU)(I + V^TU)^{-1}V^T = I.$$

□

Lemma 31 má několik důležitých důsledků.

Důsledek 8. *(Woodbury) Nechť jsou splněny předpoklady lemmatu 31 a nechť $H \in R^{n\times n}$ je regulární matice. Pak*

$$\det(H + UV^T) = \det H \ \det(I + V^TH^{-1}U)$$

a je-li matice $H + UV^T$ regulární, platí

$$(H + UV^T)^{-1} = H^{-1} - H^{-1}U(I + V^TH^{-1}U)^{-1}V^TH^{-1}.$$

Důkaz Platí $H + UV^T = H(I + H^{-1}UV^T)$, takže můžeme použít (d) a (e) z lemmatu 31 (matice U se nahradí maticí $H^{-1}U$). □

Poznámka 106. (Sherman-Morrison) Mají-li matice U a V pouze jeden sloupec (takže $U = u$ a $V = v$, kde u a v jsou vektory), můžeme předchozí vztahy zapsat ve tvaru

$$\det(H + uv^T) = \det H \ \det(1 + v^TH^{-1}u)$$

a

$$(H + uv^T)^{-1} = H^{-1} - \frac{H^{-1}uv^TH^{-1}}{1 + v^TH^{-1}u}.$$

Důsledek 9. *Nechť $U \in R^{n\times m}$ a $M \in R^{m\times m}$ je symetrická matice. Pak:*

(a) *Matice* UMU^T *má stejná nenulová vlastní čísla jako matice* MU^TU *(nebo jako matice* U^TUM*).*
(b) *Platí* $\operatorname{Tr} UMU^T = \operatorname{Tr} MU^TU = \operatorname{Tr} U^TUM$*.*
(c) *Matice* $I + UMU^T$ *má stejná nejednotková vlastní čísla jako matice* $I + MU^TU$ *(nebo jako matice* $I + U^TUM$*).*
(d) *Platí* $\det(I + UMU^T) = \det(I + MU^TU) = \det(I + U^TUM)$*. Je-li matice* M *regulární, platí* $\det(I + UMU^T) = \det M \ \det(M^{-1} + U^TU)$*.*
(e) *Jsou-li matice* M *a* $I + UMU^T$ *regulární, platí* $(I + UMU^T)^{-1} = I - U(M^{-1} + U^TU)^{-1}U^T$*.*

Důkaz Stačí v lemmatu 31 použít UM místo V (nebo UM místo U a U místo V). Tvrzení (d) a (e) jsou jednodušší verzí důsledku 12. □

Důsledek 10. *Nechť jsou splněny předpoklady lemmatu 31. Pak platí*

$$\|UV^T\|_F^2 = Tr(VU^TUV^T) = Tr(U^TUV^TV).$$

Jestliže $U = [u_1, u_2]$*,* $V = [v_1, v_2]$*, dostaneme*

$$\|UV^T\|_F^2 = \|u_1v_1^T + u_2v_2^T\|_F^2 = u_1^Tu_1v_1^Tv_1 + 2u_1^Tu_2v_1^Tv_2 + u_2^Tu_2v_2^Tv_2.$$

Důkaz (a) Frobeniova norma matice A je definovaná vztahem $\|A\|_F^2 = Tr\, A^TA$, takže $\|UV^T\|_F^2 = \operatorname{Tr}(VU^TUV^T)$. Podle důsledku 9 (b) platí $\operatorname{Tr}(VU^TUV^T = \operatorname{Tr}(U^TUV^TV$, takže $\|UV^T\|_F^2 = \operatorname{Tr}(U^TUV^TV)$.

(b) Dosadíme-li $U = [u_1, u_2]$, $V = [v_1, v_2]$, dostaneme

$$\begin{aligned}\|u_1v_1^T + u_2v_2^T\|_F^2 &= \|UV^T\|_F^2 = Tr(U^TUV^TV) = Tr\begin{bmatrix} u_1^Tu_1 & u_1^Tu_2 \\ u_2^Tu_1 & u_2^Tu_2 \end{bmatrix}\begin{bmatrix} v_1^Tv_1 & v_1^Tv_2 \\ v_2^Tv_1 & v_2^Tv_2 \end{bmatrix} \\ &= u_1^Tu_1v_1^Tv_1 + 2u_1^Tu_2v_1^Tv_2 + u_2^Tu_2v_2^Tv_2.\end{aligned}$$

□

Poznámka 107. Druhý vzorec uvedený v důsledku 10 můžeme zobecnit pomocí binomické věty. Například pro tři dvojice vektorů platí

$$\|u_1v_1^T + u_2v_2^T + u_3v_3^T\|_F^2 = u_1^Tu_1v_1^Tv_1 + u_2^Tu_2v_2^Tv_2 + u_3^Tu_3v_3^Tv_3 + 2u_1^Tu_2v_1^Tv_2 + 2u_1^Tu_3v_1^Tv_3 + 2u_2^Tu_3v_2^Tv_3.$$

Poznámka 108. Dosadíme-li do vzorce (d) v důsledku 9 matice $U = [u, v]$ a $M = \operatorname{diag}(\varepsilon_1, \varepsilon_2)$, kde čísla ε_1 a ε_2 (znaménka) nabývají hodnot 1 nebo -1, dostaneme

$$\det(I + \varepsilon_1uu^T + \varepsilon_2vv^T) = (1 + \varepsilon_1u^Tu)(1 + \varepsilon_2v^Tv) - \varepsilon_1\varepsilon_2(u^Tv)^2.$$

Dosadíme-li do téhož vzorce matice $U = [u, v]$ a $M = \varepsilon_1e_1e_2^T + \varepsilon_2e_2e_1^T$, kde $e_1 = [1, 0]^T$ a $e_2 = [0, 1]^T$, dostaneme

$$\det(I + \varepsilon_1uv^T + \varepsilon_2vu^T) = (1 + \varepsilon_1v^Tu)(1 + \varepsilon_2u^Tv) - \varepsilon_1\varepsilon_2u^Tuv^Tv,$$

takže například

$$\begin{aligned}\det(I + uv^T + vu^T) &= (1 + v^Tu)^2 - u^Tuv^Tv. \\ \det(I + uv^T - vu^T) &= 1 - (v^Tu)^2 + u^Tuv^Tv, \\ \det(I - uv^T - vu^T) &= (1 - v^Tu)^2 - u^Tuv^Tv.\end{aligned}$$

Důsledek 11. *Nechť $u \in R^n$, $v \in R^n$, $v^T u \neq 0$ a*

$$P = I - \frac{u\, v^T}{v^T u},$$

Pak $P^2 = P$, $(I-P)^2 = I - P$, $P(I-P) = 0$ a

$$\|P\| = \|I - P\| = \frac{\|u\|\|v\|}{|v^T u|}.$$

Důkaz (a) Předpokládejme nejprve, že $v^T u = 1$. Pak $P^2 = (I - uv^T)^2 = I - 2uv^T + uv^T uv^T = I - uv^T = P$, $(I-P)^2 = I - 2P + P^2 = I - P$ a $P(I-P) = P - P = 0$. Pro libovolnou matici A platí $\|A\| = \sqrt{\overline{\lambda}(AA^T)}$, kde $\overline{\lambda}(AA^T)$ je největší vlastí číslo (pozitivně semidefinitní) matice AA^T. Zřejmě

$$PP^T = (I - uv^T)(I - vu^T) = I - uv^T - vu^T + v^T vuu^T = I + [u, v]\begin{bmatrix} v^T v & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} u^T \\ v^T \end{bmatrix}.$$

Podle důsledku 9 má tato matice stejná nejednotková vlastní čísla jako matice

$$I + \begin{bmatrix} v^T v & -1 \\ -1 & 0 \end{bmatrix}\begin{bmatrix} u^T u & 1 \\ 1 & v^T v \end{bmatrix} = \begin{bmatrix} u^T uv^T v & 0 \\ -u^T u & 0 \end{bmatrix},$$

jejíž vlastní čísla jsou řešením charakteristické rovnice

$$\det \begin{bmatrix} \lambda - u^T uv^T v & 0 \\ u^T u & \lambda \end{bmatrix} = \lambda^2 - \lambda u^T uv^T v = 0.$$

Tato kvadratická rovnice má dva kořeny 0 a $u^T uv^T v \geq (v^T u)^2 = 1$, takže $\|P\| = \sqrt{u^T uv^T v} = \|u\|\|v\|$. Jelikož matice $(I-P)(I-P)^T = uv^T vu^T$ má podle lemmatu 31 stejná nenulová vlastní čísla jako matice $\left[u^T uv^T v\right]$, platí $\|I - P\| = \sqrt{u^T uv^T v} = \|u\|\|v\|$.

(b) Pokud $v^T u \neq 1$, lze předchozí postup aplikovat na matici $P = u\tilde{v}^T$, kde $\tilde{v} = v/v^T u$. □

Důsledek 12. *Nechť jsou splněny předpoklady důsledku 9 a nechť $H \in R^{n\times n}$ je regulární symetrická matice. Pak*

$$\det(H + UMU^T) = \det H\, \det(I + MU^T H^{-1}U) = \det H\, \det(I + U^T H^{-1}UM)$$

a je-li matice M regulární, platí

$$\det(H + UMU^T) = \det H\, \det M\, \det(M^{-1} + U^T H^{-1}U).$$

Jsou-li matice M a $H + UMU^T$ regulární, platí

$$(H + UMU^T)^{-1} = H^{-1} - H^{-1}U(M^{-1} + U^T H^{-1}U)^{-1}U^T H^{-1}.$$

Důkaz Označme $V = H^{-1}UM$. Jelikož $H + UMU^T = (I + UMU^T H^{-1})H = (I + UV^T)H$, můžene podle lemmatu 31 psát

$$\begin{aligned} \det(H + UMU^T) &= \det H\, \det(I + UV^T) = \det H\, \det(I + V^T U) \\ &= \det H\, \det(I + MU^T H^{-1}U) \\ (H + UMU^T)^{-1} &= H^{-1}(I + UV^T)^{-1} = H^{-1}(I - U(I + V^T U)^{-1}V^T) \\ &= H^{-1} - H^{-1}U(I + MU^T H^{-1}U)^{-1}MU^T H^{-1}. \end{aligned}$$

Je-li matice M regulární, můžeme ji vytknout před závorku (takže matice M^{-1} se po inverzi dostane za závorku). □

Nyní se vrátíme k vyšetřování metod s proměnnou metrikou z Broydenovy třídy používajících aktualizaci $H_+ = \gamma(H + UMU^T)$, kde $U = [d, Hy]$. Budeme předpokládat, že H je symetrická pozitivně definitní matice a vektory d, y jsou nenulové (stačí předpokládat, že $y \neq 0$, neboť z $y = g_+ - g \neq 0$ plyne $d = x_+ - x \neq 0$).

Věta 78. *Nechť $H_+ = \gamma(H + UMU^T)$, kde H je symetrická pozitivně definitní matice a $U = [d, Hy]$, $d \neq 0$, $y \neq 0$. Pak rovnost $H_+ y = \rho d$ platí právě tehdy, když*

$$M = \begin{bmatrix} \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right), & -\frac{\eta}{b} \\ -\frac{\eta}{b}, & \frac{\eta - 1}{a} \end{bmatrix}, \tag{282}$$

kde η je volný parametr a kde

$$a = y^T Hy, \qquad b = y^T d, \qquad c = d^T H^{-1} d.$$

Důkaz Označme m_1, m_2, m_3 prvky matice M. Platí-li (272), můžeme podle (271) psát

$$\begin{aligned} \frac{1}{\gamma} H_+ y &= Hy + [d, Hy] \begin{bmatrix} m_1, & m_2 \\ m_2, & m_3 \end{bmatrix} \begin{bmatrix} b \\ a \end{bmatrix} \\ &= Hy + (m_1 b + m_2 a)d + (m_2 b + m_3 a)Hy = \frac{\rho}{\gamma} d, \end{aligned}$$

takže nutně

$$\begin{aligned} m_1 b + m_2 a &= \rho/\gamma, \\ m_2 b + m_3 a &= -1. \end{aligned}$$

Jeden parametr je nadbytečný. Zvolíme $m_2 = -\eta/b$ a zbylé prvky m_1, m_3 určíme řešením uvedených rovnic, takže

$$m_1 = \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right), \qquad m_2 = -\frac{\eta}{b}, \qquad m_3 = \frac{\eta - 1}{a}. \tag{283}$$

Tím dostaneme matici M uvedenou ve větě 78. Z druhé strany, vynásobíme-li (271), kde matice M je dána vztahem (282), vektorem y, dostaneme (272). □

Poznámka 109. Při vyšetřování metod s proměnnou metrikou budeme často používat označení

$$\delta = \frac{\rho}{\gamma}\left(\eta\frac{c}{b} + (1 - \eta)\frac{b}{a}\right), \tag{284}$$

$$\mu = \frac{1}{ab}\left(\eta\frac{a}{b} + (1 - \eta)\frac{\rho}{\gamma}\right). \tag{285}$$

Přímým výpočtem se snadno přesvědčíme, že $\mu = -\det M$, kde M je matice určená vztahem (282). Podle poznámky 115 platí $\det((1/\gamma)H_+) = \det(H + UMU^T) = \delta \det H$.

Poznámka 110. Číslo $c = d^T H^{-1} d$ se v matici M nevyskytuje, často se však používá k určení hodnot parametrů γ a η (je to popsáno v oddílu 4.4). Realizujeme-li metody s proměnnou metrikou jako metody spádových směrů, platí $s = -Hg$, takže $d = \alpha s = -\alpha Hg$. Můžeme tedy položit $c = -\alpha d^T g$ a není třeba řešit soustavu rovnic s maticí H.

Poznámka 111. Z pozitivní definitnosti matice H a z nenulovosti vektorů d a y plyne, že $a > 0$ a $c > 0$. Vybíráme-li délku kroku podle (S3b), platí

$$b = y^T d = \alpha(g_+ - g)^T s \geq \alpha(\varepsilon_2 - 1)g^T s > 0.$$

Můžeme proto předpokládat, že $a > 0$, $b > 0$, $c > 0$. Z pozitivní definitnosti matice H a ze Schwarzovy nerovnosti plyne, že $ac - b^2 \geq 0$. Jsou-li vektory d a Hy lineárně nezávislé, platí $ac - b^2 > 0$.

Poznámka 112. V dalším textu budeme často předpokládat, že vektory d a Hy jsou lineárně nezávislé, neboli $ac - b^2 > 0$. Jsou-li vektory d a Hy lineárně závislé, má matice UMU^T hodnost 1 a všechny aktualizace z Broydenovy třídy jsou ekvivalentní. Jestliže $Hy = \lambda d$, kde $\lambda \neq 0$, platí $a = \lambda b$ a vztah (286) lze zapsat ve tvaru

$$\frac{1}{\gamma}H_+ = H + \frac{1}{b}\left(\frac{\rho}{\gamma} - \frac{a}{b}\right)dd^T.$$

Zvolíme-li $\rho/\gamma = a/b$, lze kvazinewtonovskou podmínku $H_+y = \rho d$ splnit prostým vynásobením matice H číslem $\gamma = \rho\, b/a$.

Poznámka 113. Vztah $H_+ = \gamma(H + UMU^T)$ můžeme roznásobit. Pak platí

$$\frac{1}{\gamma}H_+ = H + \frac{\rho}{\gamma b}dd^T - \frac{1}{a}Hy(Hy)^T + \frac{\eta}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \tag{286}$$

(Broydenova třída). Nejznámější členy Broydenovy třídy dostaneme, položíme-li $\eta = \eta^{DFP} = 0$ (metoda Davidona [38], Fletchera a Powella [56]), takže

$$\frac{1}{\gamma}H_+^{DFP} = H + \frac{\rho}{\gamma b}dd^T - \frac{1}{a}Hy(Hy)^T, \tag{287}$$

nebo $\eta = \eta^{BFGS} = 1$ (metoda Broydena [13], Fletchera [52], Goldfarba [68] a Shanno [141]), takže

$$\frac{1}{\gamma}H_+^{BFGS} = H + \left(\frac{\rho}{\gamma} + \frac{a}{b}\right)\frac{1}{b}dd^T - \frac{1}{b}\left(Hyd^T + d(Hy)^T\right), \tag{288}$$

nebo $\eta = \eta^{R1} = (\rho/\gamma)/(\rho/\gamma - a/b)$ (metoda hodnosti 1 [12]), takže

$$\frac{1}{\gamma}H_+^{R1} = H + \frac{1}{(\rho/\gamma)b - a}\left(\frac{\rho}{\gamma}d - Hy\right)\left(\frac{\rho}{\gamma}d - Hy\right)^T, \tag{289}$$

nebo $\eta = \eta^{H} = (\rho/\gamma)/(\rho/\gamma + a/b)$ (Hoshinova metoda [83]), takže

$$\frac{1}{\gamma}H_+^{H} = H + \frac{2\rho}{\gamma b}dd^T - \frac{1}{(\rho/\gamma)b + a}\left(\frac{\rho}{\gamma}d + Hy\right)\left(\frac{\rho}{\gamma}d + Hy\right)^T. \tag{290}$$

Z těchto čtyř konkrétních metod jsou bez dalších úprav prakticky použitelné pouze metoda BFGS a Hoshinova metoda. Metoda DFP vyžaduje přesný výběr délky kroku nebo důsledné škálování, jinak konverguje velmi pomalu. Metoda hodnosti 1 obecně nesplňuje podmínku pro pozitivní definitnost matice H_+ (zdůvodnění je uvedeno v poznámce 117), takže může dojít ke ztrátě globální konvergence vlivem porušení podmínky spádovosti (S1a).

Poznámka 114. Pro konstrukci metod s omezenou pamětí, popsaných v oddílu 9.1, je užitečný pseudosoučinový tvar

$$\frac{1}{\gamma}H_+ = \left(I - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)y^T\right)H\left(I - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)y^T\right)^T + \frac{\rho}{\gamma b}dd^T \tag{291}$$

který lze použít pouze tehdy, když $\eta \geq 0$. Tento vzorec se velmi zjednoduší pro metodu BFGS, kdy $\eta = 1$ (vzorec (292) nebo (293) pro $\eta = 1$). Platí také

$$\frac{1}{\gamma}H_+ = \left(I - \frac{1}{b}dy^T\right)\left(H + \frac{\eta - 1}{a}Hy(Hy)^T\right)\left(I - \frac{1}{b}yd^T\right) + \frac{\rho}{\gamma b}dd^T. \tag{292}$$

Vzorec (286) lze zapsat ve tvaru

$$\begin{aligned}\frac{1}{\gamma}H_+ &= \frac{1}{\gamma}H_+^{DFP} + \frac{\eta}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \\ &= \frac{1}{\gamma}H_+^{BFGS} + \frac{\eta - 1}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \\ &= \left(I - \frac{1}{b}dy^T\right)H\left(I - \frac{1}{b}yd^T\right) + \frac{\rho}{\gamma b}dd^T + \frac{\eta - 1}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T,\end{aligned} \tag{293}$$

kde první část posledního řádku je pseudosoučinový tvar pro metodu BFGS. Platí také

$$\begin{aligned}\frac{1}{\gamma}H_+ &= H - \frac{\mu}{m_3}dd^T + m_3\left(\frac{m_2}{m_3}d + Hy\right)\left(\frac{m_2}{m_3}d + Hy\right)^T \\ &= H + \frac{\mu a}{1 - \eta}dd^T - \frac{1 - \eta}{a}\left(\frac{\eta a}{(1 - \eta)b}d + Hy\right)\left(\frac{\eta a}{(1 - \eta)b}d + Hy\right)^T,\end{aligned} \tag{294}$$

nebo

$$\begin{aligned}\frac{1}{\gamma}H_+ &= H + m_1\left(d + \frac{m_2}{m_1}Hy\right)\left(d + \frac{m_2}{m_1}Hy\right)^T - \frac{\mu}{m_1}Hy(Hy)^T \\ &= H + \frac{1}{(\rho/\gamma)b + \eta a}\left(\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right)d - \eta Hy\right)\left(\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right)d - \eta Hy\right)^T \\ &\quad - \frac{\mu b^2}{(\rho/\gamma)b + \eta a}Hy(Hy)^T.\end{aligned} \tag{295}$$

kde m_1, m_2, m_3 jsou čísla určená vztahy (283). První vzorec je zobecněním vztahu pro metodu DFP (používá se pro $\eta < 1$) a druhý vzorec je zobecněním vztahu pro metodu BFGS (používá se pro $\eta > 0$). Tyto dvoučlenné vzorce jsou vhodné pro praktické použití, neboť vyžadují zhruba $2n^2$ aritmetických opperací, zatímco tříčlenný vztah (286) vyžaduje zhruba $3n^2$ aritmetických operací. Pomocí vzorce (295) lze aktualizaci metody BFGS vyjádřit ve tvaru

$$\frac{1}{\gamma}H_+^{BFGS} = H + \frac{1}{(\rho/\gamma)b + a}\left(\left(\frac{\rho}{\gamma} + \frac{a}{b}\right)d - Hy\right)\left(\left(\frac{\rho}{\gamma} + \frac{a}{b}\right)d - Hy\right)^T - \frac{1}{(\rho/\gamma)b + a}Hy(Hy)^T, \tag{296}$$

Teoretický význam má též vzorec

$$\frac{1}{\gamma}H_+^{BFGS} = H + \frac{1}{b}\left(d\left(\frac{\rho}{\gamma}d - Hy\right)^T + \left(\frac{\rho}{\gamma}d - Hy\right)d^T\right) - \left(\frac{\rho}{\gamma} - \frac{a}{b}\right)\frac{1}{b}dd^T, \tag{297}$$

který lze získat variačním odvozením (poznámka 142). Aktualizaci metody DFP lze vyjádřit v součinovém tvaru

$$\frac{1}{\gamma}H_+^{DFP} = \left(I - \frac{1}{a}\left(Hy \pm \sqrt{\frac{\rho a}{\gamma b}}d\right)y^T\right)H\left(I - \frac{1}{a}\left(Hy \pm \sqrt{\frac{\rho a}{\gamma b}}d\right)y^T\right)^T. \tag{298}$$

O správnosti všech těchto vztahů se můžeme přesvědčit jejich roznásobením a porovnáním odpovídajících si členů.

Lemma 32. *Nechť H je symetrická pozitivně definití matice, $a > 0$, $b > 0$, $c > 0$ a nechť H_+ je matice získaná pomocí aktualizace (286), kde $\gamma > 0$ a $\rho > 0$. Pak matice $(1/\gamma)H^{-1/2}H_+H^{-1/2}$ má $n-2$ jednotkových vlastních čísel a zbylá dvě vlastní čísla jsou řešením kvadratické rovnice.*

$$\lambda^2 - \sigma\lambda + \delta = 0,$$

kde

$$\sigma = \frac{1}{b^2}(\eta(ac-b^2)+b^2) + \frac{\rho}{\gamma}\frac{c}{b} = \frac{\rho}{\gamma}\frac{c}{b} + \left(1 - \frac{\eta}{\eta^*}\right), \tag{299}$$

$$\delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac-b^2)+b^2) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right) \tag{300}$$

a kde

$$\eta^* = -\frac{b^2}{ac-b^2} < 0 \tag{301}$$

je kritická hodnota parametru η (pokud $ac - b^2 = 0$, můžeme položit $1/\eta^* = 0$ a $\eta^* = -\infty$).

Důkaz Podle (271) platí

$$\frac{1}{\gamma}H^{-1/2}H_+H^{-1/2} = I + H^{-1/2}UMU^TH^{-1/2}. \tag{302}$$

Tato matice má $n-2$ jednotkových vlastních čísel odpovídajících $n-2$ vlastním vektorům kolmým k $H^{-1/2}U$. Zbylá dvě vlastní čísla jsou podle důsledku 9 vlastními čísly matice $I + MU^TH^{-1}U$, takže pro ně musí platit $\det((1-\lambda)I + MU^TH^{-1}U) = 0$. Použijeme-li vyjádření (282) a přihlédneme-li k tomu, že

$$U^TH^{-1}U = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix}, \tag{303}$$

můžeme psát

$$\begin{aligned} \det((1-\lambda)I + MU^TH^{-1}U) &= \det\left(\begin{bmatrix} 1-\lambda, & 0 \\ 0, & 1-\lambda \end{bmatrix} + M\begin{bmatrix} c, & b \\ b, & a \end{bmatrix}\right) \\ &= \det\begin{bmatrix} \eta\dfrac{ac-b^2}{b^2} + \dfrac{\rho c}{\gamma b} + 1 - \lambda, & \dfrac{\rho}{\gamma} \\ -\eta\dfrac{ac-b^2}{ab} - \dfrac{b}{a}, & -\lambda \end{bmatrix} = 0, \end{aligned}$$

což po úpravě dává $\lambda^2 - \sigma\lambda + \delta = 0$ s koeficienty uvedenými v lemmatu 32. □

Poznámka 115. Poznamenejme, že δ se jako součin vlastních čísel matice $I + MU^TH^{-1}U$ podle důsledku 9 rovná determinantu matice $(1/\gamma)H^{-1/2}H_+H^{-1/2} = I + H^{-1/2}UMU^TH^{-1/2}$. Platí tedy $\det((1/\gamma)H_+) = \delta \det H$, což lze zapsat ve tvaru.

$$\det\left(\frac{1}{\gamma}H_+\right) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right)\det H \tag{304}$$

(pokud $ac - b^2 = 0$, můžeme položit $1/\eta^* = 0$).

Věta 79. *Nechť jsou splněny předpoklady lemmatu 32. Pak matice H_+ je pozitivně definitní právě tehdy, je-li splněna nerovnost $\eta(ac-b^2)+b^2 > 0$, neboli $\eta > \eta^*$ (pokud $ac - b^2 = 0$, můžeme položit $\eta^* = -\infty$).*

Důkaz Je třeba najít podmínku pro to, aby rovnice $\lambda^2 - \sigma\lambda + \delta$ s koeficienty uvedenými v lemmatu 32 měla kladné kořeny. Označme λ_1 a λ_2 tyto kořeny. Pak $\lambda_1 + \lambda_2 = \sigma$ a $\lambda_1\lambda_2 = \delta$ takže $\lambda_1 > 0$ a $\lambda_2 > 0$ právě tehdy. když $\sigma > 0$ a $\delta > 0$. Z definice čísel σ a δ plyne, že

$$\sigma = \frac{\gamma}{\rho}\frac{a}{b}\delta + \frac{\rho}{\gamma}\frac{c}{b}. \tag{305}$$

Jelikož předpokládáme, že $a > 0$, $b > 0$, $c > 0$, $\gamma > 0$, $\rho > 0$, platí $\sigma > 0$ kdykoliv $\delta > 0$. Z $\delta > 0$ dostaneme podmínku $\eta(ac-b^2)+b^2 > 0$, neboli $\eta > \eta^*$. □

Poznámka 116. Ve větě 79 předpokládáme, že $b > 0$. Pokud $b = 0$, není matice H_+ definována. Pokud $b < 0$ a $\delta > 0$, plyne z důkazu věty 79, že $\sigma < 0$, takže matice H_+ není pozitivně definitní. Podmínka $b > 0$ je tedy pro pozitivní definitnost nutná. Naštěstí lze tuto podmínku snadno zajistit (poznámka 111).

Poznámka 117. Z věty 79 plyne, že matice H_+ je pozitivně definitní, pokud $\eta \geq 0$ (neboť $\eta^* < 0$). To znamená, že metoda DFP, metoda BFGS i Hoshinova metoda generují pozitivně definitní matice. Metoda hodnosti 1 tuto vlastnost nemá, neboť přímým dosazením hodnoty $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$ do výrazu pro δ zjistíme, že platí $\delta = ((\rho/\gamma)c - b)/(b - (\gamma/\rho)a)$, takže $\delta > 0$ pouze tehdy, když buď $0 < \rho/\gamma < b/c$, takže $\eta^* < \eta < 0$, nebo $a/b < \rho/\gamma$, takže $1 < \eta$ (ze Schwarzovy nerovnosti plyne, že $b/c \leq a/b$).

V některých aplikacích, například při použití metod s lokálně omezeným krokem nebo při minimalizaci s nelineárními omezeními, je velmi důležitý inverzní tvar rekurentního vztahu (286).

Věta 80. *(Aktualizace matice $B = H^{-1}$). Nechť jsou splněny předpoklady lemmatu 32. Nechť $B = H^{-1}$ a $B_+ = H_+^{-1}$. Pak platí*

$$\gamma B_+ = B + \frac{\gamma}{\rho b} y y^T - \frac{1}{c} Bd(Bd)^T + \frac{\beta}{c}\left(\frac{c}{b}y - Bd\right)\left(\frac{c}{b}y - Bd\right)^T, \tag{306}$$

kde

$$\beta\eta(ac - b^2) + (\beta + \eta)b^2 = b^2. \tag{307}$$

Důkaz Inverzí vztahu $(1/\gamma)H_+ = H + UMU^T$ podle důsledku 12 dostaneme

$$\gamma B_+ = B - BU(M^{-1} + U^T BU)^{-1}U^T B \stackrel{\Delta}{=} B + BUKU^T B,$$

kde $K \in R^{2\times 2}$. Jelikož podle (272) platí $H_+ y = \rho d$, musí platit $B_+ d = (1/\rho)y$ neboli

$$\gamma B_+ d = Bd + [Bd, y]\begin{bmatrix} k_1, & k_2 \\ k_2, & k_3 \end{bmatrix}\begin{bmatrix} c \\ b \end{bmatrix} = Bd + (k_1 c + k_2 b)Bd + (k_2 c + k_3 b)y = \frac{\gamma}{\rho}y,$$

takže nutně

$$k_1 c + k_2 b = -1,$$
$$k_2 c + k_3 b = \gamma/\rho.$$

Zvolíme $k_2 = -\beta/b$ a zbylé prvky k_1, k_3 určíme řešením uvedených rovnic. Tím dostaneme

$$K = \begin{bmatrix} \dfrac{\beta - 1}{c}, & -\dfrac{\beta}{b} \\ -\dfrac{\beta}{b}, & \dfrac{1}{b}\left(\beta\dfrac{c}{b} + \dfrac{\gamma}{\rho}\right) \end{bmatrix},$$

což po dasazení do $\gamma B_+ = B + BUKU^T B$ dává (306). Vztah (307), svazující β s η lze získat například z rovnosti

$$K = -(M^{-1} + U^T BU)^{-1}.$$

Jednodušší způsob je uveden v poznámce 120. □

Poznámka 118. Jelikož $Bs = -g$ (poznámka 110), lze ve vztahu (306) nahradit vektor Bd vektorem $-\alpha g$, takže odpadne maticové násobení.

Poznámka 119. (Dualita) Vztah (306) dostaneme ze vztahu (286) záměnou $\gamma \to 1/\gamma$, $\rho \to 1/\rho$, $a \to c$, $c \to a$, $d \to y$, $y \to d$, $H \to B$, $\eta \to \beta$. Metody DFP a BFGS jsou navzájem duální. Metodu DFP dostaneme pro $\beta = \beta^{DFP} = 1$, takže

$$\gamma B_+^{DFP} = B + \left(\frac{\gamma}{\rho} + \frac{c}{b}\right)\frac{1}{b}yy^T - \frac{1}{b}\left(Bdy^T + y(Bd)^T\right). \tag{308}$$

Metodu BFGS dostaneme pro $\beta = \beta^{BFGS} = 0$, takže

$$\gamma B_+^{BFGS} = B + \frac{\gamma}{\rho b} yy^T - \frac{1}{c} Bd(Bd)^T. \tag{309}$$

Metoda hodnosti 1 je samoduální, dostaneme ji pro $\beta = \beta^{R1} = (\gamma/\rho)/(\gamma/\rho - c/b)$, takže

$$\gamma B_+^{R1} = B + \frac{1}{(\gamma/\rho)b - c} \left(\frac{\gamma}{\rho} y - Bd\right) \left(\frac{\gamma}{\rho} y - Bd\right)^T. \tag{310}$$

Hoshinova metoda je také samoduální, dostaneme ji pro $\beta = \beta^{H} = (\gamma/\rho)/(\gamma/\rho + c/b)$, takže

$$\gamma B_+^{H} = B + \frac{2\gamma}{\rho b} yy^T - \frac{1}{(\gamma/\rho)b + c} \left(\frac{\gamma}{\rho} y + Bd\right) \left(\frac{\gamma}{\rho} y + Bd\right)^T. \tag{311}$$

Poznámka 120. Z duality lze snadno určit vztah mezi β a η. Platí totiž

$$\det(\gamma B_+) = \frac{\gamma}{\rho} \frac{b}{c} \left(1 - \frac{\beta}{\beta^*}\right) \det B, \tag{312}$$

kde

$$\beta^* = \eta^* = -\frac{b^2}{ac - b^2} < 0, \tag{313}$$

což spolu s výrazem (304) pro $\det H_+$ a s identitami $\det B \det H = 1$, $\det B_+ \det H_+ = 1$ dává

$$\frac{b^2}{ac} \left(1 - \frac{\beta}{\beta^*}\right) \left(1 - \frac{\eta}{\eta^*}\right) = 1. \tag{314}$$

Po dosazení, roznásobení a úpravě tohoto vztahu dostaneme rovnost (307). Z rovnosti (307) plynou převodní vztahy

$$\eta = \frac{b^2(1-\beta)}{\beta(ac - b^2) + b^2} = \frac{\beta^*(\beta - 1)}{\beta - \beta^*}, \tag{315}$$

$$\beta = \frac{b^2(1-\eta)}{\eta(ac - b^2) + b^2} = \frac{\eta^*(\eta - 1)}{\eta - \eta^*}. \tag{316}$$

Z vyjádření (312) vyplývá, že matice B_+ je pozitivně definitní právě tehdy, když $\beta > \beta^*$.

Poznámka 121. Ze vztahu (314) plyne, že pro $\eta \geq \eta*$ platí $\beta \geq \beta^*$ a tudíž $1 - \eta/\eta^* \geq 0$ a $1 - \beta/\beta^* \geq 0$. Jestliže v tomto intervalu η roste pak β klesá a naopak, takže hodnoty $\eta = \eta^*$ a $\eta = \infty$ jsou duální k hodnotám $\beta = \infty$ a $\beta = \beta^*$.

Poznámka 122. Z úvahy použité v poznámce 120 plyne, že při přechodu od vztahu (286) ke vztahu (306) provádíme záměnu $\delta \to 1/\delta$. Z důkazu věty 80 víme, že $-K^{-1} = M^{-1} + U^T BU = M^{-1}(I + MU^T BU)$, což podle lemmatu 32 dává $\det(K^{-1}) = \delta \det(M^{-1})$ (neboť $K \in R^{2\times 2}$, takže $\det(-K^{-1}) = \det K^{-1}$). Odtud plyne, že při přechodu od vztahu (286) ke vztahu (306) provádíme záměnu $\mu \to \mu/\delta$. Z duality plyne, že

$$\frac{1}{\delta} = \frac{\gamma}{\rho} \left(\beta \frac{a}{b} + (1-\beta)\frac{b}{c}\right), \tag{317}$$

$$\frac{\mu}{\delta} = \frac{1}{bc} \left(\beta \frac{c}{b} + (1-\beta)\frac{\gamma}{\rho}\right). \tag{318}$$

Jelikož vlastní čísla matice $\gamma B^{-1/2} B_+ B^{-1/2}$ jsou převrácenými hodnotami vlastních čísel matice (302), jsou to buď jednotky nebo řešení kvadratické rovnice

$$\lambda^2 - \frac{\sigma}{\delta}\lambda + \frac{1}{\delta} = 0.$$

Z duality plyne, že

$$\frac{1}{\delta} = \frac{\gamma}{\rho}\frac{1}{bc}(\beta(ac-b^2)+b^2) = \frac{\gamma}{\rho}\frac{b}{c}\left(1-\frac{\beta}{\beta^*}\right), \tag{319}$$

$$\frac{\sigma}{\delta} = \frac{1}{b^2}(\beta(ac-b^2)+b^2) + \frac{\gamma}{\rho}\frac{a}{b} = \frac{\gamma}{\rho}\frac{a}{b} + \left(1-\frac{\beta}{\beta^*}\right). \tag{320}$$

Poznámka 123. V tomto oddílu jsme se zabývali metodami s proměnnou metrikou tvaru (271), splnujícími kvazinewtonovskou podmínku (272). Ukážeme, že nelze obecně zkonstruovat symetrickou aktualizaci tvaru (271) splňující současně dvě kvazinewtonovské podmínky. Omezíme se přitom na případ, kdy $\gamma_i = 1$, $\rho_i = 1$, $i \in N$. Předpokládejme, že $H_+y = d$ a $H_+y_- = d_-$, takže $y_-^T H_+ y = y_-^T d$ a $y^T H_+ y_- = y^T d_-$. Je-li matice H_+ symetrická, dostaneme $y_-^T d = y^T d_-$, což platí, je-li minimalizovaná funkce kvadratická (kdy $y = Gd$ a $y_- = Gd_-$), nikoliv však obecně.

Poznámka 124. V následující tabulce jsou uvedeny hodnoty parametrů nejznámějších metod s proměnnou metrikou z Broydenovu třídy (OC je metoda definovaná vzorcem (377)).

	DFP	BFGS	R1	H	OC
η	0	1	$\dfrac{\rho b}{\rho b-\gamma a}$	$\dfrac{\rho b}{\rho b+\gamma a}$	$\dfrac{b}{\gamma}\dfrac{\rho c-\gamma b}{ac-b^2}$
m_1	$\dfrac{\rho}{\gamma b}$	$\dfrac{1}{b}\left(\dfrac{a}{b}+\dfrac{\rho}{\gamma}\right)$	$\dfrac{\rho}{\gamma}\dfrac{\rho}{\rho b-\gamma a}$	$\dfrac{2a}{b^2}-\dfrac{\rho}{\gamma}\dfrac{\rho}{\rho b+\gamma a}$	$\dfrac{a}{\gamma b}\left(\dfrac{\rho c-\gamma b}{ac-b^2}+\dfrac{\rho}{a}\right)$
m_2	0	$-\dfrac{1}{b}$	$-\dfrac{\rho}{\rho b-\gamma a}$	$-\dfrac{\rho}{\rho b+\gamma a}$	$-\dfrac{\rho c-\gamma b}{\gamma(ac-b^2)}$
m_3	$-\dfrac{1}{a}$	0	$\dfrac{\gamma}{\rho}\dfrac{\rho}{\rho b-\gamma a}$	$-\dfrac{\gamma}{\rho}\dfrac{\rho}{\rho b+\gamma a}$	$\dfrac{c}{\gamma a}\dfrac{\rho b-\gamma a}{ac-b^2}$
μ	$\dfrac{\rho}{\gamma ab}$	$\dfrac{1}{b^2}$	0	$\dfrac{2}{b^2}\dfrac{\rho b}{\rho b+\gamma a}$	$\dfrac{1}{b^2}\left(\dfrac{\rho^2 c}{\gamma^2 a}-\dfrac{(\rho c-\gamma b)^2}{\gamma^2(ac-b^2)}\right)$
δ	$\dfrac{\rho b}{\gamma a}$	$\dfrac{\rho c}{\gamma b}$	$\dfrac{\rho}{\gamma}\dfrac{\rho c-\gamma b}{\rho b-\gamma a}$	$\dfrac{\rho}{\gamma}\dfrac{\rho c+\gamma b}{\rho b+\gamma a}$	$\dfrac{\rho^2 c}{\gamma^2 a}$
σ	$1+\dfrac{\rho c}{\gamma b}$	$\dfrac{ac}{b^2}+\dfrac{\rho c}{\gamma b}$	$\dfrac{a}{b}\dfrac{\rho c-\gamma b}{\rho b-\gamma a}+\dfrac{\rho c}{\gamma b}$	$\dfrac{a}{b}\dfrac{\rho c+\gamma b}{\rho b+\gamma a}+\dfrac{\rho c}{\gamma b}$	$2\dfrac{\rho c}{\gamma b}$
β	1	0	$\dfrac{\gamma b}{\gamma b-\rho c}$	$\dfrac{\gamma b}{\gamma b+\rho c}$	$\dfrac{b}{\rho}\dfrac{\gamma a-\rho b}{ac-b^2}$
k_1	$\dfrac{1}{b}\left(\dfrac{c}{b}+\dfrac{\gamma}{\rho}\right)$	$\dfrac{\gamma}{\rho b}$	$\dfrac{\gamma}{\rho}\dfrac{\gamma}{\gamma b-\rho c}$	$\dfrac{2c}{b^2}-\dfrac{\gamma}{\rho}\dfrac{\gamma}{\gamma b+\rho c}$	$\dfrac{c}{\rho b}\left(\dfrac{\gamma a-\rho b}{ac-b^2}+\dfrac{\gamma}{c}\right)$
k_2	$-\dfrac{1}{b}$	0	$-\dfrac{\gamma}{\gamma b-\rho c}$	$-\dfrac{\gamma}{\gamma b+\rho c}$	$-\dfrac{\gamma a-\rho b}{\rho(ac-b^2)}$
k_3	0	$-\dfrac{1}{c}$	$\dfrac{\rho}{\gamma}\dfrac{\gamma}{\gamma b-\rho c}$	$-\dfrac{\rho}{\gamma}\dfrac{\gamma}{\gamma b+\rho c}$	$\dfrac{a}{\rho c}\dfrac{\gamma b-\rho c}{ac-b^2}$
$\dfrac{\mu}{\delta}$	$\dfrac{1}{b^2}$	$\dfrac{\gamma}{\rho bc}$	0	$\dfrac{2}{b^2}\dfrac{\gamma b}{\gamma b+\rho c}$	$\dfrac{1}{b^2}\left(\dfrac{\gamma^2 a}{\rho^2 c}-\dfrac{(\gamma a-\rho b)^2}{\rho^2(ac-b^2)}\right)$
$\dfrac{1}{\delta}$	$\dfrac{\gamma a}{\rho b}$	$\dfrac{\gamma b}{\rho c}$	$\dfrac{\gamma}{\rho}\dfrac{\gamma a-\rho b}{\gamma b-\rho c}$	$\dfrac{\gamma}{\rho}\dfrac{\gamma a+\rho b}{\gamma b+\rho c}$	$\dfrac{\gamma^2 a}{\rho^2 c}$
$\dfrac{\sigma}{\delta}$	$\dfrac{ac}{b^2}+\dfrac{\gamma a}{\rho b}$	$1+\dfrac{\gamma a}{\rho b}$	$\dfrac{c}{b}\dfrac{\gamma a-\rho b}{\gamma b-\rho c}+\dfrac{\gamma a}{\rho b}$	$\dfrac{c}{b}\dfrac{\gamma a+\rho b}{\gamma b+\rho c}+\dfrac{\gamma a}{\rho b}$	$2\dfrac{\gamma a}{\rho b}$

4.2 Součinový tvar metod s proměnnou metrikou

Zatím jsme předpokládali, že aktualizovaná matice H je vždy pozitivně definitní. V některých důležitých případech je však tato matice pozitivně semidefinitní a má hodnost $m < n$. Pak je výhodné předpokládat, že $H = SS^T$ a $H_+ = S_+S_+^T$, kde matice $S \in R^{n\times m}$ a $S_+ \in R^{n\times m}$ mají plnou hodnost, a matice S_+ se určuje pomocí aktualizace

$$\frac{1}{\sqrt{\gamma}}S_+ = S + p\tilde{q}^T, \tag{321}$$

kde $p \in R^n$ a $\tilde{q} \in R^m$. Tento součinový tvar metod s proměnnou metrikou se používá zejména při minimalizaci na lineární varietě rovnoběžné s podprostorem $\mathcal{L}(S)$ dimenze $m < n$ (oddíl 19.1), nebo při realizaci posunutých metod s omezenou pamětí, kdy $H = \zeta I + SS^T$ a $m \ll n$ (oddíl 9.6). Používáme-li k minimalizaci na lineární varietě metody spádových směrů, platí $s = -Hg = -SS^Tg$, takže $s \in \mathcal{L}(S)$ (v případě posunutých metod s omezenou pamětí však tento předpoklad neplatí). Pokud $s \in \mathcal{L}(S)$, budeme předpokládat, že

$$d = S\tilde{d} \neq 0, \qquad p = S\tilde{p} \neq 0$$

($d \neq 0$ plyne z toho, že $Hg = 0$ pouze tehdy, je-li bod x stacionárním bodem na lineární varietě rovnoběžné s podprostorem $\mathcal{L}(S)$ a $p \neq 0$ je záležitost volby tohoto vektoru). Dále budeme předpokládat, že

$$\tilde{y} = S^Ty \neq 0, \qquad \tilde{q} = S^Tq \neq 0$$

($\tilde{y} \neq 0$ plyne z toho, že nerovnost $b = y^Td = y^TS\tilde{d} = \tilde{y}^T\tilde{d} > 0$ lze zajistit vhodným výběrem délky kroku a $\tilde{q} \neq 0$ je záležitost volby tohoto vektoru).

Poznámka 125. Pokud $H = SS^T$, počítáme směrový vektor $s = -Hg$ podle vzorců

$$s = S\tilde{s}, \qquad \tilde{s} = -S^Tg.$$

Pak $\tilde{d} = \alpha\tilde{s}$ a $\tilde{y} = S^Ty$.

Poznámka 126. Pokud $p \in \mathcal{L}(S)$ (jako u metod pro minimalizaci na lineární varietě), lze vzorec (321) zapsat dvojím způsobem, buď

$$\frac{1}{\sqrt{\gamma}}S_+ = S(I + \tilde{p}\tilde{q}^T), \tag{322}$$

nebo

$$\frac{1}{\sqrt{\gamma}}S_+ = (I + pq^T)S. \tag{323}$$

Pokud ale $p \notin \mathcal{L}(S)$ (jako u posunutých metod s proměnnou metrikou), jsou tyto matice různé.

V případě, že $m < n$, je pozitivně semidefinitní matice H singulární (má hodnost $m < n$). Z tohoto důvodu nelze použít inverzní matici H^{-1}. Místo toho se používá pseudoinverzní matice $H^\dagger$

Definice 37. *Nechť M je libovolná matice. Pak matici $M^\dagger$ stejného typu jako M^T nazveme pseudoinverzí matice M, jsou-li matice $MM^\dagger$ a $M^\dagger M$ symetrické a platí-li*

$$MM^\dagger M = M, \qquad M^\dagger MM^\dagger = M^\dagger.$$

Věta 81. *Ke každé matici existuje její pseudoinverze a je určena jednoznačně.*

Důkaz (a) (Existence) Nechť matice $M \in R^{n\times m}$ má hodnost $k \leq \min(n, m)$. Je zřemé, že tuto matici lze vyjádřit ve tvaru $M = UV^T$, kde matice $U \in R^{n\times k}$ a $V \in R^{m\times k}$ mají plnou hodnost (hodnost součinu matic nepřevýší hodnost žádného činitele). Ukážeme, že $M^\dagger = V(V^TV)^{-1}(U^TU)^{-1}U^T$. Symetrie matic $M^\dagger M$ a $MM^\dagger$ je zřejmá. Dále platí

$$MM^\dagger M = UV^TV(V^TV)^{-1}(U^TU)^{-1}U^TUV^T = UV^T = M,$$

$$M^\dagger M M^\dagger = V(V^TV)^{-1}(U^TU)^{-1}U^TUV^TV(V^TV)^{-1}(U^TU)^{-1}U^T = V(V^TV)^{-1}(U^TU)^{-1}U^T = M^\dagger.$$

(b) (Jednoznačnost) Nechť $M_1^\dagger$, $M_2^\dagger$ jsou dvě matice takové, že

$$\begin{aligned} MM_1^\dagger &= (MM_1^\dagger)^T, & MM_2^\dagger &= (MM_2^\dagger)^T, \\ M_1^\dagger M &= (M_1^\dagger M)^T, & M_2^\dagger M &= (M_2^\dagger M)^T, \\ MM_1^\dagger M &= M, & MM_2^\dagger M &= M, \\ M_1^\dagger MM_1^\dagger &= M_1^\dagger, & M_2^\dagger MM_2^\dagger &= M_2^\dagger. \end{aligned}$$

Nejdříve ukážeme, že $MM_1^\dagger = MM_2^\dagger$. Platí

$$MM_1^\dagger = (M_1^\dagger)^TM^T = (M_1^\dagger)^TM^T(M_2^\dagger)^TM^T = MM_1^\dagger(M_2^\dagger)^TM^T = MM_1^\dagger MM_2^\dagger = MM_2^\dagger$$

Úplně stejně se dokáže, že $M_1^\dagger M = M_2^\dagger M$. Použijeme-li tyto vztahy, dostaneme

$$M_1^\dagger = M_1^\dagger MM_1^\dagger = M_1^\dagger MM_2^\dagger = M_2^\dagger MM_2^\dagger = M_2^\dagger$$

. □

Poznámka 127. Má-li matice $S \in R^{n\times m}$, $m \le n$, plnou hodnost, lze podle definice 37 snadno ověřit, že

$$S^\dagger = (S^TS)^{-1}S^T, \qquad S^\dagger S = I \tag{324}$$

$$(SS^T)^\dagger = (S^\dagger)^TS^\dagger = S(S^TS)^{-2}S^T, \qquad (S^TS)^{-1} = S^\dagger(S^\dagger)^T. \tag{325}$$

Z (324) plyne, že má-li matice S plnou hodnost, má i matice $S^\dagger$ plnou hodnost.

Poznámka 128. Nechť M je symetrická pozitivně semidefinitní matice. Pak existují matice $M^{1/2}$ a $(M^\dagger)^{1/2}$ takové, že $M^{1/2}M^{1/2} = M$ a $(M^\dagger)^{1/2}(M^\dagger)^{1/2} = M^\dagger$. Má-li matice $S \in R^{n\times m}$, $m \le n$, plnou hodnost, lze podle definice 37 snadno ověřit, že

$$(SS^T)^{1/2} = S(S^TS)^{-1/2}S^T, \qquad ((SS^T)^\dagger)^{1/2} = S(S^TS)^{-3/2}S^T. \tag{326}$$

Věta 82. *Nechť matice $S \in R^{n\times m}$, $m \le n$, má plnou hodnost a nechť $U \in R^{n\times k}$, $V \in R^{m\times k}$, $k \le m$, jsou matice takové, že $S + UV^T$ má plnou hodnost. Pak platí*

$$(S + UV^T)^\dagger = S^\dagger - S^\dagger U(I + V^TS^\dagger U)^{-1}V^TS^\dagger. \tag{327}$$

Důkaz (a) Nejprve ukážeme, že matice $I + V^TS^\dagger U$ je regulární. Předpokládejme naopak, že pro nějaký vektor $x \neq 0$ platí $(I + V^TS^\dagger U)x = 0$. Pak nutně $y = S^\dagger Ux \neq 0$. Musí tedy platit

$$S^\dagger U(I + V^TS^\dagger U)x = (S^\dagger SS^\dagger U + S^\dagger UV^TS^\dagger U)x = S^\dagger(S + UV^T)y = 0,$$

kde $y \neq 0$, což je ve sporu s předpokladem, že matice S (a tudíž i $S^\dagger$) a $S + UV^T$ mají plnou hodnost.

(b) Jelikož $S + UV^T$ má plnou hodnost, má i $(S + UV^T)^\dagger$ plnou hodnost a podle (324) je tato matice určena vztahem $(S + UV^T)^\dagger(S + UV^T) = I$. Protože

$$\begin{aligned} (S^\dagger - S^\dagger U(I + V^TS^\dagger U)^{-1}V^TS^\dagger)(S + UV^T) &= I + S^\dagger U(I + V^TS^\dagger U)(I + V^TS^\dagger U)^{-1}V^T \\ &\quad - S^\dagger U(I + V^TS^\dagger U)^{-1}V^T - S^\dagger UV^TS^\dagger U(I + V^TS^\dagger U)^{-1}V^T = I, \end{aligned}$$

platí $(S + UV^T)^\dagger = S^\dagger - S^\dagger U(I + V^TS^\dagger U)^{-1}V^TS^\dagger$. □

Věta 83. *Nechť H je pozitivně semidefinitní matice a $U = HV$, kde $V \in R^{n\times 2}$. Pak, jsou-li matice M a $M^{-1} + V^T HV$ regulární, platí*

$$(H + UMU^T)^\dagger = B - BU(M^{-1} + U^T BU)^{-1}U^T B, \qquad B = H^\dagger. \tag{328}$$

Důkaz Jelikož $U = HV$, můžeme vzorec (328) zapsat ve tvaru

$$(H + HVMV^T H)^\dagger = B - BHV(M^{-1} + V^T HV)^{-1}V^T HB.$$

Platí

$$\begin{aligned}\big(B - BHV(M^{-1} + V^T HV)^{-1}V^T HB\big)\big(H + HVMV^T H\big)& \\ = BH - BHV(M^{-1} + V^T HV)^{-1}V^T HBH + BHVMV^T H& \\ - BHV(M^{-1} + V^T HV)^{-1}V^T HVMV^T H = BH,&\end{aligned}$$

takže matice $(H + UMU^T)^\dagger(H + UMU^T) = BH$ je symetrická. Úplně stejným způsobem se dokáže, že matice $(H + UMU^T)(H + UMU^T)^\dagger = HB$ je symetrická. Nakonec dostaneme

$$\begin{aligned}(H + UMU^T)(H + UMU^T)^\dagger(H + UMU^T) &= (H + HVMV^T H)BH \\ &= H + HVMV^T H = H + UMU^T, \\ (H + UMU^T)^\dagger(H + UMU^T)(H + UMU^T)^\dagger &= BH(B - BHV(M^{-1} + V^T HV)^{-1}V^T HB \\ &= B - BH(M^{-1} + V^T HV)^{-1}V^T HB \\ &= (H + UMU^T)^\dagger.\end{aligned}$$

□

Poznámka 129. Podmínka $U = [d, Hy] = HV$ je splněna, pokud $s = -Hg$. Pak $d = -\alpha Hg$, takže $V = [-\alpha g, y]$.

Poznámka 130. Použijeme-li pseudoinverzní matice $S^\dagger$ a $B = H^\dagger$, můžeme psát

$$\tilde{y} = S^T y = S^\dagger Hy, \qquad \tilde{q} = S^T q = S^\dagger Hq, \tag{329}$$
$$\tilde{d} = S^\dagger d = S^T Bd, \qquad \tilde{p} = S^\dagger p = S^T Bp, \tag{330}$$
$$Hy = S\tilde{y}, \qquad Hq = S\tilde{q}, \tag{331}$$
$$Bd = (S^\dagger)^T\tilde{d}, \qquad Bp = (S^\dagger)^T\tilde{p}, \tag{332}$$

neboť podle (324) a (325) platí $S^\dagger H = S^\dagger SS^T = S^T$ a $S^T B = S^T(S^\dagger)^T S^\dagger = S^\dagger$. Při odvozování součinového tvaru metod s proměnnou metrikou budeme používat čísla

$$a = y^T SS^T y = \tilde{y}^T\tilde{y}, \qquad b = y^T d = \tilde{y}^T\tilde{d}, \qquad c = d^T(SS^T)^\dagger d = \tilde{d}^T\tilde{d}. \tag{333}$$

Nejprve je třeba zobecnit lemma 32.

Lemma 33. *Nechť $H = SS^T$, kde matice $S \in R^{n\times m}$, $m \le n$, má plnou hodnost. Nechť $(1/\gamma)H_+ = H + UMU$, kde $U = S\tilde{U} = S[\tilde{d}, \tilde{y}]$ a $M \in R^{2\times 2}$. Pak matice $(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2}$ má $n - m$ nulových vlastních čísel odpovídajících vlastním vektorům kolmým ke sloupcům matice S, $m-2$ jednotkových vlastních čísel a dvě vlastní čísla, která jsou vlastními čísly matice $I+MU^T H^\dagger U$. Tato dvě vlastní čísla jsou řešením kvadratické rovnice.*

$$\lambda^2 - \sigma\lambda + \delta = 0,$$

kde

$$\sigma = \frac{1}{b^2}(\eta(ac - b^2) + b^2) + \frac{\rho}{\gamma}\frac{c}{b}, \qquad \delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac - b^2) + b^2).$$

Důkaz Použijeme-li vztahy uvedené v poznámkách 127 a 128, můžeme psát

$$\begin{aligned}(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2} &= (H^\dagger)^{1/2}H(H^\dagger)^{1/2} + (H^\dagger)^{1/2}UMU^T(H^\dagger)^{1/2} \\ &= S(S^TS)^{-3/2}S^TSS^TS(S^TS)^{-3/2}S^T + (H^\dagger)^{1/2}UMU^T(H^\dagger)^{1/2} \\ &= S(S^TS)^{-1}S^T + S(S^TS)^{-3/2}S^TUMU^TS(S^TS)^{-3/2}S^T\end{aligned}$$

Je zřejmé, že $(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2}v = 0$, pro každý vektor v takový, že $S^Tv = 0$. Takových lineárně nezávislých vektorů je $n-m$. Zbylá vlastní čísla matice $(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2}$ tedy odpovídají vlastním vektorům tvaru $v = S\tilde{v}$. Pro tyto vektory platí

$$S(S^TS)^{-1}S^Tv = S(S^TS)^{-1}S^TS\tilde{v} = S\tilde{v} = v,$$

takže odpovídající vlastní čísla jsou o jedničku větší než vlastní čísla matice $(H^\dagger)^{1/2}UMU^T(H^\dagger)^{1/2}$. Podle důsledku 9 jsou to tedy jedničky nebo vlastní čísla matice $I + MU^TH^\dagger U$. Jelikož

$$U^TH^\dagger U = \tilde{U}^TS^TS(S^TS)^{-2}S^TS\tilde{U} = \tilde{U}^T\tilde{U} = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix},$$

můžeme postupovat stejně jako v důkazu lemmatu 32 a získat tak vztahy pro dvě zbylá vlastní čísla matice $(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2}$. □

Ukážeme, jak musí vypadat aktualizace (321), aby byla splněna kvazinewtonovská podmínka

$$S_+S_+^Ty = \rho d. \tag{334}$$

Lemma 34. *Uvažujme aktualizaci (321), s nenulovým vektorem $\tilde{q}$ zvoleným tak, že*

$$D^2 \triangleq \left(\tilde{q}^T\tilde{y}\right)^2 + \left(\frac{\rho}{\gamma}b - a\right)\tilde{q}^T\tilde{q} > 0, \tag{335}$$

kde $\tilde{y} = S^Ty$ (pokud $(\rho/\gamma)b > a$, lze volit $\tilde{q}$ libovolně a pokud $(\rho/\gamma)b = a$, stačí aby platilo $\tilde{q}^T\tilde{y} \neq 0$). Pak kvazinewtonovská podmínka (334) je splněna právě tehdy, když

$$p = \frac{(\rho/\gamma)d - S(\tilde{y} + \tau\tilde{q})}{\tilde{q}^T(\tilde{y} + \tau\tilde{q})} = \frac{(\rho/\gamma)d - S(\tilde{y} + \tau\tilde{q})}{D}.$$

Číslo $\tau = p^Ty$ se vypočte z rovnosti $\tilde{q}^T\tilde{y} + \tau\tilde{q}^T\tilde{q} = D$ (pokud $(\rho/\gamma)b = a$ a $\tilde{q}^T\tilde{y} \neq 0$, lze volit $\tau = 0$).

Důkaz Použitím vztahu (321) dostaneme

$$\frac{1}{\gamma}S_+S_+^T = (S + p\tilde{q}^T)(S^T + \tilde{q}p^T) = SS^T + p\tilde{q}^TS^T + S\tilde{q}p^T + p\tilde{q}^T\tilde{q}p^T,$$

takže kvazinewtonovskou podmínku můžeme zapsat ve tvaru

$$S\tilde{y} + p\tilde{q}^T\tilde{y} + S\tilde{q}p^Ty + p\tilde{q}^T\tilde{q}p^Ty = \frac{\rho}{\gamma}d,$$

kde $\tilde{y} = S^Ty$. Označíme-li $\tau = p^Ty$, můžeme tuto rovnost zapsat ve tvaru

$$S(\tilde{y} + \tau\tilde{q}) + p\tilde{q}^T(\tilde{y} + \tau\tilde{q}) = \frac{\rho}{\gamma}d,$$

odkud dostaneme vztah pro p. Dosadíme-li tento vztah do rovnosti $\tau = p^Ty$, můžeme psát

$$\tau^2\tilde{q}^T\tilde{q} + 2\tau\tilde{q}^T\tilde{y} = \frac{\rho}{\gamma}b - a.$$

Z druhé strany umocněním výrazu $D = \tilde{q}^T\tilde{y} + \tau\tilde{q}^T\tilde{q}$, dostaneme

$$\tau^2(\tilde{q}^T\tilde{q})^2 + 2\tau\tilde{q}^T\tilde{y}\tilde{q}^T\tilde{q} + (\tilde{q}^T\tilde{y})^2 = D^2,$$

což porovnáním dává

$$D^2 = (\tilde{q}^T\tilde{y})^2 + \left(\frac{\rho}{\gamma}b - a\right)\tilde{q}^T\tilde{q}.$$

Tento výraz musí být kladný, což poněkud omezuje volbu vektoru $\tilde{q}$. Poznamenejme, že pro $\tilde{q} = \tilde{y}$ a $(\rho/\gamma)b > 0$ je tento výraz kladný, neboť v tomto případě platí $\tilde{q}^T\tilde{y} = \tilde{q}^T\tilde{q}$ a $a = \tilde{y}^T\tilde{y} = \tilde{q}^T\tilde{q}$, takže $D^2 = (\rho/\gamma)b\,\tilde{q}^T\tilde{q} > 0$. □

Nyní se budeme zabývat součinovým tvarem metod z Broydenovy třídy, kdy $U = [d, Hy]$. Jelikož předpokládáme, že $s = S\tilde{s}$, a podle (331) platí $Hy = S\tilde{y}$, můžeme psát

$$U = S\tilde{U} \quad \Rightarrow \quad S^\dagger U = S^\dagger S\tilde{U} = \tilde{U}, \tag{336}$$

kde $\tilde{U} = [\tilde{d}, \tilde{y}]$ (používáme vzorec (324)). Abychom dostali rekurentní vztah $S_+S_+^T = \gamma(SS^T + UMU^T)$, je třeba aby vektor $\tilde{q}$ v (321) byl lineární kombinací sloupců matice $\tilde{U}$. Je zajímavé, že v součinovém tvaru lze realizovat pouze metody s proměnnou metrikou, pro které platí $\delta \geq 0$ a $\mu \geq 0$.

Lemma 35. *Nechť $U = S\tilde{U} = S[\tilde{d}, \tilde{y}]$. Uvažujme aktualizaci (321) (splňující kvazinewtonovskou podmínku (334)), kde $\tilde{q} = \tilde{U}\hat{q}$ a kde vektor $\hat{q} \in R^2$ je zvolen tak, aby byla splněna nerovvnost (335). Pak existuje symetrická matice $M \in R^{2\times 2}$ taková, že pro matici $H_+ = S_+S_+^T$ platí $H_+ = \gamma(H + UMU^T)$, přičemž $\delta = \det(I + MU^TH^\dagger U) \geq 0$ a $\mu = -\det M \geq 0$.*

Důkaz Jestliže $\tilde{q} = \tilde{U}\hat{q}$ a $D^2 > 0$, pak podle lemmatu 34 platí

$$p = \frac{(\rho/\gamma)d - S(\tilde{y} + \tau\tilde{q})}{D} = \frac{(\rho/\gamma)d - SS^Ty + \tau U\hat{q}}{D} \triangleq U\hat{p},$$

kde $\hat{p} \in R^2$ Dosadíme-li tato vyjádření do vztahu (321), můžeme psát

$$\frac{1}{\gamma}S_+S_+^T = (S + p\tilde{q}^T)(S + p\tilde{q}^T)^T = SS^T + U\hat{p}\hat{q}^TU^T + U\hat{q}\hat{p}^TU^T + U\hat{p}\tilde{q}^T\tilde{q}\hat{p}^TU^T = SS^T + UMU^T,$$

kde

$$M = \hat{p}\hat{q}^T + \hat{q}\hat{p}^T + \hat{p}\tilde{q}^T\tilde{q}\hat{p}^T = [\hat{p},\, \hat{q}]\begin{bmatrix} \tilde{q}^T\tilde{q}, & 1 \\ 1, & 0 \end{bmatrix}\begin{bmatrix} \hat{p}^T \\ \hat{q}^T \end{bmatrix}.$$

Použijeme-li větu o násobení determinantů, dostaneme

$$\det M = -(\det[\hat{p},\, \hat{q}])^2 \leq 0.$$

Podle lemmatu 33 se číslo δ rovná součinu vlastních čísel matice

$$(1/\gamma)(H^\dagger)^{1/2}H_+(H^\dagger)^{1/2} = S(S^TS)^{-3/2}S^TS(I + \tilde{p}\tilde{q}^T)(I + \tilde{q}\tilde{p}^T)S^TS(S^TS)^{-3/2}S^T$$

odpovídajících vlastním vektorům z $\mathcal{L}(S)$ (používáme vztahy (322) a (326)). Tento součin je podle lemmatu 31 (c) roven determinantu matice

$$(I + \tilde{q}\tilde{p}^T)(S^TS)^{-1/2}S^TS(S^TS)^{-1/2}(I + \tilde{p}\tilde{q}^T) = (I + \tilde{q}\tilde{p}^T)(I + \tilde{p}\tilde{q}^T),$$

takže platí

$$\delta = \det(I + \tilde{p}\tilde{q}^T)\det(I + \tilde{q}\tilde{p}^T) = (1 + \tilde{q}^T\tilde{p})^2 \geq 0. \tag{337}$$

□

Poznámka 131. Podle lemmatu 35 existuje součinový tvar pouze pro ty metody z Broydenovy třídy, pro které $\delta \geq 0$ a $\mu \geq 0$. Ve větě 85 a důsledku 13 ukážeme, že tyto nutné podmínky jsou i podmínkami postačujícími.

Nyní ukážeme, jak lze volit vektor $\tilde{q} = \tilde{U}\hat{q}$, abychom dostali jednotlivé metody z Broydenovy třídy. Jak vyplývá z důkazu lemmatu 34, je vektor p určen vektorem $\tilde{q}$ (existují obvykle dvě řešení). Navíc výsledná aktualizace nezávisí na normě vektoru $\tilde{q}$, neboť z (321) plyne, že vynásobíme-li vektor $\tilde{q}$ nějakým číslem, stačí tímto číslem vydělit vektor p. Proto budeme hledat vektor $\tilde{q}$ ve tvaru $\tilde{q} = \vartheta\tilde{y} - \tilde{d}$ (takže lze volit $\tilde{q} = \tilde{y}$, pokud $\vartheta = \infty$).

Věta 84. *Nechť jsou splněny předpoklady lemmatu 35 a nechť $\tilde{q} = \vartheta\tilde{y} - \tilde{d}$. Pak aktualizace (321) je ekvivalentní aktualizaci (286), pokud*

$$\frac{\rho}{\gamma}\frac{(\vartheta b - c)^2}{D^2} = \frac{1}{ab}(\eta(ac - b^2) + b^2), \tag{338}$$

kde

$$D^2 = \frac{\rho}{\gamma}b(\vartheta^2 a - 2\vartheta b + c) - (ac - b^2). \tag{339}$$

Jestliže $\eta = 0$, pak buď $\vartheta = \infty$, takže $\tilde{q} = \tilde{y}$, nebo $\vartheta = (\gamma/\rho + c/b)/2$. V ostatních případech platí

$$\vartheta = -\frac{m_3}{m_2} \pm \sqrt{\left(\frac{\gamma}{\rho} + \frac{m_3}{m_2}\right)\left(\frac{c}{b} + \frac{m_3}{m_2}\right)},$$

kde m_2, m_3 jsou čísla určená vztahy (283), což lze zapsat ve tvaru

$$\vartheta = \frac{b}{\eta}\left(\frac{\eta - 1}{a} \pm \frac{\gamma}{\rho}\sqrt{\delta\mu}\right).$$

Důkaz Podle lemmatu 34 platí

$$1 + \tilde{q}^T\tilde{p} = 1 + q^T p = 1 + \frac{(\rho/\gamma)q^T d - \tilde{q}^T(\tilde{y} - \tau\tilde{q})}{\tilde{q}^T(\tilde{y} - \tau\tilde{q})} = \frac{\rho}{\gamma}\frac{q^T d}{D} = \frac{\rho}{\gamma}\frac{\tilde{q}^T\tilde{d}}{D} = \frac{\rho}{\gamma}\frac{\vartheta b - c}{D}.$$

Použijeme-li tuto rovnost spolu se vztahy (300) a (337), dostaneme

$$\frac{\rho^2}{\gamma^2}\frac{(\vartheta b - c)^2}{D^2} = (1 + \tilde{q}^T\tilde{p})^2 = \delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac - b^2) + b^2).$$

což po úpravě dává (338). Jelikož $\tilde{q}^T\tilde{y} = \vartheta a - b$ a $\tilde{q}^T\tilde{q} = \vartheta^2 a - 2\vartheta b + c$, dostaneme po dosazení těchto vztahů do (335) a po úpravě výraz (339). Vynásobíme-li rovnost (338) číslem aD^2 můžeme psát

$$\frac{\rho}{\gamma}a(\vartheta b - c)^2 - bD^2 = \frac{\eta}{b}D^2(ac - b^2).$$

Dosadíme-li (339) do levé strany této rovnosti, sdružíme-li odpovídající si členy a vydělíme-li vzniklou rovnici číslem $ac - b^2$, dostaneme

$$\left(\frac{\rho}{\gamma}c + b\right) - 2\vartheta\frac{\rho}{\gamma}b = \frac{\eta}{b}D^2. \tag{340}$$

Pokud $\eta = 0$, má tato rovnice řešení $\vartheta = (\gamma/\rho + c/b)/2$. Rovnice (338) má v tomto případě další řešení $\vartheta = \infty$, neboli $\tilde{q} = \tilde{y}$, o čemž se můžeme přesvědčit, položíme-li $\eta = 0$ a ponecháme-li v (338)–(339) pouze členy obsahující ϑ^2. Pokud $\eta \neq 0$, dosazením (339) do (340) a dalšími úpravami dostaneme

$$b\left[\left(\frac{\rho c}{\gamma b} + 1\right) - 2\vartheta\frac{\rho}{\gamma}\right] = \eta\, b\left[\left(\frac{\rho c}{\gamma b} + 1\right) - 2\vartheta\frac{\rho}{\gamma}\right] - \eta\, a\left(\vartheta^2\frac{\rho}{\gamma} - \frac{c}{b}\right).$$

Převedeme-li všechny členy na pravou stranu, vydělíme-li vzniklou rovnici číslem $\eta a\rho/\gamma$ a použijeme-li vztah $(\eta-1)b/(\eta a) = -m_3/m_2$, který plyne z (283), můžeme psát

$$\frac{(\eta-1)b}{\eta a}\left[\left(\frac{c}{b}+\frac{\gamma}{\rho}\right)-2\vartheta\right]+\vartheta^2-\frac{\gamma c}{\rho b}=\vartheta^2+2\vartheta\frac{m_3}{m_2}-\left(\frac{m_3}{m_2}\left(\frac{\gamma}{\rho}+\frac{c}{b}\right)+\frac{\gamma c}{\rho b}\right)=0.$$

Tato kvadratická rovnice má řešení

$$\vartheta=-\frac{m_3}{m_2}\pm\sqrt{\left(\frac{m_3}{m_2}\right)^2+\frac{m_3}{m_2}\left(\frac{\gamma}{\rho}+\frac{c}{b}\right)+\frac{\gamma c}{\rho b}}=-\frac{m_3}{m_2}\pm\sqrt{\left(\frac{\gamma}{\rho}+\frac{m_3}{m_2}\right)\left(\frac{c}{b}+\frac{m_3}{m_2}\right)}.$$

Poslední dokazovaný vztah plyne z toho, že

$$\frac{c}{b}+\frac{m_3}{m_2}=\frac{\eta ac+(1-\eta)b^2}{\eta ab}=\frac{\eta(ac-b^2)+b^2}{\eta ab}=\frac{\gamma\delta}{\rho\eta}$$

a

$$\frac{\gamma}{\rho}+\frac{m_3}{m_2}=\frac{1}{\eta}\left(\eta\frac{\gamma}{\rho}+(1-\eta)\frac{b}{a}\right)=\frac{\gamma b}{\rho a\eta}\left(\eta\frac{a}{b}+(1-\eta)\frac{\rho}{\gamma}\right)=\frac{\gamma b^2}{\rho\eta}\mu.$$

□

Poznámka 132. Věta 84 udává způsob, jak lze k dané metodě s proměnnou metrikou (charakterizované parametrem η) nalézt součinový tvar (321). K dané hodnotě η najdeme podle věty 84 hodnotu ϑ určující vektor $\hat{q}$ a číslo D^2 (existují obvykle dvě řešení). Pak podle lemmatu 34 určíme vektor p (existují opět dvě řešení).

(a) Pro metodu DFP platí $\eta = 0$, takže lze volit $\vartheta = \infty$, neboli $\tilde{q} = \tilde{y}$.
(b) Pro metodu BFGS platí $\eta = 1$, takže $m_3 = 0$, což dává $\vartheta = \pm\sqrt{\gamma c/(\rho b)}$.
(c) Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$, takže $\mu = 0$ a $m_3/m_2 = -\gamma/\rho$, což dává $\vartheta = \gamma/\rho$. Metodu hodnosti 1 můžeme vyjádřit v součinovém tvaru pouze tehdy, když buď $0 < \rho/\gamma \le b/c$, nebo $a/b \le \rho/\gamma$ (poznámka 117).

Použití věty 84 není příliš vhodné pro explicitní vyjádření součinového tvaru. Jinou možnost udává následující věta, uvedená v práci [36], kde symboly $\sqrt{\delta}$ a $\sqrt{\mu}$ označují libovolné hodnoty (kladné i záporné) takové, že $(\sqrt{\delta})^2 = \delta$ a $(\sqrt{\mu})^2 = \mu$.

Věta 85. *Nechť $a > 0$, $b > 0$, $c > 0$ a $ac - b^2 > 0$. Uvažujme aktualizaci (286), kde $H = SS^T$, $\rho > 0$, $\gamma > 0$. Nechť $\delta \ge 0$, $\mu \ge 0$ a buď $\delta > 0$ nebo $\mu > 0$. Nechť $d = S\tilde{d}$ a $\tilde{y} = S^T y$. Pak platí $H_+ = S_+^T S_+$, kde*

$$\frac{1}{\sqrt{\gamma}}S_+ = S + U\hat{p}\hat{q}^T\tilde{U}^T, \tag{341}$$

přičemž

$$\hat{p}\hat{q}^T = \frac{1}{\lambda}\begin{bmatrix}\frac{\rho}{\gamma}+a\sqrt{\mu}\\ -1-b\sqrt{\mu}\end{bmatrix}\begin{bmatrix}\sqrt{\delta}-b\sqrt{\mu}\\ -\frac{\gamma}{\rho}\sqrt{\delta}+c\sqrt{\mu}\end{bmatrix}^T \tag{342}$$

a

$$\lambda = \left(\frac{\rho}{\gamma}c-b\right)+\left(b-\frac{\gamma}{\rho}a\right)\sqrt{\delta}+\left(ac-b^2\right)\sqrt{\mu}. \tag{343}$$

Důkaz (a) Z důkazu lemmatu 35 víme, že

$$(\hat{p}_1\hat{q}_2 - \hat{q}_1\hat{p}_2)^2 = (\det[\hat{p},\, \hat{q}])^2 = -\det M = \mu.$$

Použijeme-li tento výsledek můžeme psát

$$\hat{p}\hat{q}^T - \hat{q}\hat{p}^T = \begin{bmatrix} 0, & \hat{p}_1\hat{q}_2 - \hat{q}_1\hat{p}_2 \\ \hat{q}_1\hat{p}_2 - \hat{p}_1\hat{q}_2, & 0 \end{bmatrix} = \begin{bmatrix} 0, & +\sqrt{\mu} \\ -\sqrt{\mu}, & 0 \end{bmatrix}.$$

(b) Předpokládejme nejprve, že $\delta > 0$. Použijeme-li vztah (341), dostaneme

$$\frac{1}{\gamma} S_+ S_+^T = S(I + \tilde{U}\hat{p}\hat{q}^T\tilde{U}^T)(I + \tilde{U}\hat{q}\hat{p}^T\tilde{U}^T)S^T.$$

Z důkazu lemmatu 35 víme, že

$$\det(I + \tilde{U}\hat{p}\hat{q}^T\tilde{U}^T) = \sqrt{\delta},$$

takže podle lemmatu 31 (e) platí

$$(I + \tilde{U}\hat{p}\hat{q}^T\tilde{U}^T)^{-1} = I - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T\tilde{U}^T$$

a podmínku $S_+ S_+^T y = \rho d$ můžeme zapsat ve tvaru

$$(I + \tilde{U}\hat{q}\hat{p}^T\tilde{U}^T)\tilde{y} = \frac{\rho}{\gamma}\left(I - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T\tilde{U}^T\right)\tilde{d}.$$

Vynásobíme-li tuto rovnici zleva maticí $\tilde{U}^T$ a přihlédneme-li k tomu, že

$$\tilde{U}^T\tilde{U} = \begin{bmatrix} \tilde{d}^T\tilde{d}, & \tilde{d}^T\tilde{y} \\ \tilde{y}^T\tilde{d}, & \tilde{y}^T\tilde{y} \end{bmatrix} = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix},$$

dostaneme

$$\begin{bmatrix} b \\ a \end{bmatrix} + \begin{bmatrix} c, & b \\ b, & a \end{bmatrix}\hat{q}\hat{p}^T\begin{bmatrix} b \\ a \end{bmatrix} = \frac{\rho}{\gamma}\left(\begin{bmatrix} c \\ b \end{bmatrix} - \frac{1}{\sqrt{\delta}}\begin{bmatrix} c, & b \\ b, & a \end{bmatrix}\hat{p}\hat{q}^T\begin{bmatrix} c \\ b \end{bmatrix}\right),$$

což po úpravě dává

$$\hat{q}\hat{p}^T\begin{bmatrix} b \\ a \end{bmatrix} + \hat{p}\hat{q}^T\frac{1}{\sqrt{\delta}}\frac{\rho}{\gamma}\begin{bmatrix} c \\ b \end{bmatrix} = \frac{1}{ac - b^2}\begin{bmatrix} a, & -b \\ -b, & c \end{bmatrix}\left(\frac{\rho}{\gamma}\begin{bmatrix} c \\ b \end{bmatrix} - \begin{bmatrix} b \\ a \end{bmatrix}\right) = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}.$$

Použijeme-li nyní (a), dostaneme

$$\hat{p}\hat{q}^T\left(\begin{bmatrix} b \\ a \end{bmatrix} + \frac{1}{\sqrt{\delta}}\frac{\rho}{\gamma}\begin{bmatrix} c \\ b \end{bmatrix}\right) = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{\mu} \\ -1 - b\sqrt{\mu} \end{bmatrix}.$$

Z tohoto vyjádření je patrné, že vektor $\hat{p} \in R^2$ je skalárním násobkem vektoru na pravé straně poslední rovnosti. Jelikož skalární násobek můžeme zvolit libovolně, položíme

$$\hat{p} = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{\mu} \\ -1 - b\sqrt{\mu} \end{bmatrix}.$$

Pak pro vektor $\hat{q} \in R^2$ dostaneme rovnici

$$\hat{q}_1\left(b + \frac{1}{\sqrt{\delta}}\frac{\rho}{\gamma}c\right) + \hat{q}_2\left(a + \frac{1}{\sqrt{\delta}}\frac{\rho}{\gamma}b\right) = 1$$

a z (a) plyne

$$\hat{q}_1\,(1+b\sqrt{\mu})+\hat{q}_2\left(\frac{\rho}{\gamma}+a\sqrt{\mu}\right)=\sqrt{\mu}.$$

Řešením těchto dvou rovnic je vektor

$$\hat{q}=\frac{1}{\lambda}\begin{bmatrix}\sqrt{\delta}-b\sqrt{\mu}\\ -\dfrac{\gamma}{\rho}\sqrt{\delta}+c\sqrt{\mu}\end{bmatrix},$$

kde

$$\lambda=\left(\frac{\rho}{\gamma}c-b\right)+\left(b-\frac{\gamma}{\rho}a\right)\sqrt{\delta}+(ac-b^2)\sqrt{\mu}.$$

Jelikož $ac-b^2>0$, je alespoň jeden z výrazů $(\rho/\gamma)c-b$ a $b-(\gamma/\rho)a$ nenulový, a protože $\delta>0$, lze vhodnou volbou znamének $\sqrt{\delta}$ a $\sqrt{\mu}$ docílit toho, že $\lambda\neq 0$ (vzorec (342) má tedy smysl).

(c) Nechť $\delta=0$ (takže $\eta=\eta^*$) a $\mu>0$. Jelikož podle poznámky 109 jsou $\delta(\eta)$ a $\mu(\eta)$ lineárními funkcemi parametru η, existuje číslo $\overline{\eta}>\eta^*$ takové, že $\delta(\eta)>0$ a $\mu(\eta)>0$, pokud $\eta^*<\eta\leq\overline{\eta}$). Pro tyto hodnoty parametru η lze použít postup uvedený v (b), jehož výsledkem je vzorec (342), kde $\delta(\eta)>0$ a $\mu(\eta)>0$. Provedeme-li limitní přechod $\eta\to\eta^*$ dostaneme vzorec (342), kde $\delta=0$ a $\mu>0$. Jelikož podle předpokladu platí $ac-b^2>0$, můžeme vhodnou volbou znaménka $\sqrt{\mu}$ docílit toho, že $\lambda\neq 0$, takže tento vzorec má smysl.

(d) Pokud $\delta=0$ a $\mu=0$, musí podle poznámky 109 platit $\eta(ac-b^2)+b^2=0$ a $\eta(a-(\rho/\gamma)b)+(\rho/\gamma)b=0$, což je možné jedině tehdy, když $(\rho/\gamma)c-b=0$. Pak ale $\lambda=0$, takže vzorec (342) nemá smysl (platí $\hat{q}=0$ a $\lambda=0$). □

Obě předchozí věty obsahují poměrně komplikované výrazy. Tyto výrazy se velmi zjednoduší pro základní metody (287), (288), (289). Výsledné vzorce, které nyní odvodíme, lze nalézt také v [9].

Důsledek 13. Pro metodu DFP platí $\eta=0$, čili $\delta=\rho b/(\gamma a)$ a $\mu=\rho/(\gamma ab)$, takže

$$\frac{1}{\sqrt{\gamma}}S_+^{DFP}=S-\frac{1}{a}\left(SS^Ty\pm\sqrt{\frac{\rho a}{\gamma b}}d\right)\tilde{y}^T. \tag{344}$$

Pro metodu BFGS platí $\eta=1$, čili $\delta=\rho c/(\gamma b)$ a $\mu=1/b^2$, takže

$$\frac{1}{\sqrt{\gamma}}S_+^{BFGS}=S-\frac{1}{b}d\left(\tilde{y}\pm\sqrt{\frac{\rho b}{\gamma c}}\tilde{d}\right)^T. \tag{345}$$

Pro metodu hodnosti 1 platí $\eta=(\rho/\gamma)/(\rho/\gamma-a/b)$, čili $\delta=((\rho/\gamma)c-b)/(b-(\gamma/\rho)a)$ a $\mu=0$, takže

$$\frac{1}{\sqrt{\gamma}}S_+^{R1}=S+\frac{\sqrt{\delta}-1}{(\rho/\gamma)^2c-2(\rho/\gamma)b+a}\left(\frac{\rho}{\gamma}d-SS^Ty\right)\left(\frac{\rho}{\gamma}\tilde{d}-\tilde{y}\right)^T. \tag{346}$$

V těchto vzorcích je $d=S\tilde{d}$ a $SS^Ty=S\tilde{y}$. Jmenovatel v posledním vzorci je vždy nenulový (i když $\delta=0$).

Důkaz K odvození těchto vztahů můžeme použít buď větu 84 nebo větu 85. Použití věty 84 je vhodné pro metodu DFP, neboť pro $\vartheta=0$ se potřebné výrazy velmi zjednoduší. Pro metodu BFGS musíme použít trik spočívající v tom, že kvazinewtonovská podmínka je v tomto případě splněna, pokud $\tau=-1$. Použitím této hodnoty lze obejít výpočet čísla D^2 a jeho odmocniny. Zde použijeme větu 85. Přímé dosazení do (341) není triviální a vyžaduje speciální volbu znaménka $\sqrt{\mu}$, jinak nedostaneme jednoduchá vyjádření.

(a) Pro metodu DFP lze dosazením zjistit, že $\delta = \rho b/(\gamma a)$ a $\mu = \rho/(\gamma ab)$. Znaménko $\sqrt{\mu}$ budeme volit tak, aby platilo $\sqrt{\delta} = b\sqrt{\mu}$. Pak

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma a}{\rho b}\left(\frac{\rho}{\gamma}c - b\right)\sqrt{\delta} = \left(\frac{\rho}{\gamma}c - b\right)\frac{\sqrt{\delta}+1}{\sqrt{\delta}},$$

$$\hat{p} = \begin{bmatrix} \frac{a}{b}\left(\frac{\rho b}{\gamma a} + \sqrt{\delta}\right) \\ -1 - \sqrt{\delta} \end{bmatrix} = \begin{bmatrix} \frac{a}{b}\sqrt{\delta}\left(\sqrt{\delta}+1\right) \\ -\left(\sqrt{\delta}+1\right) \end{bmatrix}, \quad \lambda\hat{q} = \frac{1}{a}\begin{bmatrix} 0 \\ \frac{\gamma a}{\rho b}\left(\frac{\rho}{\gamma}c - b\right)\sqrt{\delta} \end{bmatrix} = \frac{1}{a}\begin{bmatrix} 0 \\ \left(\frac{\rho}{\gamma}c - b\right)\frac{1}{\sqrt{\delta}} \end{bmatrix}.$$

Po vykrácení dostaneme

$$\hat{p}\hat{q}^T = \frac{1}{a}\begin{bmatrix} \frac{a}{b}\sqrt{\delta} \\ -1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}^T = -\frac{1}{a}\begin{bmatrix} \pm\sqrt{\frac{\rho a}{\gamma b}} \\ 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}^T,$$

což po dosazení do (342) dává (344).
(b) Pro metodu BFGS lze dosazením zjistit, že $\delta = \rho c/(\gamma b)$ a $\mu = 1/b^2$. Znaménko $\sqrt{\mu}$ budeme volit tak, aby platilo $\sqrt{\mu} = -1/b$. Pak

$$\begin{aligned} \lambda &= \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} - \frac{1}{b}(ac - b^2) = \frac{c}{b}\left(\frac{\rho}{\gamma}b - a\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} \\ &= \left(\frac{\rho}{\gamma}b - a\right)\frac{\gamma}{\rho}\left(\frac{\rho c}{\gamma b} + \sqrt{\delta}\right) = \left(\frac{\rho}{\gamma}b - a\right)\frac{\gamma}{\rho}\sqrt{\delta}\left(\sqrt{\delta}+1\right), \end{aligned}$$

$$\hat{p} = \frac{1}{b}\begin{bmatrix} \frac{\rho}{\gamma}b - a \\ \gamma \\ 0 \end{bmatrix}, \quad \lambda\hat{q} = \begin{bmatrix} \sqrt{\delta}+1 \\ -\frac{\gamma}{\rho}\left(\sqrt{\delta} + \frac{\rho c}{\gamma b}\right) \end{bmatrix} = \begin{bmatrix} \sqrt{\delta}+1 \\ -\frac{\gamma}{\rho}\sqrt{\delta}\left(\sqrt{\delta}+1\right) \end{bmatrix}.$$

Po vykrácení dostaneme

$$\hat{p}\hat{q}^T = \frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}} \\ -1 \end{bmatrix}^T = -\frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \pm\sqrt{\frac{\rho b}{\gamma c}} \\ 1 \end{bmatrix}^T,$$

což po dosazení do (342) dává (345).
(c) Pro metodu hodnosti 1 lze dosazením zjistit, že

$$\delta = \frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c - b}{\frac{\rho}{\gamma}b - a}$$

a $\mu = 0$, takže

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c - b}{\frac{\rho}{\gamma}b - a} + \sqrt{\delta}\right) = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta}\left(\sqrt{\delta}+1\right),$$

$$\hat{p} = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}, \quad \lambda\hat{q} = \sqrt{\delta}\begin{bmatrix} 1 \\ -\frac{\gamma}{\rho} \end{bmatrix} = \frac{\gamma}{\rho}\sqrt{\delta}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}.$$

Po vykrácení dostaneme

$$\hat{p}\hat{q}^T = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T}{\left(\frac{\rho}{\gamma}b - a\right)(\sqrt{\delta}+1)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T (\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c-b}{\frac{\rho}{\gamma}b-a} - 1\right)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T (\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}\right)^2 c - 2\left(\frac{\rho}{\gamma}\right)b + a},$$

což po dosazení do (342) dává (346). Jmenovatel v posledním vzorci je kvadratický výraz v ρ/γ. Jeho diskriminant $b^2 - ac$ je podle předpokladu záporný, takže jmenovatel nemůže být nikdy nulový. Aby byl čitatel nulový, muselo by platit $\delta = 1$, což po dosazení a po úpravě dává $(\rho/\gamma)^2 c - 2(\rho/\gamma)b + a = 0$. Tato rovnost, jak jsme právě dokázali, nemůže nastat. Jelikož $ac - b^2 > 0$, je alespoň jeden z výrazů $(\rho/\gamma)c - b$ a $(\rho/\gamma)b - a$ nenulový. Pokud $\delta = 0$, je $(\rho/\gamma)c - b = 0$ a tedy $(\rho/\gamma)b - a \neq 0$, což zajišťuje existenci metody hodnosti 1 (konečnost hodnoty parametru η) v případě, že $\delta = 0$. □

Poznámka 133. Ve větě 85 jsme předpokládali, že buď $\delta > 0$ nebo $\mu > 0$, neboť v opačném případě není matice $\hat{p}\hat{q}^T$ vystupující v (341) definovaná. I v tomto případě je však možné vyjádřit aktualizaci (286) v součinovém tvaru. Hodnota $\mu = 0$ odpovídá metodě hodnosti 1, pro kterou (po vykrácení výrazem $\sqrt{\delta}$ umožněným spojitou závislostí δ na η) platí (346). Pokud $ac - b^2 > 0$, nemůže být jmenovatel ani čitatel v (346) nulový.

Poznámka 134. Vzorec (346) lze upravit tak, aby se v něm neodečítala blízká čísla. Dosadíme-li do (346) výraz pro δ a rozšíříme-li zlomek číslem $\sqrt{\delta} + 1$, vykrátí se nový čitatel s původním jmenovatelem a po úpravách dostaneme

$$\frac{1}{\sqrt{\gamma}}S_+^{R1} = S + \frac{1}{\frac{\rho}{\gamma}b - a \pm \sqrt{\frac{\rho}{\gamma}\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}c - b\right)}}\left(\frac{\rho}{\gamma}d - SS^T y\right)\left(\frac{\rho}{\gamma}\tilde{d} - \tilde{y}\right)^T.$$

V součinovém tvaru lze vyjádřit také vztah (306). Z praktických důvodů se inverzní součinový vztah používá pouze v případě, že matice B je regulární, tedy v případě, že $m \geq n$ (potřebujeme řešit soustavu $Bs + g = 0$).

Poznámka 135. Má-li matice $S \in R^{n\times m}$, $m \geq n$, plnou hodnost, lze použitím definice 37 snadno ověřit, že

$$S^\dagger = S^T(SS^T)^{-1}, \qquad (SS^T)^{-1} = (S^\dagger)^T S^\dagger.$$

Položíme-li $A = S^\dagger \in R^{m\times n}$, platí

$$B = H^{-1} = (SS^T)^{-1} = (S^\dagger)^T S^\dagger = A^T A$$

a použijeme-li (336) dostaneme

$$AU = \tilde{U} \quad \Rightarrow \quad A^T\tilde{U} = A^T AU = BU. \tag{347}$$

Předpokládejme, že $B = A^T A$ a $B_+ = A_+^T A_+$, kde matice $A \in R^{m\times n}$ a $A_+ \in R^{m\times n}$ mají plnou hodnost a $m \geq n$. To nastává například tehdy, když $F(x) = (1/2)f^T(x)f(x)$ (minimalizace součtu čtverců), a matice A aproximuje Jacobiovu matici zobrazení $f : R^n \to R^m$.

Věta 86. *Nechť $a > 0$, $b > 0$, $c > 0$ a $ac - b^2 > 0$. Uvažujme aktualizaci (306), kde $B = A^T A$, $\rho > 0$, $\gamma > 0$. Nechť $\delta \geq 0$, $\mu \geq 0$ a buď $\delta > 0$ nebo $\mu > 0$. Nechť $\tilde{d} = Ad$ a $\tilde{y} = A(A^T A)^{-1}y$ (takže $y = A^T\tilde{y}$). Pak platí $B_+ = A_+^T A_+$, kde*

$$\sqrt{\gamma}A_+ = A - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T(BU)^T, \tag{348}$$

přičemž $\hat{p}$ a $\hat{q}$ jsou vektory vystupující ve větě 85.

Důkaz Podle (323) platí

$$\frac{1}{\gamma}S_+S_+^T = (I+pq^T)SS^T(I+qp^T),$$

což s použitím lemmatu 31 (e), vztahu (337) a rovnosti $q^Tp = \tilde{g}^T\tilde{p}$ dává

$$\gamma A_+^TA_+ = (I+qp^T)^{-1}A^TA(I+pq^T)^{-1} = \left(I - \frac{1}{\sqrt{\delta}}qp^T\right)A^TA\left(I - \frac{1}{\sqrt{\delta}}pq^T\right).$$

Platí tedy

$$\sqrt{\gamma}A_+ = A\left(I - \frac{1}{\sqrt{\delta}}pq^T\right) = A - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T(BU)^T,$$

neboť podle (347) platí $Ap = AU\hat{p}$ a $q = A\tilde{q} = A^T\tilde{U}\hat{q} = BU\hat{q}$. □

Poznámka 136. Použijeme-li matice $\hat{p}^T\hat{q}$ získané v důkazu důsledku 13, zjistíme, že pro metodu DFP platí

$$\sqrt{\gamma}A_+^{DFP} = A - \frac{1}{b}\left(\tilde{d} \pm \sqrt{\frac{\gamma b}{\rho a}}\tilde{y}\right)y^T, \tag{349}$$

pro metodu BFGS platí

$$\sqrt{\gamma}A_+^{BFGS} = A - \frac{1}{c}\tilde{d}\left(A^TAd \pm \sqrt{\frac{\gamma c}{\rho b}}y\right)^T \tag{350}$$

a pro metodu hodnosti 1 platí

$$\sqrt{\gamma}A_+^{R1} = A + \frac{1/\sqrt{\delta}-1}{(\gamma/\rho)^2a - 2(\gamma/\rho)b + c}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho}y - A^TAd\right)^T, \tag{351}$$

kde $1/\delta = (\gamma/\rho)((\gamma/\rho)a - b)/((\gamma/\rho)b - c)$. Vzorec (351) lze upravit na tvar

$$\sqrt{\gamma}A_+^{R1} = A + \frac{1}{\frac{\gamma}{\rho}b - c \pm \sqrt{\frac{\gamma}{\rho}\left(\frac{\gamma}{\rho}b - c\right)\left(\frac{\gamma}{\rho}a - b\right)}}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho}y - A^TAd\right)^T.$$

Ve všech těchto vzorcích je $y = A^T\tilde{y}$ a $A^TAd = A^T\tilde{d}$ (neboť $\tilde{d} = Ad$). Poznamenejme, že pro minimalizaci součtu čtverců má praktický význam pouze metoda BFGS, která používá jediný redukovaný vektor $\tilde{d} = Ad$. Ostatní metody potřebují navíc vektor $\tilde{y} = A(A^TA)^{-1}y$, takže je nutné invertovat matici A^TA.

Součinový tvar $H = SS^T$, kde $S \in R^{n\times m}$ a $m \le n$, lze modifikovat tak, že se místo matice S používá matice Z, jejíž sloupce tvoří ortonormální bázi v $\mathcal{L}(S)$.

Věta 87. *Nechť $H = SS^T$, kde matice $S \in R^{n\times m}$, $m \le n$, má plnou hodnost, a $Z \in R^{n\times m}$ je matice jejíž sloupce tvoří bázi v $\mathcal{L}(S)$. Pak platí $H = Z(Z^TBZ)^{-1}Z^T$, kde $B = H^\dagger$.*

Důkaz Jelikož sloupce matice Z tvoří bázi v $\mathcal{L}(S)$, existuje čtvercová regulární matice M taková, že $S = ZM$. Platí tedy

$$SS^T = ZMM^TZ^T.$$

Použitím definice 37 se snadno ověří, že $S^\dagger = M^{-1}Z^\dagger$ (neboť $Z^\dagger Z = I$ podle (324)). Protože podle (325) platí $(SS^T)^\dagger = (S^\dagger)^TS^\dagger$, můžeme psát

$$Z^T(SS^T)^\dagger Z = Z^T(S^\dagger)^TS^\dagger Z = Z^T(M^{-1}Z^\dagger)^TM^{-1}Z^\dagger Z = (MM^T)^{-1}, \tag{352}$$

takže

$$Z(Z^TBZ)^{-1}Z^T = Z(Z^T(SS^T)^\dagger Z)^{-1}Z^T = ZMM^TZ = SS^T = H$$

. □

Poznámka 137. V předchozí větě nejsou kladeny žádné požadavky na výběr matice Z, takže lze položit $Z = S$. V tomto případě $M = I$, takže podle (352) platí $S^T BS = S^T (SS^T)^\dagger S = I$ (sloupce matice S jsou B-ortogonální). Odtud plyne, že $H = S(S^T BS)^{-1} S^T = SS^T$.

Poznámka 138. V dalším textu budeme používat redukované matice $\tilde{B} = Z^T BZ$ a $\tilde{H} = \tilde{B}^{-1}$. Směrový vektor $s = -Hg$ se vypočte podle vzorců

$$s = Z\tilde{s}, \qquad \tilde{s} = -\tilde{H}\tilde{g}, \qquad \tilde{g} = Z^T g$$

(vektor $\tilde{s}$ lze také získat řešením soustavy rovnic $\tilde{B}\tilde{s} = -\tilde{g}$). Pak $\tilde{d} = \alpha\tilde{s}$ a $\tilde{y} = Z^T y$. Poznamenejme, že v těchto vzorcích se používá pouze redukovaná matice $\tilde{H}$ (nebo $\tilde{B}$) a matice Z jejíž sloupce tvoří bázi v $\mathcal{L}(S)$.

Lemma 36. *Matice $\tilde{B}$ a $\tilde{H}$ jsou pozitivně definitní.*

Důkaz Podle (352) platí $\tilde{B} = (MM^T)^{-1}$, kde M je čtvercová regulární matice, takže matice MM^T je pozitivné definitní. Odtud plyne pozitivní definitnost matic $\tilde{B}$ a $\tilde{H}$. □

Nyní budeme předpokládat, že matice Z má ortonormální sloupce, takže $Z^T Z = I$. Použitím definice 37 snadno ověříme, že v tomto případě platí $Z^\dagger = Z^T$.

Lemma 37. *Nechť $Z^T Z = I$ a $\tilde{H} = \tilde{B}^{-1} = (Z^T BZ)^{-1}$, kde $B = H^\dagger = (SS^T)^\dagger$. Pak platí $\tilde{H} = Z^T HZ$ a $B = Z\tilde{B}Z^T$.*

Důkaz Podle (352) platí $\tilde{B} = (MM^T)^{-1}$, takže $\tilde{H} = \tilde{B}^{-1} = MM^T$. Pokud $Z^T Z = I$, můžeme psát

$$Z^T HZ = Z^T SS^T Z = Z^T (ZMM^T Z^T)Z = MM^T = \tilde{H}.$$

Jelikož $Z^T Z = I$, platí $Z^\dagger = Z^T$, takže podobně jako v důkazu věty 87 lze psát $S^\dagger = M^{-1} Z^\dagger = M^{-1} Z^T$. Platí tedy

$$B = (SS^T)^\dagger = (S^\dagger)^T S^\dagger = (M^{-1}Z^T)^T M^{-1} Z^T = Z(MM^T)^{-1} Z^T = Z\tilde{B}Z^T.$$

□

Používáme-li vyjádření $H = Z\tilde{H}Z^T$, matice Z se nemění. Místo toho se aktualizuje matice $\tilde{H} \in R^{m\times m}$.

Věta 88. *Nechť $H = Z\tilde{H}Z^T$, $M \in R^{2\times 2}$ a*

$$\frac{1}{\gamma}\tilde{H}_+ = \tilde{H} + \tilde{U}M\tilde{U}, \qquad \tilde{U} = [\tilde{d}, \tilde{H}\tilde{y}].$$

Pak pro matici $H_+ = Z\tilde{H}_+ Z^T$ platí

$$\frac{1}{\gamma}H_+ = H + UMU, \qquad U = [d, Hy].$$

Důkaz Podle předpokladu platí

$$\frac{1}{\gamma}H_+ = \frac{1}{\gamma}Z\tilde{H}_+ Z^T = Z\tilde{H}Z^T + Z\tilde{U}M\tilde{U}Z^T.$$

Ale $Z\tilde{H}Z^T = H$ a $Z\tilde{U} = [Z\tilde{d}, Z\tilde{H}Z^T y] = [d, Hy] = U$. Platí tedy $(1/\gamma)H_+ = H + UMU$. □

Poznámka 139. Metody s proměnnou metrikou, které používají redukované matice $\tilde{H}$ nebo $\tilde{B}$ a redukované gradienty $\tilde{g}$ se nazývají metodami redukovaných gradientů. Tyto metody používají v prvním iteračním kroku libovolnou pozitivně definitní matici $\tilde{H}$ nebo $\tilde{B}$, například jednotkovou matici, která se v dalších iteračních krocích aktualizuje podle vzorců

$$\frac{1}{\gamma}\tilde{H}_+ = \tilde{H} + \frac{\rho}{\gamma b}\tilde{d}\tilde{d}^T - \frac{1}{a}\tilde{H}\tilde{y}(\tilde{H}\tilde{y})^T + \frac{\eta}{a}\left(\frac{a}{b}\tilde{d} - \tilde{H}\tilde{y}\right)\left(\frac{a}{b}\tilde{d} - \tilde{H}\tilde{y}\right)^T,$$

nebo

$$\gamma\tilde{B}_+ = \tilde{B} + \frac{\gamma}{\rho b}\tilde{y}\tilde{y}^T - \frac{1}{c}\tilde{B}\tilde{d}(\tilde{B}\tilde{d})^T + \frac{\beta}{c}\left(\frac{c}{b}\tilde{y} - \tilde{B}\tilde{d}\right)\left(\frac{c}{b}\tilde{y} - \tilde{B}\tilde{d}\right)^T,$$

kde podle lemmatu 37 platí

$$\begin{aligned} a &= y^T H y = y^T Z\tilde{H}Z^T y = \tilde{y}^T\tilde{H}\tilde{y}, \\ b &= y^T d = y^T Z\tilde{d} = \tilde{y}^T\tilde{d}, \\ c &= d^T B d = d^T Z^T (Z\tilde{B}Z^T) Z\tilde{d} = \tilde{d}^T\tilde{B}\tilde{d}. \end{aligned}$$

Poznámka 140. V předchozím výkladu jsme narazili na jistá omezení, která musí splňovat některé významné metody z Broydenovy třídy. Proto se definují různé části této třídy.

(a) Semidefinitní metody, kdy $\eta \geq \eta^*$ a $\beta \geq \beta^*$.
(b) Rozložitelné metody, kdy $\delta \geq 0$ (takže $\eta \geq \eta^*$ a $\beta \geq \beta^*$) a $\mu \geq 0$. Dosazením za μ se snadno přesvědčíme, že pokud $b/c \leq \rho/\gamma \leq a/b$, je každá semidefinitní metoda rozložitelná. Označme η^{R1} hodnotu odpovídající metodě hodnosti 1. Pokud $0 < \rho/\gamma \leq b/c$, jsou rozložitelné ty metody pro něž $\eta \geq \eta^{R1}$, kde $\eta^* < \eta^{R1} < 0$. Pokud $a/b \leq \rho/\gamma$, jsou rozložitelné ty metody pro něž $\eta^* \leq \eta \leq \eta^{R1}$, kde $\eta^{R1} > 1$.
(c) Perfektní metody, kdy $\eta \geq 0$ a $\beta^* \leq \beta \leq 1$.
(d) Omezené metody, kdy $0 \leq \eta \leq 1$ a $0 \leq \beta \leq 1$. Tyto metody jsou též rozložitelné a perfektní.

Metody DFP, BFGS a Hoshinova metoda jsou omezené. Metoda hodnosti 1 je semidefinitní pouze tehdy, když buď $0 < \rho/\gamma \leq b/c$ nebo $a/b \leq \rho/\gamma$. V tomto případě je tato metoda rozložitelná a jestliže $a/b \leq \rho/\gamma$ i perfektní. Metoda hodnosti 1 není nikdy omezená.

Poznámka 141. Metody, které nejsou perfektní, jsou obvykle málo efektivní. Proto je účelné metodu hodnosti 1 kombinovat s metodou BFGS tak, že pokládáme $\eta = \eta^{R1+}$, kde

$$\eta^{R1+} = \eta^{R1}, \qquad a/b \leq \rho/\gamma, \tag{353}$$
$$\eta^{R1+} = 1, \qquad a/b > \rho/\gamma. \tag{354}$$

Numerické experimenty ukazují, že zhruba čtvrtina iteračních kroků této modifikované metody používá hodnotu $\eta = \eta^{R1}$. Ve zbylých iteracích se používá hodnota $\eta = \eta^{BFGS} = 1$.

4.3 Variační odvození metod s proměnnou metrikou

Velmi zajímavý způsob jak lze získat metody s proměnnou metrikou spočívá v použití minimalizačního principu [68], [73]. V tomto případě hledáme minimum maticové funkce $\psi : R^{n\times n} \to R$ na množině symetrických matic splňujících kvazinewtonovskou podmínku. O funkci $\psi(X)$ předpokládáme, že je symetrická (platí $\psi(X^T) = \psi(X)$) a ryze konvexní. V tomto případě je stacionární bod Lagrangeovy funkce jediným řešením dané úlohy.

Lemma 38. *Nechť $\psi(X)$ je symetrická ryze konvexní maticová funkce. Pak matice $X^* \in R^{n\times n}$ minimalizuje funkci $\psi(X)$ na množině symetrických matic řádu n splňujících podmínku $Xp = q$ právě tehdy existuje-li vektor Lagrangeových multiplikátorů $u \in R^n$ takový, že*

$$\frac{\partial\psi(X^*)}{\partial X} = up^T + pu^T. \tag{355}$$

Zde $\partial\psi(X)/\partial X$ označuje matici, která má prvky $\partial\psi(X)/\partial x_{kl}$, $1 \leq k \leq n$, $1 \leq l \leq n$ (x_{kl} jsou prvky matice X).

Důkaz Lagrangeova funkce uvažované úlohy má tvar

$$L(X,u,V) = \psi(X) + 2\sum_{i=1}^{n} u_i \left(q_i - \sum_{j=1}^{n} x_{ij}p_j \right) + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}(x_{ij} - x_{ji})$$

(poslední člen zajišťuje symetrii matice X). Podmínky optimality mají tvar

$$\begin{aligned} \frac{\partial L(X,u,V)}{\partial x_{kl}} &= \frac{\partial \psi(X)}{\partial x_{kl}} - 2u_k p_l + v_{kl} - v_{lk} = 0, \\ \frac{\partial L(X,u,V)}{\partial x_{lk}} &= \frac{\partial \psi(X)}{\partial x_{lk}} - 2u_l p_k + v_{lk} - v_{kl} = 0, \end{aligned}$$

kde k, l je libovolná dvojice indexů. Sečteme-li obě rovnosti a použijeme-li symetrii funkce $\psi(X)$, dostaneme

$$2\frac{\partial \psi(X)}{\partial x_{kl}} - 2u_k p_l - 2u_l p_k = 0,$$

což maticově zapsáno dává (355), a jelikož funkce $\psi(X)$ je ryze konvexní, dostaneme tvrzení lematu. □

Pro variační odvození metod s proměnou metrikou se nejčastěji používá Frobeniova norma matice. V tomto případě má funkce $\psi(x)$ tvar $\psi(x) = (1/2)\|X\|_F^2 = (1/2)\mathrm{Tr}X^T X$.

Lemma 39. *Funkce $\psi(X) = (1/2)\|X\|_F^2$ je symetrická a ryze konvexní na $R^{n\times n}$.*

Důkaz Symetrie je zřejmá. Nechť $X_1 \in R^{n\times n}$, $X_2 \in R^{n\times n}$, $X_1 \neq X_2$ a $X = \lambda_1 X_1 + \lambda_2 X_2$, kde $\lambda_1 > 0$, $\lambda_2 > 0$ a $\lambda_1 + \lambda_2 = 1$. Pak platí

$$\begin{aligned} \|X\|_F^2 &= \sum_{i=1}^{n}\sum_{j=1}^{n}(e_i^T X e_j)^2 = \sum_{i=1}^{n}\sum_{j=1}^{n}(\lambda_1 e_i^T X_1 e_j + \lambda_2 e_i^T X_2 e_j)^2 \\ &< \sum_{i=1}^{n}\sum_{j=1}^{n}(\lambda_1 (e_i^T X_1 e_j)^2 + \lambda_2 (e_i^T X_2 e_j)^2) = \lambda_1\|X_1\|_F^2 + \lambda_1\|X_2\|_F^2, \end{aligned}$$

neboť druhá mocnina je ryze konvexní (platí $\mathrm{d}^2(x^2)/\mathrm{d}x^2 = 2 > 0$). □

Lemma 40. *Symetrická matice X^* má minimální Frobeniovu normu na množině symetrických matic řádu n splňujících podmínku $Xp = q$ právě tehdy, když*

$$X^* = \frac{1}{p^T p}(pq^T + qp^T) - \frac{q^T p}{(p^T p)^2} pp^T. \tag{356}$$

Důkaz Jelikož pro funkci

$$\psi(x) = \frac{1}{2}\|X\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}^2$$

lze psát $\partial\psi(X)/\partial X = X$, musí podle lemmatu 38 platit $X^* = up^T + pu^T$. Z podmínky $X^*p = q$ dostaneme $up^T p + pu^T p = q$, neboli

$$u = \frac{1}{p^T p}(q - u^T pp),$$

takže

$$u^T p = \frac{1}{p^T p}(q^T p - u^T pp^T p),$$

neboli $2u^T p = q^T p/p^T p$, což dává

$$u = \frac{1}{p^T p}\left(q - \frac{1}{2}\frac{q^T p}{p^T p}p\right).$$

Dosadíme-li tento vektor do vztahu $X^* = up^T + pu^T$ a uvážíme-li konvexitu funkce $\psi(X)$, dostaneme tvrzení lemmatu. □

Věta 89. *Nechť W je symetrická pozitivně definití matice. Pak symetrická matice H_+ minimalizuje Frobeniovu normu $\|W^{-1/2}((1/\gamma)\tilde{H} - H)W^{-1/2}\|_F$ na množině symetrických matic $\tilde{H}$ řádu n splňujících kvazinewtonovskou podmínku*

$$\left(\frac{1}{\gamma}\tilde{H} - H\right) y = \frac{\rho}{\gamma} d - Hy$$

právě tehdy, když

$$\frac{1}{\gamma}H_+ = H + \frac{Wy((\rho/\gamma)d - Hy)^T + ((\rho/\gamma)d - Hy)(Wy)^T}{y^TWy} - \frac{((\rho/\gamma)d - Hy)^Ty}{y^TWy}\frac{Wy(Wy)^T}{y^TWy}. \tag{357}$$

Důkaz Položme $X = W^{-1/2}((1/\gamma)\tilde{H} - H)W^{-1/2}$ a $w = (\rho/\gamma)d - Hy$. Jelikož kvazinewtonovskou podmínku lze zapsat ve tvaru

$$W^{-1/2}\left(\frac{1}{\gamma}\tilde{H} - H\right)W^{-1/2}W^{1/2}y = W^{-1/2}w,$$

neboli $Xp = q$, kde $p = W^{1/2}y$ a $q = W^{-1/2}w$, můžeme použít lemma 40, podle kterého

$$\begin{aligned} X^* &= \frac{1}{p^Tp}(pq^T + qp^T) - \frac{q^Tp}{(p^Tp)^2}pp^T \\ &= \frac{1}{y^TWy}(W^{1/2}yw^TW^{-1/2} + W^{-1/2}wy^TW^{1/2}) - \frac{w^Ty}{(y^TWy)^2}W^{1/2}yy^TW^{1/2}. \end{aligned}$$

Jelikož $X^* = W^{-1/2}((1/\gamma)H_+ - H)W^{-1/2}$, platí $W^{1/2}X^*W^{1/2} = (1/\gamma)H_+ - H$, odkud plyne tvrzení věty. □

Poznámka 142. Zvolíme-li matici W tak, aby platilo $Wy = d$, přejde vzorec (357) na vztah (297). Metoda BFGS tedy minimalizuje Frobeniovu normu $\|W^{-1/2}((1/\gamma)H_+ - H)W^{-1/2}\|_F$, pokud $Wy = d$.

Je zřejmé, že metoda získaná aktualizací (357) patří do Broydenovy třídy právě tehdy, je-li vektor Wy lineární kombinací vektorů d a Hy. Protože aktualizace (357) nezávisí na normě vektoru Wy, budeme předpokládat, že $Wy = \vartheta d - Hy$ (takže lze volit $Wy = d$, pokud $\vartheta = \infty$).

Věta 90. *Nechť jsou splněny předpoklady věty 89 a nechť $Wy = \vartheta d - Hy$. Pak aktualizace (357) je ekvivalentní aktualizaci (286), pokud*

$$\eta = \frac{b(\vartheta^2 b - (\rho/\gamma)a)}{(\vartheta b - a)^2}. \tag{358}$$

Jestliže $\eta = 1$, pak buď $\vartheta = \infty$, takže $Wy = d$, nebo $\vartheta = (a/b + \rho/\gamma)/2$. V ostatních případech platí

$$\vartheta = \frac{a}{\eta - 1}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

Důkaz Položme $Wy = \vartheta d - Hy$ a označme $w = (\rho/\gamma)d - Hy$. Pak lze psát

$$\frac{1}{\gamma}H_+ = H + \frac{Wyw^T + w(Wy)^T}{y^TWy} - \frac{w^Ty}{y^TWy}\frac{Wy(Wy)^T}{y^TWy}, \tag{359}$$

přičemž $y^TWy = \vartheta b - a$ a $w^Ty = (\rho/\gamma)b - a$. Nyní porovnáme členy s maticí $Hy(Hy)^T$ v (286) a v (361) (v (286) je u této matice koeficient $m_3 = (\eta-1)/a$). Každý z výrazů Wyw^T, $w(Wy)^T$, $Wy(Wy)^T$ přispívá jednou maticí $Hy(Hy)^T$, takže podle (361) platí

$$\frac{2}{\vartheta b - a} - \frac{(\rho/\gamma)b - a}{(\vartheta b - a)^2} = \frac{\eta - 1}{a},$$

neboli

$$\eta = \frac{2a(\vartheta b - a) - a((\rho/\gamma)b - a) + (\vartheta b - a)^2}{(\vartheta b - a)^2} = \frac{b(\vartheta^2 b - (\rho/\gamma)a)}{(\vartheta b - a)^2}.$$

Pokud $\eta = 1$ můžeme volit $\vartheta = \infty$, takže $Wy = d$, nebo

$$\vartheta^2 b^2 - \frac{\rho}{\gamma}ab = (\vartheta b - a)^2 \;\Rightarrow\; \vartheta = \frac{1}{2}\left(\frac{a}{b} + \frac{\rho}{\gamma}\right).$$

V obecném případě lze psát

$$\eta(\vartheta^2 b^2 - 2\vartheta ab + a^2) = \vartheta^2 b^2 - \frac{\rho}{\gamma}ab,$$

což po vydělení číslem ab^2 dává

$$\frac{\eta - 1}{a}\vartheta^2 - 2\frac{\eta}{b}\vartheta + \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right) = m_3\vartheta^2 + 2\,m_2\vartheta + m_1 = 0,$$

kde m_1, m_2, m_3 jsou čísla určená vztahy (283). Tato kvadratická rovnice má řešení

$$\vartheta = \frac{-m_2 \pm \sqrt{m_2^2 - m_1 m_3}}{m_3} = \frac{-m_2 \pm \sqrt{\mu}}{m_3} = \frac{a}{\eta - 1}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

□

Poznámka 143. Věta 90 udává způsob, jak lze k dané metodě s proměnnou metrikou (charakterizované parametrem η) nalézt aktualizaci tvaru (357). K dané hodnotě η najdeme podle věty 90 hodnotu ϑ určující vektor Wy (existují obvykle dvě řešení).

(a) Pro metodu DFP platí $\eta = 0$ a $\mu = \rho/(\gamma ab)$, takže $\vartheta = -\sqrt{\rho a/(\gamma b)}$ (volíme zápornou hodnotu, aby byla splněna nerovnost $y^TWy > 0$).
(b) Pro metodu BFGS platí $\eta = 1$, takže lze volit $\vartheta = \infty$, neboli $Wy = d$.
(c) Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$ a $\mu = 0$, takže $\vartheta = \rho/\gamma$.

Poznámka 144. Analogický postup lze použít pro aktualizaci matice B. Nechť W je symetrická pozitivně definití matice. Pak symetrická matice B_+ minimalizuje Frobeniovu normu $\|W^{-1/2}(\gamma\tilde{B} - B)W^{-1/2}\|_F$ na množině symetrických matic $\tilde{B}$ řádu n splňujících kvazinewtonovskou podmínku

$$\left(\gamma\tilde{B} - B\right)d = \frac{\gamma}{\rho}y - Bd$$

právě tehdy, když

$$\gamma B_+ = B + \frac{Wd((\gamma/\rho)y - Bd)^T + ((\gamma/\rho)y - Bd)(Wd)^T}{d^TWd} - \frac{((\gamma/\rho)y - Bd)^Td}{d^TWd}\frac{Wd(Wd)^T}{d^TWd}. \tag{360}$$

Poznámka 145. Zvolíme-li matici W tak, aby platilo $Wd = y$, přejde vzorec (360) na vztah duální k (297). Metoda DFP tedy minimalizuje Frobeniovu normu $\|W^{-1/2}(\gamma B_+ - BH)W^{-1/2}\|_F$, pokud $Wd = y$.

Poznámka 146. Zvolíme-li matici W tak že $Wd = \vartheta y - Bd$, je aktualizace (360) ekvivalentní aktualizaci (306), pokud

$$\beta = \frac{b(\vartheta^2 b - (\gamma/\rho)c)}{(\vartheta b - c)^2}.$$

Jestliže $\beta = 1$, pak buď $\vartheta = \infty$, takže $Wd = y$, nebo $\vartheta = (c/b + \gamma/\rho)/2$. V ostatních případech platí

$$\vartheta = \frac{c}{\beta - 1}\left(\frac{\beta}{b} \pm \sqrt{\frac{\mu}{\delta}}\right)$$

(číslo μ/δ je určeno vzorcem (318)).

(a) Pro metodu DFP platí $\beta = 1$, takže lze volit $\vartheta = \infty$, neboli $Wd = y$.
(b) Pro metodu BFGS platí $\beta = 0$ a $\mu/\delta = \gamma/(\rho bc)$, takže $\vartheta = -\sqrt{\gamma c/(\rho b)}$ (volíme zápornou hodnotu, aby byla splněna nerovnost $d^T W d > 0$).
(c) Pro metodu hodnosti 1 platí $\beta = (\gamma/\rho)/(\gamma/\rho - c/b)$ a $\mu/\delta = 0$, takže $\vartheta = \gamma/\rho$.

Poznamenejme, že aktualizaci (360) používají strukturované metody s proměnnou metrikou pro minimalizaci součtu čtverců (vzorec (674)).

Poznámka 147. Zvolíme-li $W = I$, dostaneme metodu, která nepatří do Broydenovy třídy a která se nazývá Powellovou symetrizací Broydenovy metody. Platí

$$\gamma B_+^{PSB} = B + \frac{d((\gamma/\rho)y - Bd)^T + ((\gamma/\rho)y - Bd)d^T}{d^T d} - \frac{((\gamma/\rho)y - Bd)^T d}{d^T d}\frac{dd^T}{d^T d}.$$

Metoda PSB nezaručuje pozitivní definitnost matice B_+, takže nemusí globálně konvergovat. Přesto je této, obecně velmi neefektivní, metodě věnována velká publicita, která souvisí s její příbuzností s některými metodami pro řídké úlohy (věta 221).

Kromě Frobeniovy normy lze použít i jiná minimalizační kriteria. Velmi se osvědčila funkce $\psi(X) = \operatorname{Tr} X - \ln\det X$ definovaná na množině symetrických pozitivně definitních matic. Nechť λ_i, $1 \le i \le n$, jsou vlastní čísla symetrické pozitivně definitní matice X. Pak platí

$$\operatorname{Tr} X - \ln\det X = \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \ln \lambda_i^{-1},$$

takže minimalizací této funkce lze zajistit, že vlastní čísla matice X nebudou ani příliš malá ani příliš veliká. Je důležité, že použití této funkce vede na aktualizaci s nejvýše dvěma korekčními členy [54].

Lemma 41. *Funkce $\psi(X) = \operatorname{Tr} X - \ln\det X$ je symetrická a ryze konvexní na množině symetrických pozitivně definitních matic.*

Důkaz (a) Symetrie je zřejmá, plyne z rovností $\operatorname{Tr}(X^T) = \operatorname{Tr} X$ a $\det(X^T) = \det X$.

(b) Ukážeme nejprve, že funkce $\ln\det X$ je ryze konkávní na množině symetrických pozitivně definitních matic. Nechť $X_1 \succ 0$, $X_2 \succ 0$, $X_1 \ne X_2$ a $X = \lambda_1 X_1 + \lambda_2 X_2$, kde $\lambda_1 > 0$, $\lambda_2 > 0$ a $\lambda_1 + \lambda_2 = 1$. Pak matice X a $X_2^{-1/2} X_1 X_2^{-1/2}$ jsou pozitivně definitní a existuje regulární čtvercová matice W (jejímiž sloupci jsou ortonormální vlastní vektory matice $X_2^{-1/2} X_1 X_2^{-1/2}$) taková, že $W^T X_2^{-1/2} X_1 X_2^{-1/2} W = D$ a $W^T W = I$, kde $D = \operatorname{diag}(\delta_1, \ldots, \delta_n)$ je pozitivně definitní diagonální matice. Pak, položíme-li $V = X_2^{-1/2} W$, platí $V^T X_1 V = D$, $V^T X_2 V = I$ a

$$V^T X V = \lambda_1 V^T X_1 V + \lambda_2 V^T X_2 V = \lambda_1 D + \lambda_2 I,$$

kde $D \neq I$ (neboť $X_1 \neq X_2$). Zřejmě

$$\begin{aligned}\ln\det X_1 + 2\ln\det V &= \ln\det(V^T X_1 V) = \ln\det D = \sum_{i=1}^{n} \ln\delta_i,\\ \ln\det X_2 + 2\ln\det V &= \ln\det(V^T X_2 V) = \ln\det I = 0,\end{aligned}$$

a jelikož funkce logaritmus je ryze konkávní, můžeme psát

$$\begin{aligned}\ln\det X + 2\ln\det V &= \ln\det(V^T X V) = \ln\det(\lambda_1 D + \lambda_2 I) = \sum_{i=1}^{n} \ln(\lambda_1\delta_i + \lambda_2)\\ &> \sum_{i=1}^{n} (\lambda_1 \ln\delta_i + \lambda_2 \ln 1) = \lambda_1 \ln\det D + \lambda_2 \ln\det I\\ &= \lambda_1 \ln\det(V^T X_1 V) + \lambda_2 \det(V^T X_2 V)\\ &= \lambda_1 \ln\det X_1 + \lambda_2 \ln\det X_2 + 2\ln\det V,\end{aligned}$$

což dává $\ln\det(\lambda_1 X_1 + \lambda_2 X_2) = \ln\det X > \lambda_1 \ln\det X_1 + \lambda_2 \ln\det X_2$.

(c) Jelikož funkce $\mathrm{Tr}\, X$ je lineární (a tudíž konvexní) a funkce $\ln\det X$ je ryze konkávní, je funkce $\psi(X) = \mathrm{Tr}\, X - \ln\det X$ ryze konvexní. □

Lemma 42. *Nechť $\psi(X) = \mathrm{Tr}\, X - \ln\det X$. Pak, pokud $\det X > 0$, platí*

$$\frac{\partial\psi(X)}{\partial X} = I - (X^{-1})^T.$$

Důkaz (a) Vztah $\partial\,\mathrm{Tr}X/\partial X = I$, je zřejmý.

(b) Dokážeme nejprve, že $\partial\ln\det X/\partial X = (X^{-1})^T$. Použijeme-li Shermanův-Morrisonův vzorec, uvedený v poznámce 106, dostaneme

$$\frac{\partial\det X}{\partial x_{kl}} = \lim_{t\to 0}\frac{\det(X + te_k e_l^T) - \det X}{t} = \det X \lim_{t\to 0}\frac{1 + te_l^T X^{-1} e_k - 1}{t} = \det X\, e_l^T X^{-1} e_k,$$

takže

$$\frac{\partial\ln\det X}{\partial x_{kl}} = \frac{1}{\det X}\frac{\partial\det X}{\partial x_{kl}} = e_l^T X^{-1} e_k, \tag{361}$$

neboli $\partial\ln\det X/\partial X = (X^{-1})^T$.

(c) Spojíme-li oba výsledky dohromady, dostaneme $\partial\psi(X)/\partial X = I - (X^{-1})^T$. □

Lemma 43. *Symetrická matice X^* minimalizuje funkci $\psi(X) = \mathrm{Tr}\, X - \ln\det X$ na množině symetrických pozitivně definitních matic řádu n splňujících podmínku $Xp = q$ právě tehdy, když*

$$(X^*)^{-1} = I - \frac{1}{q^T p}(pq^T + qp^T) + \frac{1}{p^T q}pp^T + \frac{q^T q}{(p^T q)^2}pp^T. \tag{362}$$

Důkaz Jelikož podle lemmatu 42 lze psát $\partial\psi(X)/\partial X = I - (X^{-1})^T$ a matice X^* je symetrická, musí podle lemmatu 38 platit $(X^*)^{-1} = I - up^T - pu^T$. Z podmínky $(X^*)^{-1}q = p$ dostaneme $p = q - up^T q - pu^T q$, neboli

$$u = \frac{1}{p^T q}(q - p - u^T qp),$$

takže

$$u^T q = \frac{1}{p^T q}(q^T q - p^T q - u^T q p^T q),$$

neboli $2u^T q = q^T q/p^T q - 1$, což dává

$$u = \frac{1}{p^T q}\left(q - p - \frac{1}{2}\left(\frac{q^T q}{p^T q} - 1\right)p\right) = \frac{1}{p^T q}\left(q - \frac{1}{2}\left(\frac{q^T q}{p^T q} + 1\right)p\right).$$

Dosadíme-li tento vektor do vztahu $(X^*)^{-1} = I - up^T - pu^T$ a uvážíme-li konvexitu funkce $\psi(X)$, dostaneme tvrzení lemmatu. □

Věta 91. *Symetrická matice $H_+ = B_+^{-1}$ minimalizuje funkci $\psi((1/\gamma)H^{-1/2}\tilde{H}H^{-1/2})$ na množině symetrických matic $\tilde{H}$ řádu n splňujících kvazinewtonovskou podmínku $\tilde{H}y = \rho d$ právě tehdy, když*

$$\gamma B_+ = B - \frac{1}{b}(y(Bd)^T + Bdy^T) + \frac{1}{b}\left(\frac{\gamma}{\rho} + \frac{c}{b}\right)yy^T,$$

kde $B = H^{-1}$ a $a = y^T Hy$, $b = y^T d$, $c = d^T Bd$.

Důkaz Položme $X = (1/\gamma)H^{-1/2}\tilde{H}H^{-1/2}$. Jelikož kvazinewtonovskou podmínku lze zapsat ve tvaru

$$\frac{1}{\gamma}H^{-1/2}\tilde{H}H^{-1/2}H^{1/2}y = \frac{\rho}{\gamma}H^{-1/2}d,$$

neboli $Xp = q$, kde $p = H^{1/2}y$ a $q = (\rho/\gamma)H^{-1/2}d$, můžeme použít lemma 43, podle kterého

$$\begin{aligned}
(X^*)^{-1} &= I - \frac{1}{q^T p}(pq^T + qp^T) + \frac{1}{p^T q}pp^T + \frac{q^T q}{(p^T q)^2}pp^T \\
&= I - \frac{\gamma}{\rho b}\left(\frac{\rho}{\gamma}H^{1/2}yd^T H^{-1/2} + \frac{\rho}{\gamma}H^{-1/2}dy^T H^{1/2}\right) \\
&\quad + \frac{\gamma}{\rho b}H^{1/2}yy^T H^{1/2} + \left(\frac{\gamma}{\rho b}\right)^2\frac{\rho^2 c}{\gamma^2}H^{1/2}yy^T H^{1/2} \\
&= I - \frac{1}{b}(H^{1/2}yd^T H^{-1/2} + H^{-1/2}dy^T H^{1/2}) + \frac{1}{b}\left(\frac{\gamma}{\rho} + \frac{c}{b}\right)H^{1/2}yy^T H^{1/2}.
\end{aligned}$$

Jelikož $(X^*)^{-1} = \gamma B^{-1/2}B_+B^{-1/2}$, platí $B^{1/2}(X^*)^{-1}B^{1/2} = \gamma B_+$, odkud plyne tvrzení věty. □

Poznámka 148. Podle věty 91 minimalizuje metoda DFP funkci $\psi(X) = \operatorname{Tr} X - \ln\det X$, pokud $X = (1/\gamma)H^{-1/2}\tilde{H}H^{-1/2}$ a $\tilde{H}y = \rho d$.

Poznámka 149. Analogický postup lze použít pro aktualizaci matice H. Symetrická matice $B_+ = H_+^{-1}$ minimalizuje funkci $\gamma B^{-1/2}\tilde{B}B^{-1/2}$ na množině symetrických matic $\tilde{B}$ řádu n splňujících kvazinewtonovskou podmínku $\tilde{B}d = (1/\rho)y$ právě tehdy, když

$$\gamma H_+ = H - \frac{1}{b}(d(Hy)^T + Hyd^T) + \frac{1}{b}\left(\frac{\rho}{\gamma} + \frac{a}{b}\right)dd^T,$$

kde $H = B^{-1}$ a $a = y^T Hy$, $b = y^T d$, $c = d^T Bd$.

Poznámka 150. Podle poznámlky 149 minimalizuje metoda BFGS funkci $\psi(X) = \operatorname{Tr} X - \ln\det X$, pokud $X = \gamma B^{-1/2}\tilde{B}B^{-1/2}$ a $\tilde{B}d = (1/\rho)y$.

Minimalizační postup lze použít i k odvození součinového tvaru metod s proměnnou metrikou. V tomto případě dostaneme vyjádření, které je obecnější než (341) a které obsahuje i aktualizace hodnosti 2. Abychom mohli použít variační princip, zapíšeme kvazinewtonovskou podmínku ve tvaru

$$\frac{1}{\sqrt{\gamma}} S_+^T y = \tilde{z}, \qquad \frac{1}{\sqrt{\gamma}} S_+ \tilde{z} = \frac{\rho}{\gamma} d, \qquad \tilde{z}^T \tilde{z} = \frac{\rho}{\gamma} b, \tag{363}$$

kde $\tilde{z} \in R^m$ je volitelný vektor (parametr). Poznamenejme, že poslední rovnost, která je důsledkem prvních dvou rovností, je jediným omezením kladeným na volbu vektoru $\tilde{z}$.

Věta 92. *Nechť T je symetrická pozitivně definitní matice. Pak Frobeniova norma $\|T^{-1/2}(S_+/\sqrt{\gamma} - S)\|_F$ je minimální na množině všech matic splňujících kvazinewtonovskou podmínku (363) právě tehdy, platí-li*

$$\frac{1}{\sqrt{\gamma}} S_+ = S - \frac{Ty}{y^T T y} \tilde{y}^T + \left(\frac{\rho}{\gamma} d - z + \frac{y^T z}{y^T T y} Ty \right) \frac{\tilde{z}^T}{\tilde{z}^T \tilde{z}}, \tag{364}$$

kde $\tilde{y} = S^T y$ a $z = S\tilde{z}$.

Důkaz Označme $X = S_+/\sqrt{\gamma}$. Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2} \left\| T^{-1/2} (X - S) \right\|_F^2 + \tilde{u}^T \left(X^T y - \tilde{z} \right) + v^T \left(X\tilde{z} - \frac{\rho}{\gamma} d \right) \\ &= \sum_{i=1}^{m} \left[\frac{1}{2} (x_i - s_i)^T T^{-1} (x_i - s_i) + \tilde{u}_i y^T x_i + \tilde{z}_i v^T x_i \right] - \tilde{u}^T \tilde{z} - \frac{\rho}{\gamma} v^T d, \end{aligned}$$

kde $S = [s_1, \dots s_m]$ a $X = [x_1, \dots x_m]$. Derivováním Langrangeovy funkce dostaneme

$$\frac{\partial L}{\partial x_i} = T^{-1} (x_i - s_i) + \tilde{u}_i y + \tilde{z}_i v.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $T^{-1}(x_i - s_i) + \tilde{u}_i y + \tilde{z}_i v = 0$, $1 \le i \le m$, neboli

$$X - S = -Ty\tilde{u}^T - Tv\tilde{z}^T.$$

Použitím první podmínky z (363) dostaneme

$$X^T y = S^T y - y^T T y \tilde{u} - v^T T y \tilde{z} = \tilde{z} \quad \Rightarrow \quad \tilde{u} = \frac{1}{y^T T y} \left(S^T y - (1 + v^T T y) \tilde{z} \right),$$

což po dosazení do předchozí rovnosti dává

$$X - S = -\frac{Ty}{y^T T y} \tilde{y}^T + w\tilde{z}^T,$$

kde $w \in R^n$ je zatím neznámý vektor (jednoznačně určený vektorem v). Užitím druhé podmínky z (363) dostaneme

$$X\tilde{z} = S\tilde{z} - \frac{y^T S\tilde{z}}{y^T T y} Ty + \tilde{z}^T \tilde{z} w = \frac{\rho}{\gamma} d \quad \Rightarrow \quad w = \frac{1}{\tilde{z}^T \tilde{z}} \left(\frac{\rho}{\gamma} d - z + \frac{y^T z}{y^T T y} Ty \right),$$

což po dosazení do předchozí rovnosti (s využitím vztahu $X = S_+/\sqrt{\gamma}$) dává (364). Postačitelnost plyne z konvexity Frobeniovy normy. □

Poznámka 151. Zvolíme-li matici T tak, aby platilo $Ty = (\rho/\gamma)d - z$, výraz (364) se velmi zjednoduší. Po dosazení a úpravě dostaneme

$$\frac{1}{\sqrt{\gamma}} S_+ = S - \frac{(\rho/\gamma)d - z}{(\rho/\gamma)b - y^T z} (\tilde{y} - \tilde{z})^T \tag{365}$$

Položíme-li $\tilde{z} = \pm\sqrt{\rho b/(\gamma c)}\, \tilde{d}$, dostaneme metodu BFGS (vzorec (345)).

Nyní ukážeme, jak lze volit vektory Ty a $\tilde{z}$, abychom dostali jednotlivé metody z Broydenovy třídy. Za tímto účelem budeme předpokládat, že $Ty = Hv$, kde $v \in R^n$. V tomto případě lze vzorec (364) zapsat ve tvaru

$$\frac{1}{\sqrt{\gamma}} S_+ = S - \frac{Hv}{y^T Hv} y^T S + \left(\frac{\rho}{\gamma} d - z + \frac{y^T z}{y^T Hv} Hv \right) \frac{\tilde{z}^T}{\tilde{z}^T \tilde{z}}. \tag{366}$$

Věta 93. *Nechť H_+ je symetrická matice určená podle (286), kde $H = SS^T$, $d = -\alpha Hg$, $a > 0$, $b > 0$, $c > 0$ a $\eta \geq 0$ (takže $\delta > 0$). Nechť $B = H^\dagger = S(S^T S)^{-1} S^T$ a S_+ je matice určená podle (364) nebo (366), kde*

$$Ty = Hv = \frac{\sqrt{\eta}}{b} d + \frac{1 - \sqrt{\eta}}{a} Hy, \qquad \tilde{z} = \frac{\rho}{\gamma} \frac{b}{\sqrt{\delta}} S^T B Ty = \frac{\rho}{\gamma} \frac{b}{\sqrt{\delta}} S^T BHv \tag{367}$$

a kde $\sqrt{\eta}$ a $\sqrt{\delta}$ jsou libovolné hodnoty (kladné i záporné) takové, že $(\sqrt{\eta})^2 = \eta$ a $(\sqrt{\delta})^2 = \delta$ (δ je číslo definované vztahem (300)). Pak platí $H_+ = S_+ S_+^T$.

Důkaz (a) Položme $\tilde{z} = \vartheta S^T BHv$, kde hodnota ϑ se vybírá tak, aby platilo $\tilde{z}^T \tilde{z} = (\rho/\gamma) b$ (vztah (363)). Jelikož $\tilde{z}^T \tilde{z} = \vartheta^2 v^T Hv$ (neboť podle definice 37 platí $HBHBH = H$), je tato hodnota dána výrazem

$$\vartheta^2 = \frac{\rho}{\gamma} \frac{b}{v^T Hv}. \tag{368}$$

Speciální volbu $\tilde{z} = \vartheta S^T BHv$ používáme proto, že se tím velmi zjednoduší aktualizace (366), neboť v tomto případě platí $z = S\tilde{z} = \vartheta Hv$, takže

$$\frac{y^T z}{y^T Hv} Hv - z = \vartheta \frac{y^T Hv}{y^T Hv} Hv - \vartheta Hv = 0. \tag{369}$$

(b) Jelikož norma vektoru Hv neovlivní tvar aktualizace (366), budeme předpokládat, že $y^T Hv = 1$. Položíme-li

$$Hv = \frac{\alpha_1}{b} d + \frac{\alpha_2}{a} Hy \tag{370}$$

(což lze, neboť $d = -\alpha Hg$), pak z $y^T Hv = 1$ plyne $\alpha_1 + \alpha_2 = 1$. Dále platí

$$v^T Hv = v^T HBHv = \frac{\alpha_1^2}{b^2} c + 2 \frac{\alpha_1 \alpha_2}{ab} b + \frac{\alpha_2^2}{a^2} a = \frac{\alpha_1^2 (ac - b^2) + b^2}{ab^2}$$

(používáme vztah $\alpha_1 + \alpha_2 = 1$). Dosadíme-li tento výsledek do (368), dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma} \frac{b}{v^T Hv} = \frac{\rho}{\gamma} \frac{ab^3}{\alpha_1^2 (ac - b^2) + b^2} \tag{371}$$

(c) Nyní využijeme toho, že vektory $\tilde{z}$ a Hv, uvedené v (a) a (b), umožňují zapsat aktualizaci (366) ve velmi jednoduchém tvaru

$$\frac{1}{\sqrt{\gamma}} S_+ = S - Hvy^T S + \frac{\vartheta}{b} dv^T HBS \tag{372}$$

(používáme rovnost (369) a vztah $y^T Hv = 1$). Položíme-li $H = SS^T$, $H_+ = S_+ S_+^T$ a použijeme-li vztah (372), dostaneme po roznásobení

$$\begin{aligned} \frac{1}{\gamma} H_+ &= H - (Hvy^T H + Hyv^T H) + \frac{\vartheta}{b} (dv^T H + Hvd^T) + aHvv^T H - \frac{\vartheta}{b} (dv^T H + Hvd^T) + \frac{\vartheta^2}{b^2} v^T Hvdd^T \\ &= H - (Hvy^T H + Hyv^T H) + aHvv^T H + \frac{\vartheta^2}{b^2} v^T Hvdd^T. \end{aligned} \tag{373}$$

Jelikož vektor Hv je lineární kombinací vektorů d a Hy, můžeme tuto aktualizaci vyjádřit ve tvaru $(1/\gamma) H_+ = H + UMU^T$, kde použité matice mají stejný význam jako ve větě 78. K určení parametru η

stačí porovnat koeficienty u Hyy^TH v obou vyjádřeních. Podle věty 78 se tento koeficient rovná $(\eta-1)/a$ a dosazením vektoru $Hv = (\alpha_1/b)d + (\alpha_2/a)Hy$ do (373) dostaneme hodnotu $\alpha_2^2/a - 2\alpha_2/a$. Musí tedy platit

$$\frac{\alpha_2^2}{a} - 2\frac{\alpha_2}{a} = \frac{\eta-1}{a},$$

neboli $\alpha_1^2 = (1-\alpha_2)^2 = \alpha_2^2 - 2\alpha_2 + 1 = \eta$, což po dosazení do (370) dává první rovnost v (367). Dosadíme-li $\alpha_1^2 = \eta$ do (371) a použijeme-li výraz δ definovaný v poznámce 109, dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{ab^3}{\eta(ac-b^2)+b^2} = \left(\frac{\rho}{\gamma}\right)^2\frac{b^2}{\delta} \quad \Rightarrow \quad \vartheta = \frac{\rho}{\gamma}\frac{b}{\sqrt{\delta}}.$$

Dosadíme-li toto číslo do vztahu $\tilde{z} = \vartheta SBHv$, dostaneme druhou rovnost v (367). □

Důsledek 14. *Nechť jsou splněny předpoklady věty 93 a nechť*

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)\tilde{y}^T + \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}}d\left(\frac{\sqrt{\eta}}{b}\tilde{d} + \frac{1-\sqrt{\eta}}{a}\tilde{y}\right)^T \tag{374}$$

kde $\sqrt{\eta}$ a $\sqrt{\delta}$ jsou libovolné hodnoty (kladné i záporné) takové, že $(\sqrt{\eta})^2 = \eta$ a $(\sqrt{\delta})^2 = \delta$. Pak platí $H_+ = S_+S_+^T$.

Důkaz Dokazovaný vztah dostaneme prostým dosazením vektoru Hv a čísla ϑ, uvedených ve větě 93 a jejím důkazu, do vzorce (372) a použitím vztahů (329)–(330). □

Poznámka 152. Věta 93 používá jiné předpoklady než věta 85, nerovnost $\mu \geq 0$ je nahražena nerovností $\eta \geq 0$ (pak také $\delta \geq 0$). Vztah (374) lze tedy použít pro každou perfektní metodu z Broydenovy třídy. Na druhé straně matice $(1/\sqrt{\gamma})S_+ - S$ v (374) má obecně hodnost 2, takže (374) vyžaduje více numerických operací než (341). Pro metody DFP a BFGS dávají oba vzorce stejné výsledky, dosazení do (374) je však nesrovnatelně jednodušší. Dosadíme-li do (374) $\vartheta = 0$ a $\delta = \rho b/(\gamma a)$, dostaneme

$$\frac{1}{\sqrt{\gamma}}S_+^{DFP} = S - \frac{1}{a}SS^Ty\tilde{y}^T \pm \frac{1}{a}\sqrt{\frac{\rho a}{\gamma b}}d\tilde{y}^T,$$

což je (344). Dosadíme-li do (374) $\vartheta = 1$ a $\delta = \rho c/(\gamma b)$, dostaneme

$$\frac{1}{\sqrt{\gamma}}S_+^{BFGS} = S - \frac{1}{b}d\tilde{y}^T \pm \frac{1}{b}\sqrt{\frac{\rho b}{\gamma c}}d\tilde{d}^T,$$

což je (345). Je také zajímavé porovnat vztah (374) s pseudosoučinovým tvarem (291) uvedeným v poznámce 114.

4.4 Výběr parametrů (škálování a korekce)

Zatím jsme se zabývali různými vyjádřeními a základními vlastnostmi metod s proměnnou metrikou. Nyní je třeba ukázat, jak se volí podíl ρ/γ a parametr η. Vhodná volba podílu ρ/γ může mít vliv na asymptotickou rychlost konvergence diskutovanou v poznámce 41. Úvahy tohoto typu se poprvé objevily v práci [129], odkud pochází následující věta.

Věta 94. *Nechť $\tilde{G}$ je matice taková, že $\tilde{G}d = y$, tedy například*

$$\tilde{G} = \int_0^1 G(x+\lambda d)\mathrm{d}\lambda. \tag{375}$$

Označme $R = \tilde{G}^{-1/2}B\tilde{G}^{-1/2}$ a $R'_+ = \tilde{G}^{-1/2}B_+\tilde{G}^{-1/2}$. Pak jestliže $0 \leq \beta \leq 1$ a $b/a \leq \gamma/\rho \leq c/b$, platí $\kappa(R'_+) \leq \kappa(R)$.

Důkaz Označme $z = \tilde{G}^{1/2}d$, takže $y = \tilde{G}^{1/2}z$. Vynásobíme-li (306) zleva i zprava maticí $\tilde{G}^{-1/2}$, můžeme psát

$$\gamma R'_+ = R + \frac{\gamma}{\rho b} zz^T - \frac{1}{c} Rz(Rz)^T + \frac{\beta}{c}\left(\frac{c}{b}z - Rz\right)\left(\frac{c}{b}z - Rz\right)^T, \tag{376}$$

kde $b = z^T z$ a $c = z^T Rz$. Transformací kvazinewtonovské podmínky dostaneme $R'_+ z = (1/\rho)z$, takže matice R'_+ má vlastní číslo $1/\rho$ příslušné vlastnímu vektoru z. Vlastní vektory $v \in R^n$ příslušné vlastním číslům $\lambda \neq 1/\rho$ matice R'_+ můžeme volit tak, aby platilo $v^T z = 0$ a $v^T v = 1$. Potom z (376) plyne, že

$$\lambda = v^T R'_+ v = \frac{1}{\gamma} v^T Rv + \frac{1}{\gamma}\frac{\beta - 1}{c}(v^T Rz)^2 \leq \frac{1}{\gamma} v^T Rv \leq \frac{1}{\gamma}\|R\|$$

(neboť $\beta - 1 \leq 0$). Můžeme tedy psát $\|R'_+\| \leq \max(1/\rho, \|R\|/\gamma)$. Protože $c = z^T Rz$ a $b = z^T z$, platí $c/b = z^T Rz/z^T z \leq \|R\|$, takže pro $\gamma/\rho \leq c/b$ dostaneme $1/\rho \leq \|R\|/\gamma$. Platí tedy $\|R'_+\| \leq \|R\|/\gamma$. Nyní můžeme použít dualitu (poznámka 119) a provést stejnou úvahu pro matici $(R'_+)^{-1}$ (v tomto případě se používá nerovnost $\rho/\gamma \leq a/b$). Dostaneme tak $\|(R'_+)^{-1}\| \leq \gamma\|R^{-1}\|$. Spojením obou nerovností dostaneme dokazované tvrzení □

Poznámka 153. Pokud $\gamma = 1$, $\rho = 1$ a $0 \leq \beta \leq 1$, lze postupem použitým v důkazu věty 94 ukázat, že platí

$$\|R'_+\| \leq \max(1, \|R\|), \qquad \|(R'_+)^{-1}\| \leq \max(1, \|R^{-1}\|),$$

a označíme-li $\tilde{\kappa} = \max(1, \|R\|)\max(1, \|R^{-1}\|)$, $\tilde{\kappa}'_+ = \max(1, \|R'_+\|)\max(1, \|(R'_+)^{-1}\|)$, dostaneme $\tilde{\kappa}'_+ \leq \tilde{\kappa}$.

Podle věty 94 je vhodné volit podíl ρ/γ tak, aby byla splněna nerovnost $b/c \leq \rho/\gamma \leq a/b$. V tomto případě platí $\mu \geq 0$ pro libovolnou hodnotu parametru η (poznámka 140). Metoda hodnosti 1 však vyžaduje, aby $0 < \rho/\gamma < b/c$ nebo $a/b < \rho/\gamma$ (poznámka 117), neboť jinak není matice H_+ pozitivně definitní. Interval $0 < \rho/\gamma < b/c$ je nevhodný, neboť v tomto případě $\eta^{R1} < 0$. Zbývá tedy interval $a/b < \rho/\gamma$. Pak $\eta^{R1} > 1$ a metoda hodnosti 1 patří mezi perfektní metody s proměnnou metrikou. Bližší podrobnosti týkající se volby podílu ρ/γ jsou uvedeny v poznámce 156.

K volbě parametru η lze použít různé minimalizační principy. Nejvíce se ujal princip spočívající v minimalizaci čísla podmíněnosti matice $(1/\gamma)H^{-1/2}H_+H^{-1/2}$, použitý v pracech [36] a [130].

Lemma 44. *Nechť jsou splněny předpoklady lemmatu 32 a nechť vektory d a Hy jsou lineárně nezávislé (takže $ac - b^2 > 0$). Pak pro $\eta > \eta^*$ platí:*

(a) Kořeny $\underline{\lambda}(\eta) \leq \overline{\lambda}(\eta)$ kvadratické rovnice $\lambda^2 - \sigma\lambda + \delta = 0$ jsou rostoucími funkcemi parametru η.
(b) Podmínka $0 < \underline{\lambda}(\eta) \leq 1 \leq \overline{\lambda}(\eta)$ je splněna právě tehdy, když $\mu \geq 0$.
(c) Podíl $\overline{\lambda}(\eta)/\underline{\lambda}(\eta)$ nabývá svého minima právě tehdy, když

$$\eta = \eta^{OC} = \frac{bc(\rho/\gamma - b/c)}{ac - b^2} > \eta^*. \tag{377}$$

Důkaz (a) Podle lemmatu 32 jsou čísla $\underline{\lambda}(\eta)$ a $\overline{\lambda}(\eta)$ vlastními čísly matice

$$\frac{1}{\gamma}H^{-1/2}H_+H^{-1/2} = H^{-1/2}H_+^{DFP}H^{-1/2} + \frac{\eta}{a}H^{-1/2}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T H^{-1/2} \triangleq W + \eta ww^T$$

(použili jsme první rovnost v (293)). Nechť $\eta_2 > \eta_1$. Je-li $v_1 \in R^n$, $\|v_1\| = 1$, vlastním vektorem matice $W + \eta_1 ww^T$ příslušným vlastnímu číslu $\overline{\lambda}(\eta_1)$, platí

$$\overline{\lambda}(\eta_1) = v_1^T(W + \eta_1 ww^T)v_1 < v_1^T(W + \eta_2 ww^T)v_1 \leq \overline{\lambda}(\eta_2).$$

Je-li $v_2 \in R^n$, $\|v_2\| = 1$, vlastním vektorem matice $W + \eta_2 ww^T$ příslušným vlastnímu číslu $\underline{\lambda}(\eta_2)$, platí

$$\underline{\lambda}(\eta_2) = v_2^T(W + \eta_2 ww^T)v_2 > v_2^T(W + \eta_1 ww^T)v_2 \geq \overline{\lambda}(\eta_1).$$

(b) Podle (271) platí

$$(1/\gamma)H^{-1/2}H_+H^{-1/2} = I + H^{-1/2}UMU^TH^{-1/2},$$

takže podmínka $0 < \underline{\lambda}(\eta) \leq 1 \leq \overline{\lambda}(\eta)$ je splněna právě tehdy, leží-li nula mezi nejmenším a největším vlastním číslem matice $H^{-1/2}UMU^TH^{-1/2}$, což nastává právě tehdy, platí-li $\det M = -\mu \leq 0$.

(c) Poznamenejme, že diskriminant kvadratické rovnice $\lambda^2 - \sigma\lambda + \delta = 0$ je kladný, neboť s použitím (305) dostaneme

$$\sigma^2 - 4\delta = \left(\frac{\gamma a}{\rho b}\delta + \frac{\rho c}{\gamma b}\right)^2 - 4\delta = \left(\frac{\gamma a}{\rho b}\delta - \frac{\rho c}{\gamma b}\right)^2 + 4\frac{ac-b^2}{b^2}\delta > 0$$

(podle předpokladu platí $ac - b^2 > 0$ a $\eta > \eta^* \Rightarrow \delta > 0$). Kořeny kvadratické rovnice $\lambda^2 - \sigma\lambda + \delta = 0$ lze zapsat ve tvaru

$$\underline{\lambda} = \frac{\sigma}{2} - \sqrt{\left(\frac{\sigma}{2}\right)^2 - \delta}, \qquad \overline{\lambda} = \frac{\sigma}{2} + \sqrt{\left(\frac{\sigma}{2}\right)^2 - \delta}.$$

Použijeme-li substituci

$$\omega = \frac{\sigma}{2\sqrt{\delta}},$$

dostaneme po rozšíření zlomku

$$\frac{\overline{\lambda}}{\underline{\lambda}} = \frac{\omega + \sqrt{\omega^2 - 1}}{\omega - \sqrt{\omega^2 - 1}} = \left(\omega + \sqrt{\omega^2 - 1}\right)^2,$$

takže podíl $\overline{\lambda}/\underline{\lambda}$ je minimální, je-li číslo ω minimální. Minimum tohoto čísla najdeme řešením rovnice

$$\omega' = \frac{2\sigma'\delta - \sigma\delta'}{4\delta\sqrt{\delta}} = 0,$$

neboli $2\sigma'\delta - \sigma\delta' = 0$ (neboť $\delta > 0$). Použijeme-li výrazy uvedené v lemmatu 32, dostaneme

$$\begin{aligned} 2\sigma'\delta - \sigma\delta' &= \frac{\rho}{\gamma}\frac{ac-b^2}{ab^3}\left(\eta(ac-b^2)+b^2\right) - \left(\frac{\rho}{\gamma}\right)^2\frac{ac-b^2}{ab^3}bc \\ &= \frac{\rho}{\gamma}\frac{(ac-b^2)^2}{ab^3}\left(\eta - \frac{bc(\rho/\gamma - b/c)}{ac-b^2}\right), \end{aligned}$$

odkud plyne dokazované tvrzení. □

Věta 95. *Nechť jsou splněny předpoklady lemmatu 44. Označme $\kappa(\eta)$ spektrální číslo podmíněnosti matice $(1/\gamma)H^{-1/2}H_+H^{-1/2}$. Pak:*

(a) Pokud $0 < \rho/\gamma < b/c$, je $\kappa(\eta)$ minimální právě tehdy, když $\eta = \max(\eta^{R1}, \eta^{OC})$.
(b) Pokud $b/c \leq \rho/\gamma \leq a/b$, je $\kappa(\eta)$ minimální právě tehdy, když $\eta = \eta^{OC}$.
(c) Pokud $a/b < \rho/\gamma$, je $\kappa(\eta)$ minimální právě tehdy, když $\eta = \min(\eta^{R1}, \eta^{OC})$.

Důkaz Podle Lemmatu 32 platí

$$\kappa(\eta) = \frac{\max(1, \overline{\lambda}(\eta))}{\min(1, \underline{\lambda}(\eta))}. \tag{378}$$

Jestliže $\mu \geq 0$, podle (b) lemmatu 44 platí $0 < \underline{\lambda}(\eta) \leq 1 \leq \overline{\lambda}(\eta)$, takže podle (c) lemmatu 44 je $\kappa(\eta)$ minimální, pokud $\eta = \eta^{OC}$. Jestliže $\mu < 0$, lze podle (a) lemmatu 44 oba kořeny rovnice $\lambda^2 + \sigma\lambda + \delta = 0$ současně zvětšit nebo zmenšit změnou parametru η. Podíl (378) je pak minimální, pokud $\overline{\lambda}(\eta) = 1$ nebo $\underline{\lambda}(\eta) = 1$, neboli $\mu = 0$, což odpovídá metodě hodnosti 1. Zbytek tvrzení pak plyne z poznámky 117 a poznámky 140. □

Poznámka 154. Hodnota (377) je samoduální. Dosadíme-li ji do (307), dostaneme

$$\beta = \beta^{OC} = \frac{ab(\gamma/\rho - b/a)}{ac - b^2} > \beta^*. \tag{379}$$

Označme $\kappa(\beta)$ spektrální číslo podmíněnosti matice $\gamma B^{-1/2} B_+ B^{-1/2}$. Pak:

(a) Pokud $0 < \gamma/\rho < b/a$, je $\kappa(\beta)$ minimální právě tehdy, když $\beta = \max(\beta^{R1}, \beta^{OC})$.
(b) Pokud $b/a \leq \gamma/\rho \leq c/b$, je $\kappa(\beta)$ minimální právě tehdy, když $\beta = \beta^{OC}$.
(c) Pokud $c/b < \gamma/\rho$, je $\kappa(\beta)$ minimální právě tehdy, když $\beta = \min(\beta^{R1}, \beta^{OC})$.

Poznámka 155. Položíme-li $\delta = 1$, dostaneme použitím (300) hodnotu

$$\eta = \eta^{OD} = \frac{ab(\gamma/\rho - b/a)}{ac - b^2} > \eta^*, \tag{380}$$

která je shodná s výrazem vystupujícím ve vzorci (379). Tato hodnota je samoduální. Dosadíme-li ji do (307), dostaneme

$$\beta = \beta^{OD} = \frac{bc(\rho/\gamma - b/c)}{ac - b^2} > \beta^*, \tag{381}$$

což je zase výraz vystupující ve vzorci (377).

Poznámka 156. Větu 95 lze použít k volbě parametru η. Jestliže $b/c \leq \rho/\gamma \leq a/b$, je vhodné použít hodnotu $\eta = \eta^{OC}$. Mnohem praktičtější aplikací věty 95 je však určení vhodného podílu ρ/γ pro danou hodnotu parametru η. V tomto případě z $\eta = \eta^{OC}$ plyne

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{b}{c}\left(1 - \frac{\eta}{\eta^*}\right). \tag{382}$$

Řešením rovnice $2\sigma'\delta - \sigma\delta' = 0$, kde tentokrát derivujeme podle podílu ρ/γ, se lze přesvědčit, že tato hodnota minimalizuje podíl $\overline{\lambda}/\underline{\lambda}$ pro zadanou hodnotu parametru η. Vyšetříme nyní některé konkrétní metody. Budeme přitom používat vzorce z tabulky uvedené na konci oddílu 4.1

(1) Pro metodu DFP platí $\eta = 0$, takže je vhodné volit $\rho/\gamma = b/c$. Pro tuto hodnotu dostaneme $\delta = b^2/(ac) > 0$.
(2) Pro metodu BFGS platí $\eta = 1$, takže je vhodné volit $\rho/\gamma = a/b$. Pro tuto hodnotu dostaneme $\delta = ac/b^2 \geq 1$.
(3) Pro Hoshinovu metodu platí $\eta = (\rho/\gamma)/(\rho/\gamma + a/b)$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{(\rho/\gamma)(ac - b^2) + b^2(\rho/\gamma + a/b)}{(\rho/\gamma + a/b)bc} = \frac{a}{b}\frac{\rho/\gamma + b/c}{\rho/\gamma + a/b},$$

což je kvadratická rovnice, která má kladný kořen $\rho/\gamma = \sqrt{a/c}$. Pro tuto hodnotu dostaneme

$$\eta = \frac{\sqrt{a/c}}{\sqrt{a/c} + a/b} = \frac{b}{b + \sqrt{ac}}$$

$$\delta = \sqrt{a/c}\,\frac{\sqrt{a/c}\,(c/b) + 1}{\sqrt{a/c} + a/b} = \frac{a/b + \sqrt{a/c}}{\sqrt{a/c} + a/b} = 1$$

(4) Pro metodu R1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{(\rho/\gamma)(ac - b^2) + b^2(\rho/\gamma - a/b)}{(\rho/\gamma - a/b)bc} = \frac{a}{b}\frac{\rho/\gamma - b/c}{\rho/\gamma - a/b},$$

což je kvadratická rovnice, která má dva kladné kořeny

$$\frac{\rho}{\gamma} = \frac{a}{b}\,(1+\lambda)\,, \qquad \lambda = \pm\sqrt{1-b^2/(ac)}.$$

Pro tyto hodnoty dostaneme

$$\eta = \frac{(a/b)(1+\lambda)}{(a/b)(1+\lambda)-a/b} = \frac{1+\lambda}{\lambda} = 1+\frac{1}{\lambda} = 1 \pm \frac{1}{\sqrt{1-b^2/(ac)}}.$$

Menší z těchto hodnot je záporná a tudíž nevhodná. Dosadíme-li získané hodnoty ρ/γ a η do (284), dostaneme

$$\begin{aligned}\delta &= \frac{1+\lambda}{b^2}\left(\frac{1+\lambda}{\lambda}(ac-b^2)+b^2\right) = (1+\lambda)\frac{ac(1+\lambda)-b^2}{\lambda\, b^2} \geq (1-|\lambda|)\left(\frac{ac}{b^2}-\frac{ac-b^2}{|\lambda|\, b^2}\right)\\ &= \frac{\sqrt{ac}-\sqrt{ac-b^2}}{\sqrt{ac}}\left(\frac{ac}{b^2}-\frac{\sqrt{ac}\sqrt{ac-b^2}}{b^2}\right) = \frac{(\sqrt{ac}-\sqrt{ac-b^2})^2}{b^2} \geq \frac{b^2}{4ac} > 0,\end{aligned}$$

neboť $|\lambda| = \sqrt{1-b^2/(ac)}$, $0 \leq |\lambda| < 1$, $ac - b^2 \geq 0$, $b > 0$ a $\sqrt{ac-b^2} \leq \sqrt{ac} - b^2/(2\sqrt{ac})$. Metodu hodnosti 1 nelze škálovat pomocí standardních hodnot ρ/γ. Položíme-li $\rho/\gamma = a/b$, dostaneme $\eta = \infty$ a $\delta = \infty$. Položíme-li $\rho/\gamma = b/c$, dostaneme $\eta = \eta^*$ a $\delta = 0$. Položíme-li $\rho/\gamma = \sqrt{a/c}$, dostaneme

$$\eta = \frac{\sqrt{a/c}}{\sqrt{a/c}-a/b} = \frac{b}{b-\sqrt{ac}} < 0,$$

$$\delta = \sqrt{a/c}\,\frac{\sqrt{a/c}\,(c/b)-1}{\sqrt{a/c}-a/b} = \frac{a/b-\sqrt{a/c}}{\sqrt{a/c}-a/b} = -1 < 0.$$

(5) Pro metodu OC podle vzorce (377) platí $\eta = bc(\rho/\gamma - b/c)/(ac-b^2)$, což dává

$$\frac{\rho}{\gamma} = \frac{\eta(ac-b^2)+b^2}{bc} = \frac{bc(\rho/\gamma-b/c)+b^2}{bc} = \frac{\rho}{\gamma}-\frac{b}{c}+\frac{b}{c} = \frac{\rho}{\gamma},$$

takže metoda OC je optimální pro jakýkoliv podíl ρ/γ splňující podmínku $b/c \leq \rho/\gamma \leq a/b$. Lze tedy použít standardní hodnoty $\rho/\gamma = a/b$, $\rho/\gamma = b/c$ a $\rho/\gamma = \sqrt{a/c}$ (pak dostabeme metody BFGS, DFP a Hoshinovu metodu). Zřejmě $\delta = (\rho/\gamma)^2 c/a > 0$.

(6) Pro metodu OD podle vzorce (380) platí $\eta = ab(\gamma/\rho - b/a)/(ac-b^2)$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\eta(ac-b^2)+b^2}{bc} = \frac{ab(\gamma/\rho-b/a)+b^2}{bc} = \frac{a}{c}\frac{\gamma}{\rho},$$

což je kvadratická rovnice, která má kladný kořen $\rho/\gamma = \sqrt{a/c}$. Pro tuto hodnotu platí

$$\eta = \frac{ab(\sqrt{c/a}-b/a)}{ac-b^2} = \frac{b(\sqrt{ac}-b)}{(\sqrt{ac}-b)(\sqrt{ac}+b)} = \frac{b}{b+\sqrt{ac}}.$$

Optimálně škálovaná metoda OD je tedy totožná s optimálně škálovanou Hoshinovou metodou. Obě používají stejný podíl $\rho/\gamma = \sqrt{a/c}$ a platí pro ně $\delta = 1$.

(7) Pro metodu VD podle vzorce (391) platí $\eta = ((\rho/\gamma)bc+b^2)/((\rho/\gamma)bc+b^2-ac)$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\eta(ac-b^2)+b^2}{bc} = \frac{((\rho/\gamma)bc+b^2)(ac-b^2)+b^2((\rho/\gamma)bc+b^2-ac)}{((\rho/\gamma)bc+b^2-ac)\,bc} = \frac{(\rho/\gamma)ac}{(\rho/\gamma)bc+b^2-ac},$$

což je (po vykrácení) lineární rovnice, která má řešení $\rho/\gamma = (2ac-b^2)/(bc)$. Pro tuto hodnotu dostaneme

$$\eta = \frac{(2ac-b^2)+b^2}{(2ac-b^2)+b^2-ac} = 2$$

$$\delta = \frac{\rho}{\gamma}\frac{\eta(ac-b^2)+b^2}{ab} = \frac{2ac-b^2}{bc}\frac{2ac-b^2}{ab} = \frac{(2ac-b^2)^2}{ab^2c} \geq \frac{ac}{b^2} \geq 1.$$

(8) Pro metodu VL podle vzorce (395) platí $\eta = (\lambda ac - b^2)/(ac - b^2)$, kde $\lambda = \sqrt{c/a}$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\lambda ac - b^2 + b^2}{bc} = \lambda \frac{a}{b} = \sqrt{\frac{ac}{b^2}}.$$

V tomto případě platí

$$\delta = \frac{\rho}{\gamma} \frac{\eta(ac - b^2) + b^2}{ab} = \frac{\rho}{\gamma} \frac{\lambda ac - b^2 + b^2}{ab} = \lambda \frac{\rho}{\gamma} \frac{a}{b} = \frac{ac}{b^2} \geq 1.$$

Pokud $\eta > 1$, což nastává u metod (7)–(8), platí podle věty 95 $\rho/\gamma > a/b$. Použití této hodnoty však není vhodné, neboť v tomto případě nejsou splněny předpoklady věty 94. Lepší praktické výsledky dává hodnota $\rho/\gamma = a/b$ (nejbližší možná hodnota z intervalu $b/c \leq \rho/\gamma \leq a/b$). Poznamenejme, že pro metodu s $\eta > 1$ a s optimální volbou podílu ρ/γ platí $\mu \geq 0$ právě tehdy, když $\eta^2 - 2\eta + \eta^* \leq 0$ (přesvědčíme se o tom dosazením optimální hodnoty podílu ρ/γ do výrazu pro μ).

Poznámka 157. Jak je ukázáno v předchozí posnámce, lze větu 95 použít k získání některých dalších metod s proměnnou metrikou. Dosadíme-li optimální hodnotu podílu ρ/γ do vztahu určujícího parametr η, dostaneme výraz, který již neobsahuje tento podíl a definuje (pro neoptimální hodnotu ρ/γ) novou metodu z Broydenovy třídy.

(a) Dosadíme-li hodnotu $\rho/\gamma = \sqrt{a/c}$ do vztahu $\eta = (\rho/\gamma)/(\rho/\gamma + a/b)$ (Hoshinova metoda), dostaneme

$$\eta = \eta^{OS} = \frac{b}{b + \sqrt{ac}}. \tag{383}$$

Tato metoda, uvedená v [130]), patří mezi omezené metody s proměnnou metrikou.

(b) Dosadíme-li hodnotu $\rho/\gamma = (a/b)(1 + \sqrt{1 - b^2/(ac)})$ do vztahu $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$ (metoda hodnosti 1), dostaneme

$$\eta = \eta^{RS} = 1 + \frac{1}{\sqrt{1 - b^2/(ac)}}. \tag{384}$$

Tato metoda, uvedená v [97], patří mezi perfektní metody s proměnnou, metrikou ale není omezená (platí $\eta > 1$).

(c) Dosadíme-li hodnotu $\rho/\gamma = (2ac - b^2)/(bc)$ do vztahu $\eta = ((\rho/\gamma)bc + b^2)/((\rho/\gamma)bc + b^2 - ac)$ (metoda VD), dostaneme

$$\eta = \eta^{VS} = 2. \tag{385}$$

Tato metoda, uvedená v [98], patří mezi perfektní metody s proměnnou metrikou, ale není omezená (platí $\eta > 1$).

Existují další minimalizační postupy, kterými lze získat hodnotu parametru η a škálovací podíl ρ/γ. Jeden z nich je založen na minimalizaci normy $\|H^{1/2}(\gamma B_+ - B)H^{1/2}\|_F = \|\gamma H^{1/2} B_+ H^{1/2} - I\|_F$. Jelikož matice $H^{1/2} B_+ H^{1/2}$ a $B_+^{1/2} H B_+^{1/2}$ mají stejná vlastní čísla, platí

$$\begin{aligned} \|H^{1/2}(\gamma B_+ - B)H^{1/2}\|_F &= \|\gamma H^{1/2} B_+ H^{1/2} - I\|_F = \|\gamma B_+^{1/2} H B_+^{1/2} - I\|_F \\ &= \|B_+^{1/2}(H_+ - \gamma H)B_+^{1/2}\|_F \approx \|\tilde{G}^{1/2}(H_+ - \gamma H)\tilde{G}^{1/2}\|_F, \end{aligned}$$

což zdůvodňuje použití zvolené normy.

Lemma 45. *Nechť jsou splněny předpoklady lemmatu 32. Pak platí*

$$\|H^{1/2}(\gamma B_+ - B)H^{1/2}\|_F^2 = \left(\frac{\gamma a}{\rho b}\right)^2 - 2\frac{\gamma b}{\rho c} + 1 + \left(\frac{ac}{b^2} - 1\right)^2 \left(\beta^2 + 2\beta \frac{\gamma b}{\rho c}\right) \tag{386}$$

Důkaz Použijeme-li (306), dostaneme

$$H^{1/2}(\gamma B_+ - B)H^{1/2} = \frac{\gamma}{\rho b}H^{1/2}y(H^{1/2}y)^T - \frac{1}{c}B^{1/2}d(B^{1/2}d)^T + \frac{\beta}{c}\left(\frac{c}{b}H^{1/2}y - B^{1/2}d\right)\left(\frac{c}{b}H^{1/2}y - B^{1/2}d\right)^T.$$

Podle důsledku 10 a poznámky 107 platí

$$\begin{aligned}\|H^{1/2}(\gamma B^+ - B)H^{1/2}\|_F^2 &= \left(\frac{\gamma}{\rho b}\right)^2 a^2 + \frac{1}{c^2}c^2 + \frac{\beta^2}{c^2}\left(\frac{c^2}{b^2}a - 2\frac{c}{b}b + c\right)^2 \\ &- 2\frac{\gamma}{\rho b}\frac{1}{c}b^2 + 2(\gamma/\rho b)\frac{\beta}{c}\left(\frac{c/b}{a} - b\right)^2 - 2\frac{1}{c}\frac{\beta}{c}\left(\frac{c}{b}b - c\right)^2 \\ &= \left(\frac{\gamma a}{\rho b}\right)^2 + 1 + \beta^2\left(\frac{ac}{b^2} - 1\right)^2 - 2\frac{\gamma b}{\rho c} + 2\beta\frac{\gamma b}{\rho c}\left(\frac{ac}{b^2} - 1\right)^2,\end{aligned}$$

což dává (386). □

Věta 96. *Nechť jsou splněny předpoklady lemmatu 44. Pak norma $\|H^{1/2}(\gamma B^+ - B)H^{1/2}\|_F$ nabývá svého minima, pokud*

$$\beta = -\frac{b^2}{2ac - b^2} \quad \Rightarrow \quad \eta = 2, \tag{387}$$

$$\frac{\gamma}{\rho} = \frac{bc}{2ac - b^2} \quad \Rightarrow \quad \frac{\rho}{\gamma} = \frac{2ac - b^2}{bc}. \tag{388}$$

Důkaz Výraz (386) je poměrně komplikovaný. Abychom zjednodušili zápis zavedeme označení $\hat{a} = a/b$, $\hat{c} = c/b$ a $\hat{\gamma} = \gamma/\rho$. Dále označíme $\Delta(\beta, \hat{\gamma}) = \|H^{1/2}(\gamma B^+ - B)H^{1/2}\|_F$. Pak lze rovnost (386) zapsat ve tvaru

$$\Delta^2(\beta, \hat{\gamma}) = (\hat{\gamma}\hat{a})^2 - 2\frac{\hat{\gamma}}{\hat{c}} + 1 + (\hat{a}\hat{c} - 1)^2\left(\beta^2 + 2\beta\frac{\hat{\gamma}}{\hat{c}}\right).$$

Derivováním dostaneme

$$\frac{\partial\Delta^2(\beta, \hat{\gamma})}{\partial\beta} = 2(\hat{a}\hat{c} - 1)^2\left(\beta + \frac{\hat{\gamma}}{\hat{c}}\right), \tag{389}$$

$$\frac{\partial\Delta^2(\beta, \hat{\gamma})}{\partial\hat{\gamma}} = 2\hat{a}^2\hat{\gamma} - \frac{2}{\hat{c}}\left(1 - \beta(\hat{a}\hat{c} - 1)^2\right), \tag{390}$$

takže nutné podmínky pro extrém (věta 3) mají tvar

$$\beta = -\frac{\hat{\gamma}}{\hat{c}}, \qquad \hat{\gamma} = \frac{1}{\hat{a}^2\hat{c}}\left(1 - \beta(\hat{a}\hat{c} - 1)^2\right) = \frac{1}{\hat{a}^2\hat{c}}\left(1 + \frac{\hat{\gamma}}{\hat{c}}(\hat{a}\hat{c} - 1)^2\right).$$

Poslední rovnost lze zapsat ve tvaru

$$\hat{\gamma}\left(1 - \frac{(\hat{a}\hat{c} - 1)^2}{\hat{a}^2\hat{c}^2}\right) = \frac{1}{\hat{a}^2\hat{c}} \quad \Rightarrow \quad \hat{\gamma}\frac{2\hat{a}\hat{c} - 1}{\hat{a}^2\hat{c}^2} = \frac{1}{\hat{a}^2\hat{c}},$$

což po úpravě dává

$$\hat{\gamma} = \frac{\hat{c}}{2\hat{a}\hat{c} - 1}, \qquad \beta = -\frac{\hat{\gamma}}{\hat{c}} = -\frac{1}{2\hat{a}\hat{c} - 1}.$$

Dosadíme-li za $\hat{a}$, $\hat{c}$ a $\hat{\gamma}$ příslušné podíly, dostaneme levé vztahy v (387) a (388). Levý vztah v (387) dostaneme též dosazením $\eta = 2$ do (316), což potvrzuje správnost pravého vztahu v (387). Abychom

ukázali, že nalezený stacionární bod je lokálním minimem funkce $\Delta^2(\beta,\hat\gamma)$ a tedy i funkce $\Delta(\beta,\hat\gamma)$, vypočteme prvky Hessovy matice a její determinant. Platí

$$\frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\beta^2}=2(\hat a\hat c-1)^2>0,\qquad \frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\beta\partial\hat\gamma}=\frac{2}{\hat c}(\hat a\hat c-1)^2>0,\qquad \frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\gamma^2}=2\hat a^2>0,$$

$$\frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\beta^2}\frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\hat\gamma^2}-\left(\frac{\partial^2\Delta^2(\beta,\hat\gamma)}{\partial\beta\partial\hat\gamma}\right)^2=4\hat a^2(\hat a\hat c-1)^2-\frac{4}{\hat c^2}(\hat a\hat c-1)^4=4\frac{(\hat a\hat c-1)^2}{\hat c^2}(2\hat a\hat c-1)>0,$$

neboť podle předpokladu platí $\hat a>0$, $\hat c>0$ a $\hat a\hat c-1>0$. Hessova matice je tedy pozitivně definitní, takže nalezený stacionární bod je izolovaným lokálním minimem. □

Poznámka 158. Je-li zadán podíl γ/ρ, je funkce $\Delta^2(\beta,\hat\gamma)$ minimální, pokud

$$\beta=\beta^{VD}=-\frac{\gamma b}{\rho c}\quad\Rightarrow\quad \eta=\eta^{VD}=\frac{(\rho/\gamma)bc+b^2}{(\rho/\gamma)bc+b^2-ac},\tag{391}$$

přičemž $\beta>\beta^*$, pokud $\gamma/\rho<bc/(ac-b^2)$ (vztah pro η plyne z (315)). Například pro $\gamma/\rho=b/a$ dostaneme $\beta=-b^2/(ac)$, čemuž odpovídá $\eta=(ac+b^2)/b^2$. Je-li zadán parametr β, je funkce $\Delta^2(\beta,\hat\gamma)$ minimální, pokud

$$\frac{\gamma}{\rho}=\frac{b^4-\beta(ac-b^2)^2}{a^2bc},$$

přičemž $\gamma/\rho>0$, pokud $\beta<b^2/(ac-b^2)$. Například pro $\beta=0$ dostaneme $\gamma/\rho=b^3/(a^2c)$, neboli $\rho/\gamma=a^2c/b^3$.

Poznámka 159. Ze vztahů (382) a (388) plyne, že pro hodnotu $\eta=\eta^{VS}=2$ podíl $\rho/\gamma=(2ac-b^2)/(bc)$ minimalizuje nejen číslo podmíněnosti matice $(1/\gamma)H^{-1/2}H_+H^{-1/2}$ ale i normu $\|H^{1/2}(\gamma B_+-B)H^{1/2}\|_F$.

Další postup je založen na použití nerovnosti mezi geometrickým a aritmetickým průměrem

$$\frac{1}{n}\mathrm{Tr}(X)=\frac{1}{n}\sum_{i=1}^n\lambda_i\ge\left(\prod_{i=1}^n\lambda_i\right)^{1/n}=(\det(X))^{1/n}$$

(lemma 2), která platí pro libovolnou symetrickou pozitivně definitní matici X s vlastními čísly $\lambda_i>0$, $1\le i\le n$. Protože rovnost nastává pouze tehdy, mají-li všechna vlastní čísla stejnou hodnotu, je účelné minimalizovat funkci

$$\frac{\mathrm{Tr}(\gamma B^{-1/2}B_+B^{-1/2})}{n(\det(\gamma B^{-1/2}B_+B^{-1/2}))^{1/n}}=\frac{n-2+\sigma/\delta}{n\,(1/\delta)^{1/n}},\tag{392}$$

kde σ a δ jsou čísla určená podle vzorců (299) a (300).

Věta 97. *Hodnota β minimalizuje funkci (392) právě tehdy, platí-li*

$$\beta=\beta^*\,\frac{1-(\gamma/\rho)(a/b)}{n-1}$$

Důkaz Protože platí

$$\left(\frac{n-2+\sigma/\delta}{n\,(1/\delta)^{1/n}}\right)'=\frac{(\sigma/\delta)'n(1/\delta)^{1/n}-(1/\delta)^{1/n}(1/\delta)^{-1}(1/\delta)'(n-2+\sigma/\delta)}{n^2(1/\delta)^{2/n}},$$

je tato derivace nulová právě tehdy, když $n(\sigma/\delta)'(1/\delta)-(1/\delta)'(n-2+\sigma/\delta)=0$. Použijeme-li hodnoty (319) a (320), dostaneme

$$n\left(1-\frac{\beta}{\beta^*}\right)=n-2+\frac{\gamma a}{\rho b}+1-\frac{\beta}{\beta^*}$$

(neboť $(1/\delta)' = (\gamma/\rho)(b/c)(\sigma/\delta)'$), takže

$$1 - \frac{\beta}{\beta^*} = \frac{n-2+(\gamma/\rho)(a/b)}{n-1} = 1 + \frac{(\gamma/\rho)(a/b)-1}{n-1}, \tag{393}$$

odkud plyne tvrzení věty. □

Poznámka 160. Hodnota β uvedená ve větě 97 je obvykle velmi malá v absolutní hodnotě a příslušná metoda se tedy chová jako metoda BFGS a často jí i předčí. Odpovídající hodnotu η pak určíme ze vztahů (314) a (393) (je však vhodné nahradit tuto hodnotu nulou, vyjde-li záporná).

Odvodíme ještě jednu hodnotu parametru β, která definuje velmi efektivní metodu, překonávající všechny zatím popsané metody s proměnnou metrikou. Použijeme přitom funkci $\psi(X) = \mathrm{Tr}X - \ln\det X$. Kdybychom zvolili $X = \gamma B^{-1/2} B_+ B^{-1/2}$, dostali bychom metodu BFGS (poznámka 150). Volba $X = \gamma B_+$ je nevhodná neboť získaná matice může být příliš vzdálená od B. Použijeme tedy matici $X = \gamma G^{-1/2} B_+ G^{-1/2}$, kde G je nějaká aproximace Hessovy matice minimalizované funkce. Tato myšlenka byla použita v práci [17].

Lemma 46. *Hodnota β minimalizuje funkci $\psi(\gamma G^{-1/2} B_+ G^{-1/2})$ právě tehdy když*

$$\beta = \beta^* + \frac{c}{v^T G^{-1} v},$$

kde $v = (c/b)y - Bd$.

Důkaz Použijeme-li (306), můžeme psát

$$\gamma G^{-1/2} B_+ G^{-1/2} = \tilde{B} + \frac{\gamma}{\rho}\frac{1}{b}\tilde{y}\tilde{y}^T - \frac{1}{c}\tilde{B}\tilde{d}(\tilde{B}\tilde{d})^T + \frac{\beta}{c}\tilde{v}\tilde{v}^T,$$

kde $\tilde{B} = G^{-1/2} B G^{-1/2}$, $\tilde{d} = G^{1/2} d$, $\tilde{y} = G^{-1/2} y$ a $\tilde{v} = G^{-1/2} v$, což spolu s (312) dává

$$\mathrm{Tr}(\gamma G^{-1/2} B_+ G^{-1/2}) = C_1 + \frac{\beta}{c}\tilde{v}^T\tilde{v} = C_1 + \frac{\beta}{c} v^T G^{-1} v,$$

$$\ln\det(\gamma G^{-1/2} B_+ G^{-1/2}) = C_2 + \ln\left(1 - \frac{\beta}{\beta^*}\right),$$

kde C_1 a C_2 jsou nějaké konstanty. Podmínka pro extrém funkce $\psi(\gamma G^{-1/2} B_+ G^{-1/2})$ má tedy tvar

$$\psi'(\gamma G^{-1/2} B_+ G^{-1/2}) = \frac{1}{c} v^T G^{-1} v + \frac{1}{\beta^* - \beta} = 0,$$

což dává tvzení lemmatu. □

Jelikož neznáme inverzi Hessovy matice, musíme použít nějakou její aproximaci. Jednou z možností je položit $G^{-1} = \lambda H$, kde λ je zatím nespecifikovaný parametr [99].

Lemma 47. *Zvolíme-li $G^{-1} = \lambda H$, dostaneme*

$$\beta = \beta^{VL} = \frac{b^2}{ac - b^2}\left(\frac{1}{\lambda} - 1\right), \tag{394}$$

$$\eta = \eta^{VL} = \frac{\lambda ac - b^2}{ac - b^2} \tag{395}$$

Důkaz Zvolíme-li $G^{-1} = \lambda H$, můžeme psát

$$\frac{v^T G^{-1} v}{c} = \frac{\lambda}{c}((c/b)y - Bd)^T H((c/b)y - Bd) = \frac{\lambda}{c}\left(\frac{c^2}{b^2}a - 2\frac{c}{b}b + c\right) = \lambda\frac{ac - b^2}{b^2}. \tag{396}$$

Použijeme-li lemma 46 a vzorec (396), dostaneme $\beta^* - \beta = \beta^*/\lambda$, neboli $1 - \beta/\beta^* = 1/\lambda$, což spolu s (313) a (315) dává (394) a (395). □

Poznámka 161. Dobré výsledky dostaneme, položíme-li $\lambda = \sqrt{c/a}$. Poněkud horší výsledky dávají volby $\lambda = b/a$ a $\lambda = c/b$. Hodnota (395) je definována pokud $ac - b^2 > 0$, v opačném případě pokládáme $\eta = 1$. Pokud je hodnota (395) záporná, pokládáme $\eta = 0$.

Zatím jsme popsali, jak lze volit parametr η a podíl ρ/γ. Nyní ukážeme jak se určují parametry ρ a γ. Parametr ρ slouží ke korekci kvadratického modelu minimalizované funkce, který odpovídá kvazinewtonovské podmínce $B_+ d = y$. V této podmínce vystupují pouze gradienty g_+ a g. Korekce kvadratického modelu je založena na dodatečném použití funkčních hodnot F_+ a F. Postupuje se tak, že se pomocí funkčních hodnot a gradientů určí aproximace čísla $d^T B_+ d$ a z modifikované kvazinewtonovské podmínky $B_+ d = (1/\rho)y$ (která byla zavedena v práci [84]) se vypočte hodnota

$$\rho = \frac{d^T y}{d^T B_+ d}. \tag{397}$$

Hodnota (397) se používá pouze tehdy, když $\underline{\rho} \le \rho \le \overline{\rho}$ (obvykle $\underline{\rho} = 0.01$ a $\overline{\rho} = 100$). V opačném případě se volí hodnota $\rho = 1$. Při výpočtu čísla $d^T B_+ d$ se používají výrazy

$$A = \frac{F_+ - F}{d^T g}, \qquad B = \frac{d^T g_+}{d^T g}.$$

Pro metody spádových směrů splňující slabou Wolfeho podmínku platí $d^T g < 0$, $F_+ - F \le \varepsilon_1 d^T g$, a $d^T g_+ \ge \varepsilon_2 d^T g$, takže $A \ge \varepsilon_1 > 0$, $B \le \varepsilon_2 < 1$ a $B - 1 < 0$.

Poznámka 162. Použijeme-li větu o střední hodnotě (tvrzení 2) ve zpětném směru, dostaneme

$$F = F_+ - d^T g^+ + \frac{1}{2} d^T G_+ d + o(\|d\|^2).$$

Zanedbáme-li člen $o(\|d\|^2)$ a aproximujeme-li $d^T G_+ d$ pomocí $d^T B_+ d$, můžeme psát

$$d^T B_+ d = 2(F - F_+ + d^T g_+),$$

což po dosazení do (397) dává

$$\rho = \frac{d^T y}{d^T B_+ d} = \frac{d^T y}{2(F - F_+ + d^T g_+)} = \frac{B - 1}{2(B - A)}. \tag{398}$$

Poznámka 163. K přesnějšímu odhadu čísla $d^T B_+ d$ můžeme použít více členů Taylorova rozvoje. Platí

$$F = F_+ - d^T g_+ + \frac{1}{2} d^T G_+ d - \frac{1}{6} d^T (T_+ d) d + o(\|d\|^3),$$

$$d^T g = d^T g_+ - d^T G_+ d + \frac{1}{2} d^T (T_+ d) d + o(\|d\|^3),$$

kde

$$d^T (T_+ d) d = \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{k=1}^{n} \frac{\partial^3 F(x_+)}{\partial x_i \partial x_j \partial x_k} d_i d_j d_k.$$

Zanedbáme-li člen $o(\|d\|^3)$ a aproximujeme-li $d^T G_+ d$ pomocí $d^T B_+ d$, dostaneme po úpravě

$$\begin{aligned} 6(F - F_+ + d^T g_+) &= 3d^T B_+ d - d^T (T_+ d)d, \\ 2(d^T g - d^T g_+) &= -2d^T B_+ d + d^T (T_+ d)d, \end{aligned}$$

což po sečtení dává

$$d^T B_+ d = 6(F - F_+) + 4d^T g_+ + 2d^T g,$$

takže s použitím (397) dostaneme

$$\rho = \frac{d^T y}{d^T B_+ d} = \frac{d^T y}{6(F - F_+) + 4d^T g_+ + 2d^T g} = \frac{B - 1}{4B - 6A + 2} = \frac{B - 1}{4(B - 1) - 6(A - 1)}. \tag{399}$$

Existují další způsoby, jak lze určit hodnotu parametru ρ. Označme $\varphi(\alpha) = F(x + \alpha s)$. Pak platí

$$d^T B_+ d = \frac{1}{\rho} d^T y = \frac{\alpha}{\rho}(\varphi'(\alpha) - \varphi'(0))$$

a použijeme-li aproximaci $\varphi''(\alpha) = s^T B_+ s = d^T B_+ d/\alpha^2$, můžeme psát

$$\rho = \frac{\varphi'(\alpha) - \varphi'(0)}{\alpha \varphi''(\alpha)}. \tag{400}$$

Zvolíme-li vhodný tvar funkce $\varphi(\alpha)$ a spočteme-li $\varphi'(0)$, $\varphi'(\alpha)$, $\varphi''(\alpha)$, můžeme podle předchozího vzorce určit odpovídající hodnotu parametru ρ. Pro metody spádových směrů splňující slabou Wolfeho podmínku platí $\varphi'(0) < 0$, $\varphi(\alpha) - \varphi(0) \leq \varepsilon_1 \alpha \varphi'(0)$ a $\varphi'(\alpha) \geq \varepsilon_2 \varphi'(0)$.

Poznámka 164. Jednou z možností, uvedenou v práci [5], je použití kubického modelu

$$\varphi(\alpha) = a\alpha^3 + b\alpha^2 + c\alpha + d,$$

jehož čtyři koeficienty a, b, c, d lze určit pomocí hodnot $\varphi(0)$, $\varphi(\alpha)$ a derivací $\varphi'(0)$, $\varphi'(\alpha)$. Pak dosazením $\varphi'(0)$, $\varphi'(\alpha)$ a $\varphi''(\alpha) = 6a\alpha + 2b$ do (400) dostaneme vzorec (399), stejný jako v případě použití čtyř členů zpětného Taylorova rozvoje.

Poznámka 165. Velmi se osvědčilo použití homogenního modelu

$$\varphi(\alpha) = a\alpha^r + b\alpha + c,$$

kde a, b, c jsou neznáme koeficienty a $r > 1$ je neznámý exponent.

Věta 98. *Uvažujme homogenní model $\varphi(\alpha) = a\alpha^r + b\alpha + c$, $a \neq 1$, kde délka kroku α je získána metodou spádových směrů spňující slabou Wolfeho podmínku. Pak platí*

$$\rho = \frac{A - 1}{B - A}, \tag{401}$$

kde

$$\begin{aligned} A &= \frac{\varphi(\alpha) - \varphi(0)}{\alpha \varphi'(0)} = \frac{F_+ - F}{d^T g}, \\ B &= \frac{\varphi'(\alpha)}{\varphi'(0)} = \frac{d^T g_+}{d^T g}. \end{aligned}$$

Pro hodnotu (401) platí $\rho > 0$ právě tehdy, odpovídá-li tato hodnota homogennímu modelu s $r > 1$.

Důkaz Zřejmě

$$\begin{aligned}\varphi'(\alpha) &= ar\alpha^{r-1} + b,\\ \varphi''(\alpha) &= ar(r-1)\alpha^{r-2}\end{aligned}$$

a protože $a \neq 0$ a $b = \varphi'(0) < 0$, můžeme psát

$$\begin{aligned}B - 1 &= \frac{\varphi'(\alpha)}{\varphi'(0)} - 1 = \frac{ar\alpha^{r-1} + b}{b} - 1 = \frac{ar\alpha^{r-1}}{b},\\ A - 1 &= \frac{\varphi(\alpha) - \varphi(0)}{\alpha\varphi'(0)} - 1 = \frac{a\alpha^r + b\alpha}{\alpha b} - 1 = \frac{a\alpha^{r-1}}{b},\end{aligned}$$

odkud plyne

$$r = \frac{B-1}{A-1}.$$

Dále platí

$$\begin{aligned}\frac{\alpha\varphi''(\alpha)}{\varphi'(0)} &= \frac{ar(r-1)\alpha^{r-1}}{b} = (B-1)(r-1)\\ &= (B-1)\left(\frac{B-1}{A-1} - 1\right) = (B-1)\frac{B-A}{A-1},\end{aligned}$$

což po dosazení do (400) dává

$$\rho = \frac{\varphi'(\alpha) - \varphi'(0)}{\varphi'(0)}\frac{\varphi'(0)}{\alpha\varphi''(\alpha)} = (B-1)\frac{A-1}{(B-1)(B-A)} = \frac{A-1}{B-A}.$$

Nyní je třeba ukázat že $\rho > 0$ platí právě tehdy když $r > 1$. Jelikož délka kroku α je získána metodou spádových směrů spňující slabou Wolfeho podmínku, platí $\varphi'(\alpha) \geq \varepsilon_2\varphi'(0)$, takže $B \leq \varepsilon_2 < 1$ a $B - 1 < 0$. Předpokládejme nejprve, že $r = (B-1)/(A-1) > 1$, takže nutně $A - 1 < 0$. Můžeme tedy psát $B - 1 < A - 1$, neboli $B - A < 0$, což dává $\rho > 0$. Předpokládejme nyní, že $r = (B-1)/(A-1) < 1$. Pokud $A - 1 > 0$ (takže $r < 0$), dostaneme $B - 1 < A - 1$, neboli $B - A < 0$, což dává $\rho < 0$. Pokud $A - 1 < 0$ (takže $0 < r < 1$), dostaneme $B - 1 > A - 1$, neboli $B - A > 0$, což dává $\rho < 0$. Poznamenejme, že pro $r = 1$ platí $B - A = 0$, takže hodnota ρ není definována. Podobně pro $\rho = 0$ platí $A - 1 = 0$ a hodnota r není definována. □

Parametr γ slouží ke škálování matice H, neboť aktualizace (286) s $\gamma \neq 1$ je ekvivalentní aktualizaci (286) s $\gamma = 1$ aplikované na matici γH. Důvod pro škálování poskytuje věta 94 a následující věta.

Věta 99. *Nechť $\tilde{F}(\tilde{x}) = F(T^{-1}x)$, kde T je regulární čtvercová matice. Nechť $\tilde{x}_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s proměnnou metrikou z Broydenovy třídy s počáteční maticí $\tilde{H}_1$ aplikovanou na funkci $\tilde{F}(\tilde{x})$ a $x_i \in R^n$, $i \in N$, je posloupnost generovaná toutéž metodou s proměnnou metrikou aplikovanou na funkci $F(x)$. Pak pokud používáme stejný výběr délky kroku a pokud $H_1 = T\tilde{H}_1T^T$, platí $x_i = T\tilde{x}_i$ (metoda s proměnnou metrikou je invariantní vzhledem k lineární transformaci proměnných).*

Důkaz Snadno se dokáže (derivováním složené funkce $\tilde{F}(\tilde{x}) = F(T^{-1}x)$), že platí $\tilde{g}(\tilde{x}) = T^Tg(x)$ a $\tilde{G}(\tilde{x}) = T^TG(x)T$. Ukážeme, že $H_i = T\tilde{H}_iT^T$, $\forall i \in N$ (podle předpokladu to platí pro $i = 1$). Pak

$$x_{i+1} = x_i - \alpha_iH_ig_i = T(\tilde{x}_i - \alpha_i\tilde{H}_iT^Tg_i) = T(\tilde{x}_i - \alpha_i\tilde{H}_i\tilde{g}_i) = T\tilde{x}_{i+1}.$$

Důkaz provedeme indukcí. Předpokládejme, že $H = T\tilde{H}T^T$ (platí to v první iteraci). Protože $d = T\tilde{d}$ a $y = (T^T)^{-1}\tilde{y}$, můžeme psát $U = [d, Hy] = [T\tilde{d}, T\tilde{H}T^T(T^T)^{-1}\tilde{y}] = T[\tilde{d}, \tilde{H}\tilde{y}] = T\tilde{U}$, takže

$$\frac{1}{\gamma}H_+ = H + UMU^T = T\tilde{H}T^T + T\tilde{U}M\tilde{U}^TT^T = \frac{1}{\gamma}T\tilde{H}_+T^T.$$

□

Poznámka 166. Zvolíme-li $T = G^{-1/2}$, platí $\tilde{G} = T^T G T = I$. Odtud plyne, že pro libovolně špatně podmíněnou úlohu, můžeme lineární transformací proměnných docílit toho, že nová úloha je dobře podmíněná a zvolíme-li vhodně počáteční matici H_1, konverguje metoda s proměnnou metrikou velmi rychle. Proto je účelné matici H_1 a (jelikož násobení skalárem nedokáže dobře vystihnout transformaci $T\tilde{H}_1 T^T$) také matice H_i v dalších iteračních krocích vhodně škálovat. Vzhledem k tomu, že aproximujeme podmínku $\tilde{G}^{-1}y = \rho d$, je výhodné volit γ tak aby $\gamma H y \approx \rho d$, což po vynásobení zleva vektorem y^T dává $\rho/\gamma = a/b$ a po vynásobení zleva vektorem $H^{-1}d^T$ dává $\rho/\gamma = b/c$. Vhodný je také geometrický střed $\rho/\gamma = \sqrt{a/c}$.

Poznámka 167. Škálování provádíme tak, že určíme podíl ρ/γ (poznámka 156) a hodnotu parametru ρ (buď hodnotu $\rho = 1$ nebo některou z hodnot (398), (399), (401)). Pak $\gamma = \rho/(\rho/\gamma)$. Z předchozího výkladu by se mohlo zdát, že je výhodné škálovat matici H v každém iteračním kroku. To však odporuje předpokladům zaručujícím superlineární rychlost konvergence (poznámka 174). Proto se používají různé strategie škálování, kdy se hodnota $\gamma \neq 1$ používá pouze v některých iteračních krocích.

(NS) Žádné škálování. V každém iteračním kroku pokládáme $\gamma = 1$.

(PS) Počáteční škálování [142]. V prvním iteračním kroku (nebo po restartu) určíme γ tak, aby podíl ρ/γ splňoval vhodné podmínky (například $b/c \leq \rho/\gamma \leq a/b$). V ostatních iteračních krocích pokládáme $\gamma = 1$.

(IS) Intervalové škálování [109]. V prvním iteračním kroku (nebo po restartu) postupujeme stejně jako v případě (PS). V ostatních iteračních krocích testujeme zda získaná hodnota γ leží v intervalu $0 < \underline{\gamma} \leq \gamma \leq \overline{\gamma}$ (kde například $\underline{\gamma} = 1$ a $\overline{\gamma} = 6$). Neleží-li hodnota γ v tomto intervalu pokládáme $\gamma = 1$.

(CS) Řízené škálování [99]. Postupujeme v zásadě stejně jako v případě (IS). Hodnotu $\gamma = 1$ však používáme mnohem častěji. Nechť α_1 je počáteční odhad délky kroku (obvykle $\alpha_1 = 1$), nechť $F_1 = F(x + \alpha_1 s)$, $g_1 = g(x + \alpha_1 s)$, $\lambda_1 = s^T g_1 / s^T g$, a nechť $\lambda > 0$ je vhodná konstanta (například $\lambda = 0.2$). Pak hodnotu $\gamma = 1$ použijeme navíc v následujících případech (γ je původní hodnota určená podle (IS)):
   (a) Jestliže $|\lambda_1| \leq \lambda$ a $F_1 \leq F$.
   (b) Jestliže $\gamma > 1$ a buď $F_1 > F$ nebo $\lambda_1 < 0$.
   (c) Jestliže $\gamma < 1$ a $F_1 \leq F$ a $\lambda_1 > 0$.

(AS) Permanentní škálování. V každém iteračním kroku postupujeme tak jako v prvním iteračním kroku strategie (PS).

Poznámka 168. Podíly a/b, b/c a $\sqrt{a/c}$ používané při škálování lze nahradit spektrálními hodnotami uvedenými v poznámce 87 (kde $\lambda_i = 1$). Účinost takového škálování je však nižší. Kromě škálování popsaného v poznámce 167 lze použít inverzní škálování [21]. V tomto případě určíme podíl ρ/γ a položíme $\gamma = 1$, takže $\rho = \rho/\gamma$. V tomto případě je vhodnější nahradit podíly a/b, b/c a $\sqrt{a/c}$ spektrálními hodnotami uvedenými v poznámce 87. Význam inverzního spektrálního škálování lze demonstrovat na metodě BFGS. Položíme-li v (306) $\beta = 0$, $\gamma = 1$ a $\rho = y^T d / y^T y$, dostaneme

$$B_+ = B + \frac{y^T d}{y^T y} \frac{y\, y^T}{y^T d} - \frac{Bd(Bd)^T}{d^T B d},$$

což dává

$$Tr\, B_+ = Tr\, B + \frac{y^T d}{y^T y} \frac{y^T y}{y^T d} - \frac{d^T B^2 d}{d^T B d} \leq Tr\, B + 1.$$

takže stopa symetrické pozitivně definitní matice B, a tedy ani její spektrální norma, nemohou příliš rychle narůstat.

4.5 Globální konvergence

Nyní se budeme zabývat globální konvergencí metod s proměnnou metrikou. Omezíme se přitom na metody spádových směrů, jejichž iterační krok má tvar $x_{i+1} = x_i - \alpha_i H_i g_i$, $i \in N$, kde délka kroku $\alpha_i > 0$ se

určuje tak, aby byla splněna slabá Wolfeho podmínka. Pak $B_i d_i = -\alpha_i g_i$ a $b_i = y_i^T d_i > 0$, $i \in N$. Důkaz globální konvergence metod s proměnnou metrikou vyžaduje, aby minimalizovaná funkce byla stejnoměrně konvexní (předpoklad F5).

Nejprve dokážeme globální konvergenci standardní metody BFGS, kdy $\gamma_i = 1$, $\rho_i = 1$ a $\beta_i = 0$, $i \in N$, takže

$$B_{i+1} = B_i + \frac{y_i y_i^T}{y_i^T d_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i}, \tag{402}$$

neboť myšlenku tohoto důkazu, publikovaného v [136], lze použít k důkazu globální konvergence metody BFGS pro separovatelné úlohy (věta 229).

Věta 100. *Uvažujme metodu BFGS používající aktualizaci (402), kde matice B_1 je pozitivně definitní. Splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5, je tato metoda globálně konvergentní (definice 14).*

Důkaz (a) Jelikož $y_i = \tilde{G}_i d_i$ (věta 94), kde matice $\tilde{G}_i$ vyhovuje nerovnostem v předpokladech F4, F5, a jelikož stopa je lineární maticovou funkcí a pro libovolné dva vektory $u \in R^n$, $v \in R^n$ platí $Tr(uv^T) = v^T u$, můžeme psát

$$Tr\, B_{i+1} = Tr\, B_i + \frac{d_i^T \tilde{G}_i^2 d_i}{d_i^T \tilde{G}_i d_i} - \frac{d_i^T B_i^2 d_i}{d_i^T B_i d_i} \le Tr\, B_1 + i\overline{G} - \sum_{j=1}^{i} \frac{d_i^T B_i^2 d_i}{d_i^T B_i d_i},$$

což dává

$$Tr\, B_{i+1} \le 2i\overline{G}, \qquad \sum_{j=1}^{i} \frac{d_i^T B_i^2 d_i}{d_i^T B_i d_i} \le 2i\overline{G} \tag{403}$$

(pokud volíme $\overline{G}$ tak, aby platilo $\overline{G} \ge Tr\, B_1$), a podle lemmatu 2 též

$$\prod_{j=1}^{i} \frac{d_i^T B_i^2 d_i}{d_i^T B_i d_i} \le (2\overline{G})^i \tag{404}$$

(b) Předpokládejme, že uvažovaná metoda není globálně konvergentní. Pak není splněna podmínka (38) a existuje tedy číslo $c > 0$ takové, že pro libovolný index $i \in N$ platí

$$\sum_{j=1}^{i} \cos^2 \theta_j \le c$$

a podle lemmatu 2 též

$$\prod_{j=1}^{i} \cos^2 \theta_j \le \left(\frac{c}{i}\right)^i. \tag{405}$$

Jelikož

$$\cos^2 \theta_j = \frac{(g_j^T d_j)^2}{\|g_j\|^2 \|d_j\|^2} = \frac{d_j^T B_j d_j}{d_j^T B_j^2 d_j} \frac{d_j^T B_j d_j}{d_j^T d_j}$$

(neboť $B_j d_j = -\alpha_j g_j$), můžeme podle (404) a (405) psát

$$\prod_{j=1}^{i} \frac{d_j^T B_j d_j}{d_j^T d_j} = \prod_{j=1}^{i} \frac{d_j^T B_j^2 d_j}{d_j^T B_j d_j} \cos^2 \theta_j \le (2\overline{G})^i \left(\frac{c}{i}\right)^i = \left(\frac{2c\overline{G}}{i}\right)^i \tag{406}$$

(c) Pro metodu BFGS podle (312) (kde $\gamma_i = 1$, $\rho_i = 1$ a $\beta_i = 0$) platí

$$\frac{\det B_{i+1}}{\det B_i} = \frac{y_i^T d_i}{d_i^T B_i d_i} = \frac{d_i^T \tilde{G}_i d_i}{d_i^T B_i d_i},$$

takže s použitím vztahů (403) a (410) dostaneme

$$\prod_{j=1}^{i} \frac{d_j^T \tilde{G}_j d_j}{d_j^T B_j d_j} = \prod_{j=1}^{i} \frac{\det\, B_{j+1}}{\det\, B_j} = \frac{\det\, B_{i+1}}{\det\, B_1} \leq \frac{1}{\det\, B_1} \left(\frac{Tr\, B_{i+1}}{n} \right)^n \leq \frac{1}{\det\, B_1} \left(\frac{2i\overline{G}}{n} \right)^n \triangleq \tilde{c}\, i^n \leq \overline{c}^i, \quad (407)$$

pokud $\tilde{c}\, i^n \leq \overline{c}^i$, neboli $\log \tilde{c} + n \log i \leq i \log \overline{c}$, $\forall i \in N$. Tato nerovnost je splněna, pokud pro libovolný index $i \in N$ platí $\log \overline{c} \geq (\log \tilde{c} + n \log i)/i \geq \log \tilde{c} + n \log i/i$, a jelikož výraz $\log t/t$ nabývá maxima $1/e$ pro $t = e$, stačí volit $\log \overline{c} \geq \log \tilde{c} + n/e$. Spojíme-li (406) a (407), dostaneme

$$\prod_{j=1}^{i} \frac{d_j^T \tilde{G}_j d_j}{d_j^T d_j} = \prod_{j=1}^{i} \frac{d_j^T \tilde{G}_j d_j}{d_j^T B_j d_j} \frac{d_j^T B_j d_j}{d_j^T d_j} \leq \left(\frac{2c\overline{c}\overline{G}}{i} \right)^i,$$

a jelikož podle (F5) platí $d_j^T \tilde{G}_j d_j / d_j^T d_j \geq \underline{G}$, $j \in N$, můžeme psát

$$\underline{G}^i \leq \left(\frac{2c\overline{c}\overline{G}}{i} \right)^i \quad \Rightarrow \quad \underline{G} \leq \frac{2c\overline{c}\overline{G}}{i} \qquad \forall i \in N,$$

což je však spor, neboť pravá strana poslední nerovnosti konverguje k nule pokud $i \to \infty$. □

Nyní budeme vyšetřovat obecnější metody s proměnnou metrikou. Omezíme se přitom na perfektní metody z Broydenovy třídy takové, že matice B_1 je pozitivně definitní a pro $i \in N$ platí

$$0 < \underline{\gamma} \leq \gamma_i \leq \overline{\gamma}, \qquad 0 < \underline{\rho} \leq \rho_i \leq \overline{\rho}, \qquad (1-\lambda)\beta_i^* \leq \beta_i \leq 1 - \lambda, \quad (408)$$

kde $0 < \lambda \leq 1$. Navíc budeme předpokládat, že $\underline{\gamma} = 1$, i když lze tento předpoklad nahradit vhodnou strategií škálování (poznámka 169). Za těchto předpokladů jsou všechny matice B_i, $i \in N$, pozitivně definitní. Označíme-li $0 < \lambda_1(B_i) \leq \cdots \leq \lambda_n(B_i)$ vlastní čísla symetrické pozitivně definitní matice B_i, platí

$$\|B_i\| = \lambda_n(B_i) \leq \sum_{j=1}^{n} \lambda_j(B_i) = Tr\, B_i \quad (409)$$

a použijeme-li lemma 2, dostaneme

$$\det\, B_i = \prod_{j=1}^{n} \lambda_j(B_i) \leq \left(\frac{1}{n} \sum_{j=1}^{n} \lambda_j(B_i) \right)^n = \left(\frac{1}{n} Tr\, B_i \right)^n. \quad (410)$$

K důkazu globální konvergence použijeme postup uvedený v práci [19].

Lemma 48. *Uvažujme metodu s proměnnou metrikou z Broydenovy třídy (306) takovou, že pro $i \in N$ platí (408) (zde nepotřebujeme, aby platilo $\underline{\gamma} = 1$). Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1, F4, F5. Pak:*

(a) Existuje konstanta $\overline{C}$ taková, že $Tr\, B_{i+1} \leq \overline{C}^i$.
(b) Existuje konstanta $\underline{C}$ taková, že

$$\prod_{j=1}^{i} \frac{c_j}{b_j} \geq \underline{C}^i, \qquad \sum_{j=1}^{i} \frac{c_j}{b_j} \geq i\underline{C}.$$

Důkaz (a) Vztah (306) můžeme po roznásobení zapsat ve tvaru

$$\gamma_i B_{i+1} = B_i + \frac{\gamma_i}{\rho_i} \frac{y_i y_i^T}{y_i^T d_i} + \beta_i \frac{d_i^T B_i d_i}{y_i^T d_i} \frac{y_i y_i^T}{y_i^T d_i} - \frac{\beta_i}{y_i^T d_i} (B_i d_i y_i^T + y_i (B_i d_i)^T) + \frac{\beta_i - 1}{d_i^T B_i d_i} B_i d_i (B_i d_i)^T. \quad (411)$$

Využijeme-li toho, že stopa je lineární maticovou funkcí a toho, že pro libovolné dva vektory $u \in R^n$, $v \in R^n$ platí $Tr(uv^T) = v^T u$, dostaneme

$$\begin{aligned} Tr\, B_{i+1} &= \frac{1}{\gamma_i}\left(Tr\, B_i + \frac{\gamma_i}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i)\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i}\right) \\ &\leq \frac{1}{\underline{\gamma}}\left(Tr\, B_i + \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T d_i}{y_i^T d_i}\|B_i\| + 2\frac{\|y_i\|\|d_i\|}{y_i^T d_i}\|B_i\|\right) + \frac{1}{\underline{\rho}}\frac{y_i^T y_i}{y_i^T d_i} \end{aligned} \tag{412}$$

neboť $\beta_i \leq 1$. Protože $y_i = \tilde{G}_i d_i$, kde matice $\tilde{G}_i$ vyhovuje nerovnostem v předpokladech F4, F5, můžeme psát

$$\begin{aligned} \frac{y_i^T y_i}{y_i^T d_i} &= \frac{d_i^T \tilde{G}_i^2 d_i}{d_i^T \tilde{G}_i d_i} \leq \overline{G}, \\ \frac{d_i^T d_i}{y_i^T d_i} &= \frac{d_i^T d_i}{d_i^T \tilde{G}_i d_i} \leq \frac{1}{\underline{G}}, \\ \frac{\|y_i\|\|d_i\|}{y_i^T d_i} &= \sqrt{\frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T d_i}{y_i^T d_i}} \leq \sqrt{\frac{\overline{G}}{\underline{G}}}. \end{aligned}$$

Dosadíme-li tyto nerovnosti spolu s (409) do (412), dostaneme

$$\begin{aligned} Tr\, B_{i+1} &\leq Tr\, B_{i+1} + 1 \leq \frac{1}{\underline{\gamma}}\left(1 + \frac{\overline{G}}{\underline{G}} + 2\sqrt{\frac{\overline{G}}{\underline{G}}}\right) Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + 1 \\ &\leq \overline{K}(Tr\, B_i + 1) \leq \overline{K}^i (Tr\, B_1 + 1) \leq \overline{C}^i, \end{aligned} \tag{413}$$

kde

$$\overline{K} = \max\left(\frac{1}{\underline{\gamma}}\left(1 + \frac{\overline{G}}{\underline{G}} + 2\sqrt{\frac{\overline{G}}{\underline{G}}}\right),\ \frac{\overline{G}}{\underline{\rho}} + 1\right), \qquad \overline{C} = \overline{K}(Tr\, B_1 + 1).$$

(b) Použijeme-li vztahy (312) a (408), můžeme psát

$$\frac{\det\, B_{i+1}}{\det\, B_i} = \left(\frac{1}{\gamma_i}\right)^n \frac{\gamma_i}{\rho_i}\frac{b_i}{c_i}\left(1 - \frac{\beta_i}{\beta_i^*}\right) \geq \left(\frac{1}{\overline{\gamma}}\right)^n \frac{\underline{\gamma}}{\overline{\rho}}\frac{b_i}{c_i}\lambda \triangleq \underline{K}\frac{b_i}{c_i}, \tag{414}$$

takže

$$\frac{\det\, B_{i+1}}{\det\, B_1} \geq \underline{K}^i \prod_{j=1}^{i} \frac{b_j}{c_j}$$

a protože $\det\, H_{i+1} = 1/\det\, B_{i+1}$, platí

$$\det\, H_{i+1} \leq \frac{\det\, H_1}{\underline{K}^i} \prod_{j=1}^{i} \frac{c_j}{b_j} \leq \frac{\det\, H_1 + 1}{\underline{K}^i} \prod_{j=1}^{i} \frac{c_j}{b_j} \leq \frac{1}{K^i} \prod_{j=1}^{i} \frac{c_j}{b_j},$$

kde $K = \underline{K}/(\det H_1 + 1)$. Použijeme-li (410) a (a), dostaneme

$$\det\, B_{i+1} \leq \left(\frac{1}{n} Tr\, B_{i+1}\right)^n \leq (Tr\, B_{i+1})^n \leq \overline{C}^{in} \triangleq C^i,$$

takže

$$\frac{1}{C^i} \leq \det H_{i+1} \leq \frac{1}{K^i} \prod_{j=1}^{i} \frac{c_j}{b_j},$$

neboli

$$\prod_{j=1}^{i} \frac{c_j}{b_j} \geq \left(\frac{K}{C}\right)^i \triangleq \underline{C}^i$$

a podle lemmatu 2 platí

$$\sum_{j=1}^{i} \frac{c_j}{b_j} \geq i \left(\prod_{j=1}^{i} \frac{c_j}{b_j} \right)^{1/i} \geq i\underline{C}.$$

□

Věta 101. *(Globální konvergence) Uvažujme metodu s proměnnou metrikou z Broydenovy třídy (306) takovou, že pro $i \in N$ platí (408), kde $\underline{\gamma} = 1$. Splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1, F4, F5, je tato metoda globálně konvergentní (definice 14).*

Důkaz Použijeme opět vztah (412). Protože $y_i = \tilde{G}_i d_i$ a $B_i d_i = -\alpha_i g_i$, můžeme psát

$$\frac{|y_i^T B_i d_i|}{y_i^T d_i} = \frac{|y_i^T B_i d_i|}{d_i^T B_i d_i} \frac{c_i}{b_i} \leq \frac{\|\tilde{G} d_i\| \|\alpha_i g_i\|}{-\alpha_i d_i^T g_i} \frac{c_i}{b_i} \leq \frac{\overline{G}}{\cos\theta_i} \frac{c_i}{b_i}, \tag{415}$$

$$\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} = \frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \frac{y_i^T d_i}{d_i^T B_i d_i} \frac{c_i}{b_i} \geq \frac{\alpha_i^2 \|g_i\|^2 \underline{G} \|d_i\|^2}{\alpha_i^2 (d_i^T g_i)^2} \frac{c_i}{b_i} = \frac{\underline{G}}{\cos^2\theta_i} \frac{c_i}{b_i}, \tag{416}$$

neboť $y_i^T d_i \geq \underline{G} \|d_i\|^2$, což spolu s $1 \leq \gamma_i \leq \overline{\gamma}$ a $\beta_i \leq 1 - \lambda < 1$ dává

$$\begin{aligned} Tr\, B_{i+1} &\leq Tr\, B_i + \frac{1}{\underline{\rho}} \frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i} \frac{d_i^T B_i d_i}{y_i^T d_i} - 2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i) \frac{1}{\overline{\gamma}} \frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \\ &\leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \left(\overline{G} + 2\frac{\overline{G}}{\cos\theta_i} - \frac{\lambda \underline{G}}{\overline{\gamma} \cos^2\theta_i} \right) \frac{c_i}{b_i} = Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \xi_i \frac{c_i}{b_i}, \end{aligned} \tag{417}$$

kde

$$\xi_i = \overline{G} + 2\frac{\overline{G}}{\cos\theta_i} - \frac{\lambda \underline{G}}{\overline{\gamma} \cos^2\theta_i}.$$

Přepokládejme nyní, že uvažovaná metoda není globálně konvergentní. Pak podle věty 11 platí

$$\sum_{i=1}^{\infty} \cos^2\theta_i < \infty,$$

takže $\cos\theta_i \to 0$ a tedy $\xi_i \to -\infty$. Existuje tedy index $k \in N$ takový, že $\xi_i < -2\overline{G}/(\underline{\rho}\underline{C})$ $\forall i \geq k$. Abychom důkaz formálně zjednodušili, budeme bez újmy na obecnosti předpokládat, že $k = 1$ (v opačném případě můžeme indexy posunout). Pak podle (417) a lemmatu 48 (b) platí

$$Tr\, B_{i+1} \leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \xi_i \frac{c_i}{b_i} \leq Tr\, B_1 + i\frac{\overline{G}}{\underline{\rho}} - 2\frac{\overline{G}}{\underline{\rho}\underline{C}} \sum_{j=1}^{i} \frac{c_j}{b_j} \leq Tr\, B_1 - i\frac{\overline{G}}{\underline{\rho}} \triangleq C_1 - iC.$$

Zvolíme-li index i tak aby platilo $i > C_1/C$, dostaneme $Tr\, B_{i+1} < 0$, což je spor, neboť stopa symetrické pozitivně definití matice je kladná. □

Poznámka 169. Podmínka $\underline{\gamma} = 1$ slouží k tomu, abychom mohli použít odhad $Tr\, B_{i+1} \leq C_1 - iC$, neboť v opačném případě bychom museli konstantu $C_1 = Tr\, B_1$ nahradit číslem $Tr\, B_1/\underline{\gamma}^i$, nebo výrazem $Tr\, B_1/\omega_i$, kde $\omega_i = \prod_{j=1}^{i} \gamma_j$, a nebyla by zaručena omezenost shora. Tento nedostatek lze odstranit vhodnou škálovací strategií. Položíme-li $\gamma_i = 1$, kdykoliv původně zvolená hodnota nevyhovuje nerovnosti $\gamma_i\omega_i \geq \underline{\omega}$, kde $\underline{\omega}$ je vhodně zvolená dolní mez, platí $Tr\, B_{i+1} \leq C_1 - iC$, kde $C_1 = Tr\, B_1/\underline{\omega}$ (čísla ω_i, $i \in N$, lze získat rekurentním předpisem $\omega_1 = 1$ a $\omega_{i+1} = \gamma_i\omega_i$, $i \in N$).

Poznámka 170. Věta 101 teoreticky zdůvodňuje špatné konvergenční vlastnosti metody DFP (s nepřesným výběrem délky kroku). Metoda DFP odpovídá volbě $\beta_i = 1$, $i \in N$, takže není splněn předpoklad (408). Ze vztahu (411) vymizí poslední člen a nelze použít princip důkazu.

Poznámka 171. Věta 101 je nejobecnějším tvrzením o globální konvergenci (nemodifikovaných a nerestartovaných) metod s proměnnou metrikou. Tato věta vyžaduje, aby byl splněn předpoklad F5 (existence konstanty $\underline{G} > 0$), takže ji lze použít pouze pro konvexní funkce. Z důkazu věty 101 je zřejmé, že tento požadavek slouží pouze k tomu, aby platilo $y_i^T d_i \geq \underline{G}\|d_i\|^2$. Jednou z možností jak předpoklad (F5) obejít, je zvolit malé číslo $\underline{\tau} > 0$ a matici B_i aktualizovat pouze tehdy, když $y_i^T d_i \geq \underline{\tau}\|d_i\|^2$. Další možností je postupovat jako v poznámce 89 a nahradit v aktualizaci (411) vektor $y_i = g_{i+1} - g_i$ vektorem $\tilde{y}_i = y_i + \tau_i d_i$, kde $\tau_i = \max(0, \underline{\tau} - y_i^T d_i/d_i^T d_i)$. Používáme-li slabou Wolfeho podmínku, platí $\underline{\tau} - y_i^T d_i/d_i^T d_i \leq \tau_i \leq \underline{\tau}$ a

$$\tilde{y}_i^T d_i = y_i^T d_i + \tau_i\|d_i\|^2 \geq \frac{y_i^T d_i}{d_i^T d_i}\|d_i\|^2 + \left(\underline{\tau} - \frac{y_i^T d_i}{d_i^T d_i}\right)\|d_i\|^2 = \underline{\tau}\|d_i\|^2.$$

Jelikož $\tilde{y}_i = (\tilde{G}_i + \tau_i I)d_i$, můžeme psát

$$\frac{\tilde{y}_i^T \tilde{y}_i}{\tilde{y}_i^T d_i} = \frac{\tilde{y}_i^T \tilde{y}_i}{\tilde{y}_i^T (\tilde{G}_i + \tau_i I)^{-1}\tilde{y}_i} \leq \overline{G} + \underline{\tau}.$$

Pro aktualizace s vektory $\tilde{y}_i = y_i + \tau_i d_i$, $i \in N$, jsou tedy splněny všechny nerovnosti vystupující v důkazu věty 101 (kde místo $\underline{G}$ a $\overline{G}$ píšeme $\underline{\tau}$ a $\overline{G} + \underline{\tau}$). Místo konstanty $\underline{\tau}$ můžeme, stejně jako v poznámce 89, použít proměnnou hodnotu $\underline{\tau}_i = \overline{\tau}\min(1, \|g_i\|)$, kde $\overline{\tau} > 0$. Protože důkaz věty 101 provádíme sporem a předpokládáme, že $\|g_i\| > \underline{\varepsilon}$, platí v tomto případě $\underline{\tau}_i \geq \overline{\tau}\min(1, \underline{\varepsilon}) \stackrel{\Delta}{=} \underline{\tau}$.

Ukážeme nyní, že jsou-li splněny předpoklady věty 101, je konvergence (alespoň) lineární v tom smyslu, že platí (420). Použijeme přitom označení

$$R_i^* = G_*^{-1/2} B_i G_*^{-1/2}, \qquad u_i = G_*^{1/2} d_i. \qquad v_i = G_*^{-1/2} y_i, \tag{418}$$

kde $G_* = G(x^*)$, takže vzorec (411), vynásobený zleva i zprava maticí $G_*^{-1/2}$, lze zapsat ve tvaru

$$\begin{aligned}\gamma_i R_{i+1}^* &= R_i^* + \frac{\gamma_i}{\rho_i}\frac{v_i v_i^T}{v_i^T u_i} - \frac{1}{u_i^T R_i^* u_i} R_i^* u_i (R_i^* u_i)^T + \frac{\beta_i}{u_i^T R_i^* u_i}\left(\frac{u_i^T R_i^* u_i}{v_i^T u_i} v_i - R_i^* u_i\right)\left(\frac{u_i^T R_i^* u_i}{v_i^T u_i} v_i - R_i^* u_i\right)^T \\ &= R_i^* + \frac{\gamma_i}{\rho_i}\frac{v_i v_i^T}{v_i^T u_i} + \beta_i \frac{u_i^T R_i^* u_i}{v_i^T u_i}\frac{v_i v_i^T}{v_i^T u_i} - \frac{\beta_i}{v_i^T u_i}(R_i^* u_i v_i^T + v_i (R_i^* u_i)^T) + \frac{\beta_i - 1}{u_i^T R_i^* u_i} R_i^* u_i (R_i^* u_i)^T. \end{aligned} \tag{419}$$

Věta 102. *(Lineární konvergence) Nechť jsou splněny předpoklady věty 101. Pak platí $x_i \to x^*$ a*

$$\sum_{i=1}^{\infty} \|e_i\| = \sum_{i=1}^{\infty} \|x_i - x^*\| < \infty. \tag{420}$$

Důkaz (a) Funkce $F : \mathcal{D} \to R$ splňující předpoklad F5 je ryze konvexní na $\mathcal{D}$ a má tam tedy jediný stacionární bod x^*, který je jejím globálním minimem. Jelikož posloupnost F_i, $i \in N$, je podle (S2a) nerostoucí, platí $F_i \to F^*$, $g_i \to g^*$ a $x_i \to x^*$.

(b) Zřejmě $y_i = \tilde{G}_i d_i$, kde

$$\tilde{G}_i = \int_0^1 G(x_i + \lambda d_i)\mathrm{d}\lambda, \tag{421}$$

takže

$$v_i = G_*^{-1/2} y_i = G_*^{-1/2}\tilde{G}_i d_i = G_*^{-1/2}(G_* d_i + (\tilde{G}_i - G_*)d_i) = u_i + G_*^{-1/2}(\tilde{G}_i - G_*)G_*^{-1/2}u_i.$$

Jelikož

$$u_i^T G_*^{-1/2}(\tilde{G}_i - G_*)G_*^{-1/2}u_i \le \|G_*\|\|\tilde{G}_i - G_*\|u_i^T u_i,$$

a $\tilde{G}_i \to G_*$, platí $u_i^T v_i = u_i^T u_i(1 + o(1))$. Podobně platí $v_i^T v_i = u_i^T u_i(1 + o(1))$.

(c) Podle poznámky 10 platí $(1+o(1))(1+o(1)) = 1+o(1)$ a $(1+o(1))/(1+o(1)) = 1+o(1)$. Použijeme-li první vztah v (419) pro $(1-\lambda)\beta_i^* \le \beta_i \le 0$, dostaneme

$$\begin{aligned} Tr\, R_{i+1}^* &\le Tr\, R_i^* + \frac{1}{\underline{\rho}}\frac{v_i^T v_i}{v_i^T u_i} - \frac{1}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i} = Tr\, R_i^* + \frac{1}{\underline{\rho}}(1+o(1)) - \frac{1}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i} \\ &\le Tr\, R_i^* + C - \frac{\lambda}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i}, \end{aligned}$$

kde $C > 0$ je nějaká konstanta (která existuje, neboť $o(1) \to 0$) a $0 < \lambda \le 1$. Podobně, použijeme-li druhý vztah v (419) pro $0 < \beta_i \le 1 - \lambda$, můžeme pro dostatečně velké indexy psát

$$\begin{aligned} Tr\, R_{i+1}^* &\le Tr\, R_i^* + \frac{1}{\underline{\rho}}\frac{v_i^T v_i}{v_i^T u_i} + \beta_i \frac{v_i^T v_i}{v_i^T u_i}\frac{u_i^T R_i^* u_i}{v_i^T u_i} - 2\beta_i \frac{v_i^T R_i^* u_i}{v_i^T u_i} - (1-\beta_i)\frac{1}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i} \\ &= Tr\, R_i^* + \frac{1}{\underline{\rho}}(1+o(1)) - \beta_i \frac{u_i^T R_i^* u_i}{u_i^T u_i}(1+o(1)) - (1-\beta_i)\frac{1}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i} \\ &\le Tr\, R_i^* + C - \frac{\lambda}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i}, \end{aligned}$$

neboť $o(1) \to 0$, takže pro dostatečně velké indexy platí $1 + o(1) > 0$ (protože $\beta_i > 0$, je člen s β_i záporný a lze ho vynechat). V dalších úvahách budeme bez újmy na obecnosti předpokládat, že $1 + o(1) > 0$, $i \in N$ (v opačném případě můžeme indexy posunout). Pak lze pro $(1-\lambda)\beta_i^* \le \beta_i \le 1 - \lambda$ psát

$$\begin{aligned} Tr\, R_{i+1}^* &\le Tr\, R_i^* + C - \frac{\lambda}{\overline{\gamma}}\frac{(R_i^* u_i)^T R_i^* u_i}{u_i^T R_i^* u_i} \le Tr\, R_i^* + C - \frac{\lambda}{\overline{\gamma}}\frac{d_i^T B_i G_*^{-1} B_i d_i}{d_i^T B_i d_i} \\ &\le Tr\, R_i^* + C - \frac{\lambda \underline{G}}{\overline{\gamma}\overline{G}\cos^2\theta_i}\frac{c_i}{b_i} \le Tr\, R_1^* + iC - \frac{\lambda \underline{G}}{\overline{\gamma}\overline{G}}\sum_{j=1}^{i}\frac{1}{\cos^2\theta_j}\frac{c_j}{b_j}, \end{aligned}$$

neboť podle (418) platí $(R_i^* u_i)^T R_i^* u_i = d_i^T B_i G_*^{-1} B_i d_i$ a použitím (416) dostaneme

$$\frac{d_i^T B_i G_*^{-1} B_i d_i}{d_i^T B_i d_i} = \frac{d_i^T B_i G_*^{-1} B_i d_i}{(B_i d_i)^T B_i d_i}\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \ge \frac{1}{\overline{G}}\frac{\underline{G}}{cos^2\theta_i}\frac{c_i}{b_i}.$$

Protože matice R_{i+1}^* je pozitivně definitní, platí $Tr\, R_{i+1}^* \ge 0$, takže

$$\sum_{j=1}^{i}\frac{1}{\cos^2\theta_j}\frac{c_j}{b_j} \le \frac{\overline{\gamma}\overline{G}}{\lambda\underline{G}}\left(Tr\, R_1^* + iC - Tr\, R_{i+1}^*\right) \le \frac{\overline{\gamma}\overline{G}}{\lambda\underline{G}}\left(Tr\, R_1^* + C\right) i \triangleq Li.$$

Použijeme-li lemma 2, dostaneme

$$\prod_{j=1}^{i} \frac{1}{\cos^2 \theta_j} \frac{c_j}{b_j} \leq \left(\frac{1}{i} \sum_{j=1}^{i} \frac{1}{\cos^2 \theta_j} \frac{c_j}{b_j} \right)^i = L^i$$

a podle (b) lemmatu 48 platí

$$\prod_{j=1}^{i} \cos^2 \theta_j \geq \frac{1}{L^i} \prod_{j=1}^{i} \frac{c_j}{b_j} \geq \frac{\underline{C}^i}{L^i} \triangleq \underline{c}^i.$$

Použijeme-li znovu lemma 2, můžeme psát

$$\sum_{j=1}^{i} \cos^2 \theta_j \geq i \left(\prod_{j=1}^{i} \cos^2 \theta_j \right)^{1/i} = i\underline{c},$$

takže dokazované tvrzení plyne z věty 17 a poznámky 36. □

Poznámka 172. Ve větě 102 můžeme předpoklad F5 nahradit předpokladem použitým ve větě 14 (hromadný bod $x^* \in R^n$ posloupnosti x_i, $i \in N$, vyhovuje postačujícím podmínkám pro lokální minimum).

Platí-li (420), je možné s výhodou použít princip omezeného znehodnocení zformulovaný v následujícím lemmatu.

Lemma 49. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů konvergujících lineárně k bodu $x^* \in R^n$ (platí (420)) a nechť $\kappa_i \in R^n$, $i \in N$, je posloupnost kladných čísel taková, že $\kappa_{i+1} = \kappa_i(1 + O(\|e_i\|))$, kde $e_i = x_i - x^*$ (čili existuje číslo $C > 0$ takové, že $\kappa_{i+1} \leq \kappa_i(1 + C\|e_i\|)$, $i \in N$). Pak existuje konstanta $\overline{C} > 0$ taková, že $\kappa_i \leq \kappa_1 \exp(\overline{C})$, $i \in N$.*

Důkaz Podle lemmatu 2 pro $i \in N$ platí

$$\kappa_{i+1} \leq \kappa_1 \prod_{j=1}^{i} (1 + C\|e_j\|) \leq \kappa_1 \left(\frac{1}{i} \sum_{j=1}^{i} (1 + C\|e_j\|) \right)^i = \kappa_1 \left(1 + \frac{C}{i} \sum_{j=1}^{i} \|e_j\| \right)^i$$

a z (420) plyne existence konstanty $\overline{C}$ takové, že

$$\sum_{j=1}^{i} \|e_j\| \leq \sum_{j=1}^{\infty} \|e_j\| = \frac{\overline{C}}{C}.$$

Můžeme tedy pro $i \in N$ psát

$$\kappa_{i+1} \leq \kappa_1 \left(1 + \frac{\overline{C}}{i} \right)^i.$$

V základním kurzu analýzy se dokazuje, že posloupnost tvořená pravými stranami těchto nerovností je rostoucí a má limitu $\kappa_1 \exp(\overline{C})$ (tato limita se snadno určí pomocí l'Hospitalova pravidla). Platí tedy $\kappa_i \leq \kappa_1 \exp(\overline{C})$, $i \in N$. □

4.6 Superlineární konvergence

Nyní se budeme zabývat superlineární konvergencí metod s proměnnou metrikou. Při výkladu budeme používat výsledky uvedené v práci [76]. Jelikož superlineární konvergence vyžaduje, aby v jistém smyslu platilo $B_i \to G_i$ (věta 20), budeme předpokládat, že $\rho_i = \gamma_i = 1$, $i \in N$. Při vyšetřování superlineární

konvergence budeme pracovat s maticemi R_i, $i \in N$, zavedenými ve větě 94. Budeme opět používat označení

$$R_i = \tilde{G}_i^{-1/2} B_i \tilde{G}_i^{-1/2}, \qquad z_i = \tilde{G}_i^{1/2} d_i = \tilde{G}_i^{-1/2} y_i \tag{422}$$

a $R'_{i+1} = \tilde{G}_i^{-1/2} B_{i+1} \tilde{G}_i^{-1/2}$, takže vzorec (411) s $\rho_i = \gamma_i = 1$, vynásobený zleva i zprava maticí $\tilde{G}_i^{-1/2}$, lze zapsat ve tvaru

$$R'_{i+1} \quad = \quad R_i + \frac{z_i z_i^T}{z_i^T z_i} + \beta_i \frac{z_i^T R_i z_i}{z_i^T z_i} \frac{z_i z_i^T}{z_i^T z_i} - \frac{\beta_i}{z_i^T z_i}(R_i z_i z_i^T + z_i (R_i z_i)^T) + \frac{\beta_i - 1}{z_i^T R_i z_i} R_i z_i (R_i z_i)^T \tag{423}$$

Je však třeba mít na paměti, že $R_{i+1} = \tilde{G}_{i+1}^{-1/2} B_{i+1} \tilde{G}_{i+1}^{-1/2} \neq R'_{i+1}$. K důkazu superlineární konvergence metod s proměnnou metrikou lze s výhodou použít Frobeniovu normu. Froberniova norma matice M je definovaná vztahem $\|M\|_F = \sqrt{Tr(M^T M)}$. Z této definice plyne, že $\|M\| \leq \|M\|_F \leq \sqrt{n}\|M\|$. Využijeme toho, že pro libovolné matice M_1, M_2 platí $\|M_1\|_F^2 = Tr(M_1^T M_1)$, $\|M_2\|_F^2 = Tr(M_2^T M_2)$, takže

$$\|M_1 + M_2\|_F^2 = Tr((M_1 + M_2)^T (M_1 + M_2)) = \|M_1\|_F^2 + \|M_2\|_F^2 + 2Tr(M_1^T M_2). \tag{424}$$

Dále platí

$$\|uv^T\|_F^2 = Tr(vu^T uv^T) = u^T u Tr(vv^T) = u^T u v^T v. \tag{425}$$

Je-li matice M symetrická, je $\|M\|_F^2$ součtem druhých mocnin jejích vlastních čísel. Z toho plyne, že symetrické matice, které mají stejná vlastní čísla, mají stejné Frobeniovy normy.

Lemma 50. *Uvažujme aktualizaci*

$$R'_+ = R + \frac{zz^T}{z^T z} + \beta \frac{z^T R z}{z^T z} \frac{zz^T}{z^T z} - \beta \left(\frac{zz^T R}{z^T z} + \frac{R z z^T}{z^T z} \right) + (\beta - 1) \frac{R z z^T R}{z^T R z}$$

(vztah (423)). Pak platí

$$\begin{aligned} \|R'_+ - I\|_F^2 \;=\; \|R - I\|_F^2 \quad &- \quad (1-\beta) \left(\left(1 - \frac{z^T R^2 z}{z^T R z} \right)^2 + 2 \left(\frac{z^T R^3 z}{z^T R z} - \left(\frac{z^T R^2 z}{z^T R z} \right)^2 \right) \right) \\ &- \quad \beta \left(\left(1 - \frac{z^T R z}{z^T z} \right)^2 + 2\beta \left(\frac{z^T R^2 z}{z^T z} - \left(\frac{z^T R z}{z^T z} \right)^2 \right) \right) \\ &- \quad \beta(1-\beta) \left(\left(\frac{z^T R^2 z}{z^T R z} \right)^2 - \left(\frac{z^T R z}{z^T z} \right)^2 \right). \end{aligned} \tag{426}$$

Důkaz Aplikujeme-li pravidla (424) a (425) na vztah

$$R'_+ - I = R - I + \frac{zz^T}{z^T z} + \beta \frac{z^T R z}{z^T z} \frac{zz^T}{z^T z} - \beta \frac{zz^T R}{z^T z} - \beta \frac{R z z^T}{z^T z} + (\beta - 1) \frac{R z z^T R}{z^T R z},$$

dostaneme

$$\begin{aligned}
\|R'_+ - I\|_F^2 = \|R - I\|_F^2 &+ 1 + \beta^2 \left(\frac{z^T R z}{z^T z}\right)^2 + 2\beta^2 \left(\frac{z^T R z}{z^T z}\right)^2 + (\beta-1)^2 \left(\frac{z^T R^2 z}{z^T R z}\right)^2 \\
&+ 2\frac{z^T R z}{z^T z} + 2\beta \left(\frac{z^T R z}{z^T z}\right)^2 - 4\beta \frac{z^T R^2 z}{z^T z} + 2(\beta-1)\frac{z^T R^3 z}{z^T R z} \\
&- 2 - 2\beta\frac{z^T R z}{z^T z} + 4\beta\frac{z^T R z}{z^T z} - 2(\beta-1)\frac{z^T R^2 z}{z^T R z} + 2\beta\frac{z^T R z}{z^T z} \\
&- 4\beta\frac{z^T R z}{z^T z} + 2(\beta-1)\frac{z^T R z}{z^T z} - 4\beta^2 \left(\frac{z^T R z}{z^T z}\right)^2 \\
&+ 2\beta(\beta-1)\left(\frac{z^T R z}{z^T z}\right)^2 + 2\beta^2 \frac{z^T R^2 z}{z^T z} - 4\beta(\beta-1)\frac{z^T R^2 z}{z^T z} \\
= \|R - I\|_F^2 &- 1 + 2\beta\frac{z^T R z}{z^T z} - 2\beta^2 \frac{z^T R^2 z}{z^T z} - 2(\beta-1)\frac{z^T R^2 z}{z^T R z} \\
&+ 2(\beta-1)\frac{z^T R^3 z}{z^T R z} + \beta^2 \left(\frac{z^T R z}{z^T z}\right)^2 + (\beta-1)^2 \left(\frac{z^T R^2 z}{z^T R z}\right)^2.
\end{aligned}$$

Stejný výsledek dostaneme roznásobením vztahu (426). □

Důsledek 15. *Jsou-li splněny předpoklady lemmatu 50 s* $0 \le \beta \le 1$, *platí* $\|R'_+ - I\|_F \le \|R - I\|_F$.

Důkaz Použijeme-li Schwarzovu nerovnost, dostaneme

$$\begin{aligned}
\frac{z^T R^3 z}{z^T R z} - \left(\frac{z^T R^2 z}{z^T R z}\right)^2 &= \frac{z^T R^3 z z^T R z - (z^T R^2 z)^2}{(z^T R z)^2} \ge 0, \\
\frac{z^T R^2 z}{z^T z} - \left(\frac{z^T R z}{z^T z}\right)^2 &= \frac{z^T R^2 z z^T z - (z^T R z)^2}{(z^T z)^2} \ge 0, \\
\left(\frac{z^T R^2 z}{z^T R z}\right)^2 - \left(\frac{z^T R z}{z^T z}\right)^2 &= \frac{(z^T R^2 z z^T z)^2 - (z^T R z)^4}{(z^T R z z^T z)^2} \\
&= \frac{z^T R^2 z z^T z + (z^T R z)^2}{z^T R z z^T z} \frac{z^T R^2 z z^T z - (z^T R z)^2}{z^T R z z^T z} \ge 0.
\end{aligned}$$

Všechny závorky ve vztahu (426) jsou tedy nezáporné a jelikož $0 \le \beta \le 1$, platí $\|R'_+ - I\|_F^2 \le \|R - I\|_F^2$. □

Poznámka 173. Vynásobíme-li vztah (286) (s $\rho = \gamma = 1$) zleva i zprava maticí $\tilde{G}^{1/2}$, dostaneme

$$(R'_+)^{-1} = R^{-1} + \frac{z z^T}{z^T z} + \eta \frac{z^T R^{-1} z}{z^T z} \frac{z z^T}{z^T z} - \eta \left(\frac{z z^T R^{-1}}{z^T z} + \frac{R^{-1} z z^T}{z^T z}\right) + (\eta - 1)\frac{R^{-1} z z^T R^{-1}}{z^T R^{-1} z}.$$

Použijeme-li stejné úvahy jeko v důkazu lemmatu 50, můžeme psát

$$\begin{aligned}
\|(R'_+)^{-1} - I\|_F^2 = \|R^{-1} - I\|_F^2 &- (1-\eta)\left(\left(1 - \frac{z^T R^{-2} z}{z^T R^{-1} z}\right)^2 + 2\left(\frac{z^T R^{-3} z}{z^T R^{-1} z} - \left(\frac{z^T R^{-2} z}{z^T R^{-1} z}\right)^2\right)\right) \\
&- \eta\left(\left(1 - \frac{z^T R^{-1} z}{z^T z}\right)^2 + 2\eta\left(\frac{z^T R^{-2} z}{z^T z} - \left(\frac{z^T R^{-1} z}{z^T z}\right)^2\right)\right) \\
&- \eta(1-\eta)\left(\left(\frac{z^T R^{-2} z}{z^T R^{-1} z}\right)^2 - \left(\frac{z^T R^{-1} z}{z^T z}\right)^2\right). \qquad (427)
\end{aligned}$$

Pokud $0 \le \eta \le 1$, platí $\|(R'_+)^{-1} - I\|_F \le \|R^{-1} - I\|_F$.

Lemma 51. *Nechť bod $x^* \in R^n$ je hromadným bodem posloupnosti $x_i \in R^n$, $i \in N$, a nechť B_i, $i \in N$, je posloupnost pozitivně definitních matic získaná aktualizacemi z Broydenovy třídy s $\rho_i = 1$, $\gamma_i = 1$ a $0 \le \beta_i \le 1$. Pak, splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F4–F6, platí*

$$\|R_{i+1} - I\|_F + 1 = (\|R_i - I\|_F + 1)(1 + O(\|e_i\|)), \tag{428}$$
$$\|R_{i+1}^{-1} - I\|_F + 1 = (\|R_i^{-1} - I\|_F + 1)(1 + O(\|e_i\|)). \tag{429}$$

Důkaz Označme

$$\tilde{R}_i = B_i^{1/2} \tilde{G}_i^{-1} B_i^{1/2}, \qquad \tilde{R}'_{i+1} = B_{i+1}^{1/2} \tilde{G}_i^{-1} B_{i+1}^{1/2}.$$

Matice $\tilde{R}_i$ má stejná vlastní čísla jako matice R_i, neboť z $\tilde{G}_i^{-1/2} B_i \tilde{G}_i^{-1/2} x = \lambda x$, kde $x \neq 0$, plyne $B_i^{1/2} \tilde{G}_i^{-1} B^{1/2} y = \lambda y$, kde $y = B_i^{1/2} \tilde{G}_i^{-1/2} x \neq 0$. Platí tedy $\|\tilde{R}_i\|_F = \|R_i\|_F$ a $\|\tilde{R}_i - I\|_F = \|R_i - I\|_F$. Totéž platí pro matice $\tilde{R}'_{i+1}$ a R'_{i+1}. Použijeme-li důsledek 15 dostaneme

$$\begin{aligned} \|R_{i+1} - I\|_F &= \|\tilde{R}_{i+1} - I\|_F \le \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|\tilde{R}'_{i+1} - I\|_F \\ &= \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R'_{i+1} - I\|_F \le \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R_i - I\|_F. \end{aligned}$$

Stačí tedy dokázat, že $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F = (\|R_i - I\|_F + 1) O(\|e_i\|)$. Použijeme-li vztah (421) a nerovnost (6), můžeme psát

$$\begin{aligned} \|\tilde{G}_{i+1} - \tilde{G}_i\| &= \left\| \int_0^1 G(x_{i+1} + \lambda d_{i+1}) \mathrm{d}\lambda - \int_0^1 G(x_i + \lambda d_i) \mathrm{d}\lambda \right\| \\ &\le \int_0^1 \|G(x_{i+1} + \lambda d_{i+1}) - G(x_i + \lambda d_i)\| \mathrm{d}\lambda \\ &\le \overline{L} \int_0^1 \|e_{i+1} + \lambda d_{i+1} - e_i - \lambda d_i\| \mathrm{d}\lambda \\ &\le \overline{L} \left(\|e_{i+1}\| + \frac{1}{2}\|d_{i+1}\| + \|e_i\| + \frac{1}{2}\|d_i\| \right) = O(\|e_i\|), \end{aligned}$$

neboť $\|e_{i+1}\| = O(\|e_i\|)$ a $\|d_i\| = O(\|e_i\|)$ (poznámka 35), a z (7) plyne $\|\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}\| = O(\|e_i\|)$. Platí tedy

$$\begin{aligned} \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| &\le \|B_{i+1}^{1/2} (\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}) B_{i+1}^{1/2}\| \le \|B_{i+1}\| \|\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}\| \\ &= \|\tilde{G}_i^{1/2} \tilde{G}_i^{-1/2} B_{i+1} \tilde{G}_i^{-1/2} \tilde{G}_i^{1/2}\| \|\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}\| \le \|\tilde{G}_i\| \|R'_{i+1}\| \|\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}\| \\ &\le \overline{G} \|R'_{i+1}\| \|\tilde{G}_{i+1}^{-1} - \tilde{G}_i^{-1}\| = \|R'_{i+1}\| O(\|e_i\|). \end{aligned}$$

Ale

$$\|R'_{i+1}\| = \|I + R'_{i+1} - I\| \le 1 + \|R'_{i+1} - I\| \le 1 + \|R'_{i+1} - I\|_F \le 1 + \|R_i - I\|_F,$$

takže

$$\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F \le \sqrt{n} \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| = \|R'_{i+1}\| O(\|e_i\|) = (\|R_i - I\|_F + 1) O(\|e_i\|).$$

Jelikož z $0 \le \beta_i \le 1$ plyne $0 \le \eta_i \le 1$, lze podle poznámky 173 použít stejný postup pro posloupnost matic R_i^{-1}, $i \in N$, čímž dostaneme (429). □

Důsledek 16. *Jsou-li splněny předpoklady lemmatu 51 a platí-li (420), existují konstanty $\overline{R}$ a $\overline{B}$ takové, že $\|R_i\| \le \overline{R}$, $\|B_i\| \le \overline{B}$, $i \in N$, a konstanty $\underline{R}$, $\underline{B}$ takové, že $\|R_i^{-1}\| \le 1/\underline{R}$ a $\|H_i\| = \|B_i^{-1}\| \le 1/\underline{B}$. $i \in N$.*

Důkaz Jelikož $\|R_i\|_F \le \|I\|_F + \|R_i - I\|_F \le \sqrt{n}(\|R_i - I\|_F + 1)$ a posloupnost $\|R_i - I\|_F + 1$, $i \in N$, je podle (428) a podle lemmatu 49 omezená, je i posloupnost $\|R_i\| \le \|R_i\|_F$, $i \in N$, omezená. Jelikož

$$\|B_i\| = \|\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2}B_i\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2}\| \le \|\tilde{G}_i\|\|R_i\| \le \overline{G}\|R_i\|,$$

je i posloupnost $\|B_i\|$, $i \in N$, omezená. Podobným způsoben z (429) a z lemmatu 49 plyne, že i posloupnosti $\|R_i^{-1}\|$ a $\|H_i\| = \|B_i^{-1}\|$ jsou omezené. □

Věta 103. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů konvergujících lineárně k bodu $x^* \in R^n$ (platí (420)), kde x^* je stacionárním bodem funkce $F : \mathcal{D} \to R$ splňující předpoklady F4–F6, a nechť B_i, $i \in N$, je posloupnost pozitivně definitních matic získaná aktualizacemi z Broydenovy třídy s $\rho_i = 1$, $\gamma_i = 1$ a $0 \le \beta_i \le 1$. Pak platí*

$$\lim_{i\to\infty} \frac{\|(B_i - G_i)d_i\|}{\|d_i\|} = 0. \tag{430}$$

Důkaz Z důkazu lemmatu 51 (první nerovnost) víme, že

$$\|R_i - I\|_F - \|R'_{i+1} - I\|_F \le \|R_i - I\|_F - \|R_{i+1} - I\|_F + \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F,$$

kde $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F = (\|R_i - I\|_F + 1)O(\|e_i\|)$. Jelikož posloupnost $\|R_i - I\|_F + 1$, $i \in N$, je podle (428) a podle lemmatu 49 shora omezená, existuje konstanta $C > 0$ taková, že $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F \le C\|e_i\|$. Použijeme-li (420), dostaneme

$$\begin{aligned}\sum_{i=1}^{\infty}\left(\|R_i - I\|_F - \|R'_{i+1} - I\|_F\right) &\le \sum_{i=1}^{\infty}\left(\|R_i - I\|_F - \|R_{i+1} - I\|_F + C\|e_i\|\right)\\ &\le \|R_1 - I\|_F + C\sum_{i=1}^{\infty}\|e_i\| < \infty,\end{aligned}$$

takže platí

$$\lim_{i\to\infty}\left(\|R_i - I\|_F - \|R'_{i+1} - I\|_F\right) = 0.$$

a jelikož normy $\|R_i - I\|_F$ a $\|R'_{i+1} - I\|_F \le \|R_i - I\|_F$ jsou omezené, také

$$\lim_{i\to\infty}\left(\|R_i - I\|_F^2 - \|R'_{i+1} - I\|_F^2\right) = 0.$$

Nyní použijeme vztah (426). Protože poslední tři členy na pravé straně tohoto vztahu mají podle důsledku 15 (a jeho důkazu) stejné znaménko, musí konvergovat k nule, neboť jsme právě dokázali, že jejich součet konverguje k nule. Nechť $N = N_1 \cup N_2$, $N_1 \cap N_2 = \emptyset$ je rozklad množiny N takový, že

$$\limsup_{i\overset{N_1}{\to}\infty} \beta_i < 1, \qquad \liminf_{i\overset{N_2}{\to}\infty} \beta_i > 0$$

(například $N_1 = \{i \in N : 0 \le \beta_i \le 1/2\}$, $N_2 = \{i \in N : 1/2 < \beta_i \le 1\}$). Z konvergence zmíněných tří členů plyne, že

$$\begin{aligned}\lim_{i\overset{N_1}{\to}\infty}\frac{z_i^T R_i^3 z_i}{z_i^T R_i z_i} &= \lim_{i\overset{N_1}{\to}\infty}\frac{z_i^T R_i^2 z_i}{z_i^T R_i z_i} = 1,\\ \lim_{i\overset{N_2}{\to}\infty}\frac{z_i^T R_i^2 z_i}{z_i^T z_i} &= \lim_{i\overset{N_2}{\to}\infty}\frac{z_i^T R_i z_i}{z_i^T z_i} = 1,\end{aligned}$$

neboli

$$\begin{aligned}\lim_{i\overset{N_1}{\to}\infty}\frac{\|R_i^{1/2}(R_i - I)z_i\|^2}{\|R_i^{1/2}z_i\|^2} &= \lim_{i\overset{N_1}{\to}\infty}\frac{z_i^T(R_i^3 - 2R_i^2 + R_i)z_i}{z_i^T R_i z_i} = 0,\\ \lim_{i\overset{N_2}{\to}\infty}\frac{\|(R_i - I)z_i\|^2}{\|z_i\|^2} &= \lim_{i\overset{N_2}{\to}\infty}\frac{z_i^T(R_i^2 - 2R_i + I)z_i}{z_i^T z_i} = 0.\end{aligned}$$

Jelikož podle důsledku 16 platí $\|R_i\| \leq \overline{R}$ a $\|R_i^{-1}\| \leq 1/\underline{R}$, $i \in N$, můžeme obě tyto limity nahradit jedinou limitou

$$\lim_{i\to\infty} \frac{\|(R_i - I)z_i\|}{\|z_i\|} = \lim_{i\to\infty} \frac{\|(\tilde{G}_i^{-1/2} B_i \tilde{G}_i^{-1/2} - \tilde{G}_i^{-1/2}\tilde{G}_i\tilde{G}_i^{-1/2})z_i\|}{\|\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2} z_i\|} = \lim_{i\to\infty} \frac{\|(\tilde{G}_i^{-1/2}(B_i - \tilde{G}_i)d_i\|}{\|\tilde{G}_i^{1/2} d_i\|} = 0.$$

Protože $x_i \to x^*$ implikuje $G_i \to G_*$ a $\tilde{G}_i \to G_*$, dostaneme použitím předpokladů (F4) a (F5) vztah (430). □

Důsledek 17. *Nechť jsou splněny předpoklady věty 103, přičemž $\|B_i s_i + g_i\|/\|g_i\| \to 0$ a $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2a) a (S3a). Pak $x_i \to x^*$ superlineárně.*

Důkaz Toto tvrzení je bezprostředním důsledkem věty 103 a věty 20. □

Věta 103 a důsledek 17 vyžadují platnost vztahu (420). Tento vztah není přímým důsledkem aktualizací matic B_i, $i \in N$. Je třeba, aby tyto matice určovaly posloupnost x_i, $i \in N$ tak, že $B_i s_i = -g_i$, $i \in N$. Tento požadavek je obsažen v předpokladech věty 101.

Věta 104. *(Superlineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s proměnnou metrikou vyhovující předpokladům věty 101. Nechť jsou navíc splněny předpoklady lemmatu 51 (platí (6) a $\rho_i = 1$, $\gamma_i = 1$, $\beta_i \geq 0$, $i \in N$) a nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2a) a (S3a). Pak $x_i \to x^*$ superlineárně.*

Důkaz Protože jsou splněny předpoklady věty 101 a tedy i věty 102, platí (420), což spolu s předpoklady lemmatu 51 implikuje platnost věty 103. Lze tedy použít důsledek 17. □

Poznámka 174. Ve větě 104 předpokládáme, že platí $\beta_i \geq 0$ (neboli $\eta_i \leq 1$) $\forall i \in N$. Tento předpoklad nelze příliš zeslabit. Dá se pouze dokázat, že věta 104 zůstane v platnosti, pokud

$$\sum_{\substack{i=1 \\ \beta_i<0}}^{\infty} \frac{\beta_i}{\beta_i^*} < \infty.$$

Také je nutné, aby platilo $\rho_i = 1$ a $\gamma_i = 1$, v opačném případě nelze použít princip důkazu. Podrobnějším rozborem lze ukázat, že pro $\gamma_i \neq 1$ věta 104 neplatí a to zejména proto, že volba $\alpha_i = 1$ nemá při použití škálování žádné výsadní postavení.

Podle věty 104 jsou omezené metody s proměnnou metrikou z Broydenovy třídy (kdy $0 \leq \beta < 1$ nebo $0 \leq \eta < 1$) superlineárně konvergentní za předpokladů, které se příliš neliší od předpokladů zaručujících globální konvergenci (věta 101). Je však třeba, aby platilo $B_i s_i = -g_i$ (nebo $s_i = -H_i g_i$), $i \in N$, takže jsou vyloučeny metody, kdy se soustavy rovnic $B_i s_i = -g_i$, $i \in N$, řeší nepřesně metodou sdružených gradientů, nebo metody s lokálně omezeným krokem. Také je důležité, aby v rovnicích $B_i s_i = -g_i$, $i \in N$, vystupovaly matice B_i, $i \in N$, získané aktualizacemi z Broydenovy třídy, takže jsou vyloučeny metody s proměnnou metrikou pro separovatelné úlohy (oddíl 10.5). Proto dokážeme ještě větu o lokální konvergenci metod s proměnnoiu metrikou, která je použitelná i v uvedených nestandardních případech. Budeme přitom používat označení (418) a (422) a postup uvedený v práci [148] a použitý také v práci [76].
.

Lemma 52. *Nechť jsou splněny předpoklady lemmatu 51. Pak pro $i \in N$ platí*

$$\|R_{i+1}^*\| \leq \max(1, \|R_i^*\|)(1 + O(\|e_i\|)), \qquad \|(R_{i+1}^*)^{-1}\| \leq \max(1, \|(R_i^*)^{-1}\|)(1 + O(\|e_i\|)).$$

Důkaz Označíme-li

$$\tilde{R}_i = B_i^{1/2}\tilde{G}_i^{-1}B_i^{1/2}, \qquad \tilde{R}'_{i+1} = B_{i+1}^{1/2}\tilde{G}_i^{-1}B_{i+1}^{1/2}.$$

$$\tilde{R}_i^* = B_i^{1/2} G_*^{-1} B_i^{1/2}, \qquad \tilde{R}_{i+1}^* = B_{i+1}^{1/2} G_*^{-1} B_{i+1}^{1/2},$$

lze podobně jako v důkazu lemmatu 51 psát

$$\begin{aligned} \|R_{i+1}^* - R_{i+1}'\| &= \|\tilde{R}_{i+1}^* - \tilde{R}_{i+1}'\| = \|B_{i+1}^{1/2}(G_*^{-1} - \tilde{G}_{i+1}^{-1})B_{i+1}^{1/2}\| \le \|B_{i+1}\|\|G_*^{-1} - \tilde{G}_{i+1}^{-1}\| \\ &= \|\tilde{G}_{i+1}^{1/2} R_{i+1}' \tilde{G}_{i+1}^{1/2}\|\|G_*^{-1} - \tilde{G}_{i+1}^{-1}\| \le \overline{G}\|R_{i+1}'\|\|G_*^{-1} - \tilde{G}_{i+1}^{-1}\| = \|R_{i+1}'\|\, O(\|e_i\|), \end{aligned}$$

takže

$$\|R_{i+1}^*\| \le \|R_{i+1}'\| + \|R_{i+1}^* - R_{i+1}'\| \le \|R_{i+1}'\|(1 + O(\|e_i\|)).$$

Podobným způsobem dostaneme

$$\|R_i\| \le \|R_i^*\| + \|R_i - R_i^*\| \le \|R_i^*\|(1 + O(\|e_i\|)),$$

takže podle poznámky 153 platí

$$\|R_{i+1}^*\| \le \|R_{i+1}'\|(1 + O(\|e_i\|)) \le \max(1, \|R_i\|)(1 + O(\|e_i\|)) \le \max(1, \|R_i^*\|)(1 + O(\|e_i\|)),$$

což je první dokazovaná nerovnost. Stejný postup můžeme použít pro posloupnost matic R_i^{-1}, $i \in N$, čímž dostaneme druhou dokazovanou nerovnost. □

Věta 105. *(Lokální konvergence) Nechť bod $x^* \in R^n$ vyhovuje předpokladům věty 4 (postačujícím podmínkám pro lokální minimum) a v nějakém okolí bodu x^* platí (6). Uvažujme metodu spádových směrů takovou, že $\cos^2\theta_i \le 1/\kappa(B_i)$, kde B_i, $i \in N$, je posloupnost pozitivně definitních matic získaná aktualizacemi z Broydenovy třídy s $\rho_i = 1$, $\gamma_i = 1$ a $0 \le \beta_i \le 1$. Pak existuje číslo $\delta > 0$ takové, že pokud $\|e_1\| < \delta$, platí $x_i \to x^*$ a $\sum_{i=1}^{\infty} \|e_i\| < \infty$. Jestliže navíc $\|B_i s_i + g_i\|/\|g_i\| \to 0$ a $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2a) a (S3a), pak $x_i \to x^*$ superlineárně.*

Důkaz Protože bod $x^* \in R^n$ vyhovuje předpokladům věty 4 a platí předpoklad F6*, existuje číslo $\varepsilon > 0$ takové, že platí předpoklady F4*–F6*, pokud $x_i \in \mathcal{B}(x^*, \varepsilon)$ (poznámka 5 a poznámka **??**). V $\mathcal{B}(x^*, \varepsilon)$ jsou tedy splněny předpoklady lemmatu 51 a tedy i lemmatu 52.

(a) Označme $\tau_i = \max(1, \|R_i^*\|)$. Pak z první nerovnosti v lemmatu 52 plyne existence konstanty $C > 0$ takové, že pro $i \in N$ platí $\tau_{i+1} \le \tau_i(1 + (C/2)\|e_i\|)$. Můžeme tedy psát

$$\tau_i = \tau_1 \prod_{j=1}^{i-1} \left(1 + \frac{C}{2}\|e_j\|\right) \le \tau_1 \exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right)$$

(používáme první nerovnost v (41)), takže

$$\|B_i\| = \|G_*^{1/2} R_i^* G_*^{1/2}\| \le \overline{G}\|R_i^*\| \le \overline{G}\tau_i \le \overline{G}\tau_1 \exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right) \triangleq \overline{B}\exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right).$$

Stejným způsobem z druhé nerovnosti v lemmatu 52 dostaneme

$$\|H_i\| \le \overline{H}\exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right)$$

(konstantu C volíme tak velkou, aby byly splněny obě nerovnosti). Spojením těchto nerovností a použitím vztahu (43) dostaneme

$$\kappa_i = \kappa(B_i) = \|B_i\|\|H_i\| \le \overline{B}\,\overline{H}\exp\left(C\sum_{j=1}^{i-1}\|e_j\|\right) \le \overline{\kappa}\exp\left(C\sqrt{\frac{2}{\underline{G}}}\sum_{j=1}^{i-1}\sqrt{F_j - F^*}\right),$$

kde $\overline{\kappa} = \overline{B}\,\overline{H}$. Podle poznámky 37 a věty 18 tedy platí $x_i \to x^*$ a $\sum_{i=1}^{\infty} \|e_i\| < \infty$, pokud $x_1 \in \mathcal{B}(x^*, \delta)$, kde číslo δ je určeno vztahem (53).

(b) Podle (a) platí (420), takže jsou splněny předpoklady věty 103. Superlineární konvergence je pak bezprostředním důsledkem věty 103 a věty 20. □

Poznámka 175. Poněkud neobvyklý předpoklad $\cos^2 \theta_i \leq 1/\kappa(B_i)$, $i \in N$, znamená, že rovnici $B_i s_i = -g_i$ není nutné řešit přesně (poznámka 19).

4.7 Aktualizace trojúhelníkového rozkladu

Používáme-li inverzní metody s proměnnou metrikou (306), je třeba určovat směrový vektor řešením soustavy rovnic $Bs = -g$, kde B je symetrická pozitivně definitní matice. V tomto případě je výhodné pracovat s trojúhelníkovým rozkladem $B = LDL^T$, kde L je dolní trojúhelníková matice s jednotkami na hlavní diagonále a D je pozitivné definitní diagonální matice. Pak řešení soustavy rovnic $LDL^T s = -g$ vyžaduje $O(n^2)$ operací násobení a sčítání. Ukážeme nyní, jak lze určit trojúhelníkový rozklad matice $\bar{B} = B + \sigma z z^T$ z trojúhelníkového rozkladu matice B s použitím $O(n^2)$ operací násobení a sčítání. Použijeme přitom výsledky uvedené v práci [67].

Věta 106. *Nechť L, $\bar{L}$ jsou dolní trojúhelníkové matice s jednotkami na hlavní diagonále a D, $\bar{D}$ jsou pozitivně definitní diagonální matice, přičemž*

$$\bar{L}\bar{D}\bar{L}^T = LDL^T + \sigma z z^T. \tag{431}$$

Nechť l_i, $\bar{l}_i$, $1 \leq i \leq n$, jsou sloupce matic L, $\bar{L}$ a d_i, $\bar{d}_i$, $1 \leq i \leq n$, jsou diagonální prvky matic D, $\bar{D}$. Pak pro $1 \leq i \leq n$ platí

$$\begin{aligned} v_i &= z_{ii}, \\ \bar{d}_i &= d_i + \sigma_i v_i^2, \qquad (432) \\ \bar{l}_i &= \frac{d_i}{\bar{d}_i} l_i + \frac{\sigma_i v_i}{\bar{d}_i} z_i \qquad (433) \end{aligned}$$

(z_{ii} je i-tý prvek vektoru z_i), kde $\sigma_1 = \sigma$, $z_1 = z$ a pro $1 \leq i \leq n$ platí

$$\begin{aligned} \sigma_{i+1} &= \frac{d_i}{\bar{d}_i}\sigma_i, \qquad (434) \\ z_{i+1} &= z_i - v_i l_i. \qquad (435) \end{aligned}$$

Přitom v_i, $1 \leq i \leq n$, jsou prvky vektoru v, který je řešením soustavy rovnic $Lv = z$.

Důkaz Větu dokážeme indukcí. Předpokládejme, že pro nějaký index $i < n$ platí

$$\sum_{j=i}^{n} \bar{d}_j \bar{l}_j \bar{l}_j^T = \sum_{j=i}^{n} d_j l_j l_j^T + \sigma_i z_i z_i^T. \tag{436}$$

Zřejmě $\sigma_1 = \sigma$ a $z_1 = z$, neboť rovnost (431) lze zapsat ve tvaru

$$\sum_{j=1}^{n} \bar{d}_j \bar{l}_j \bar{l}_j^T = \sum_{j=1}^{n} d_j l_j l_j^T + \sigma z z^T.$$

Protože vektory l_j, $\bar{l}_j$, $i \leq j \leq n$, mají prvních $j-1$ prvků nulových, má matice (436) prvních $i-1$ sloupců nulových a její i-tý sloupec je určen vztahem $\bar{d}_i \bar{l}_i \bar{l}_{ii} = d_i l_i l_{ii} + \sigma_i z_i z_{ii}$, což spolu s $l_{ii} = 1$, $\bar{l}_{ii} = 1$ dává

$$\begin{aligned} \bar{d}_i &= d_i + \sigma_i z_{ii}^2, \qquad (437) \\ \bar{l}_i &= \frac{d_i}{\bar{d}_i} l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_i. \qquad (438) \end{aligned}$$

Vztah (438) můžeme ještě upravit. Položíme-li

$$z_{i+1} = z_i - z_{ii} l_i \tag{439}$$

a použijeme-li (437)–(439), dostaneme

$$\bar{d}_i \bar{l}_i = d_i l_i + \sigma_i z_{ii} z_i = \bar{d}_i l_i - \sigma_i z_{ii}^2 l_i + \sigma_i z_{ii} z_i = \bar{d}_i l_i + \sigma_i z_{ii} z_{i+1},$$

což dává

$$\bar{l}_i = l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_{i+1}. \tag{440}$$

Ukážeme nyní, že platí

$$d_i l_i l_i^T - \bar{d}_i \bar{l}_i \bar{l}_i^T + \sigma_i z_i z_i^T = \sigma_{i+1} z_{i+1} z_{i+1}^T, \tag{441}$$

kde číslo σ_{i+1} je určeno vzorcem (434). Použijeme-li vztahy (437)–(440), dostaneme

$$\begin{aligned}
& d_i l_i l_i^T - \bar{d}_i \bar{l}_i \bar{l}_i^T + \sigma_i z_i z_i^T \\
= \;& d_i l_i l_i^T - \bar{d}_i (l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_{i+1})(l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_{i+1})^T + \sigma_i (z_{i+1} + z_{ii} l_i)(z_{i+1} + z_{ii} l_i)^T \\
= \;& (d_i - \bar{d}_i) l_i l_i^T - \sigma_i z_{ii} (l_i z_{i+1}^T + z_{i+1} l_i^T) - \frac{\sigma_i^2 z_{ii}^2}{\bar{d}_i} z_{i+1} z_{i+1}^T \\
& + \sigma_i z_{i+1} z_{i+1}^T + \sigma_i z_{ii} (l_i z_{i+1}^T + z_{i+1} l_i^T) + \sigma_i z_{ii}^2 l_i l_i^T \\
= \;& \left(\sigma_i - \frac{\sigma_i^2 z_{ii}^2}{\bar{d}_i} \right) z_{i+1} z_{i+1}^T = \sigma_{i+1} z_{i+1} z_{i+1}^T,
\end{aligned}$$

kde

$$\sigma_{i+1} = \sigma_i - \frac{\sigma_i^2 z_{ii}^2}{\bar{d}_i} = \frac{\bar{d}_i - \sigma_i z_{ii}^2}{\bar{d}_i} \sigma_i = \frac{d_i}{\bar{d}_i} \sigma_i. \tag{442}$$

Platí tedy (441), kde číslo σ_{i+1} je určeno vzorcem (434). Porovnáme-li vztahy (436) a (441), získáme

$$\sum_{j=i+1}^{n} \bar{d}_i \bar{l}_i \bar{l}_i^T = \sum_{j=i+1}^{n} d_i l_i l_i^T + \sigma_i z_i z_i^T,$$

čímž jsme provedli indukční krok. Podle (437) a (438) tedy platí (432) a (433) a vzorce (442) a (439) jsou totožné se vzorci (434) a (435). Zbývá dokázat, že $Lv = z$, kde $v_i = z_{ii}$, $1 \leq i \leq n$. To je však velmi snadné, neboť podle (435) pro $1 \leq i \leq n$ platí

$$z_i = z - \sum_{j=1}^{i-1} v_j l_j,$$

což pro i-tý prvek $v_i = z_{ii}$ vektoru v dává stejný vzorec jako zpětný chod Gaussovy eliminační metody pro řešení soustavy rovnic $Lv = z$. □

Důsledek 18. *Nechť jsou splněny předpoklady věty 106. Pak pro* $1 \leq i \leq n$ *platí*

$$\begin{aligned}
v_i &= z_{ii}, \\
\bar{d}_i &= \frac{\tau_{i+1}}{\tau_i} d_i, && (443) \\
\bar{l}_i &= l_i + \frac{v_i}{\tau_{i+1} d_i} z_{i+1}, && (444)
\end{aligned}$$

kde $\tau_1 = 1/\sigma$, $z_1 = z$ *a pro* $1 \le i \le n$ *platí*

$$\tau_{i+1} = \tau_i + \frac{v_i^2}{d_i}, \tag{445}$$

$$z_{i+1} = z_i - v_i l_i \tag{446}$$

(přímé rekurence). Nechť vektor v *řešením soustavy rovnic* $Lv = z$. *Pak pro* $n \ge i \ge 1$ *platí*

$$\bar{d}_i = \frac{\tau_{i+1}}{\tau_i} d_i, \tag{447}$$

$$\bar{l}_i = l_i + \frac{v_i}{\tau_{i+1} d_i} z_{i+1}, \tag{448}$$

kde $\tau_{n+1} = 1/\sigma + v^T D^{-1} v$, $z_{n+1} = 0$ *a pro* $n \ge i \ge 1$ *platí*

$$\tau_i = \tau_{i+1} - \frac{v_i^2}{d_i}, \tag{449}$$

$$z_i = z_{i+1} + v_i l_i \tag{450}$$

(zpětné rekurence).

Důkaz Položme $\tau_i = 1/\sigma_i$, $1 \le i \le n$. Pak podle (432) a (434) pro $1 \le i \le n$ platí

$$\tau_{i+1} = \tau_i \frac{\bar{d}_i}{d_i} = \tau_i \frac{d_i + \sigma_i v_i^2}{d_i} = \tau_i + \frac{v_i^2}{d_i},$$

$$\bar{d}_i = \frac{\sigma_i}{\sigma_{i+1}} d_i = \frac{\tau_{i+1}}{\tau_i} d_i,$$

takže rekurentní vztahy (432)–(435) můžeme zapsat ve tvaru (443)–(446) (vzorec (444) plyne ze vzorce (440)). K odvození zpětných rekurencí aplikujeme důsledek 12, na matici (431). Dostaneme

$$\det \bar{B} = \det(LDL^T + \sigma z z^T) = \det L(D + \sigma v v^T) L^T = (1 + \sigma v^T D^{-1} v) \det B,$$

kde $Lv = z$ (jelikož trojúhelníková matice L má jednotky na hlavní diagonále, platí $\det L = 1$). Použijeme-li (443), můžeme psát

$$\frac{\tau_{n+1}}{\tau_1} = \prod_{i=1}^{n} \frac{\tau_{i+1}}{\tau_i} = \prod_{i=1}^{n} \frac{\bar{d}_i}{d_i} = \frac{\det \bar{D}}{\det D} = \frac{\det \bar{B}}{\det B} = 1 + \sigma v^T D^{-1} v,$$

což dává $\tau_{n+1} = 1/\sigma + v^T D^{-1} v$. Jelikož $\bar{l}_n = l_n = e_n$ (poslední prvek jednotkové matice řádu n), musí platit $z_{n+1} = 0$. Vztahy (449)–(450) dostaneme obrácením vztahů (445)–(446). □

Poznámka 176. Rekurentní vztahy (432)–(435) jsou nejpřirozenější, lze je však použít pouze tehdy, když $\sigma > 0$. V případě, že $\sigma < 0$, může vlivem zaokrouhlovacích chyb dojít ke ztrátě stability (prvky $\bar{d}_i$ mohou vycházet nulové nebo záporné). Proto se v tomto případě používají zpětné rekurence (447)–(450). Přímé rekurence (443)–(446), které lze použít pro $\sigma > 0$, mají tu výhodu, že jsou prakticky stejné jako zpětné rekurence (447)–(450), což umožňuje implementovat obě rekurence jedním algoritmem.

4.8 Modifikace a implementace metod s proměnnou metrikou

Nejprve se budeme zabývat úpravami, které umožní snížit počet operací v iteračních krocích metod s proměnnou metrikou a zvýšit tak jejich účinnost. Jednou z možností je odstranění maticového násobení při výpočtu směrového vektoru

$$s_+ = -H_+ g_+ \tag{451}$$

(tím lze snížit počet operací zhruba o čtvrtinu). Použijeme-li vztahy (270)–(271) a rovnost $d = \alpha s = -\alpha Hg$, můžeme psát

$$s_+ = -H_+ g_+ = -\gamma(H + UMU^T)g_+ = -\gamma(Hg + Hy + UMU^T g_+) = \frac{\gamma}{\alpha} d - \gamma(Hy + UMU^T g_+),$$

takže vektor s_+ lze spočítat pomocí vektorů d, Hy a sloupců matice U (v případě metod z Broydenovy třídy má matice U sloupce d, Hy). Protože vektor Hy známe z předchozího iteračního kroku, odpadá násobení maticí H_+ ve vztahu (451). Následující věta udává příslušné vzorce.

Věta 107. *Nechť H_+ je matice určená vztahem (286), kde $d = -\alpha Hg$, $\alpha > 0$ je délka kroku a matice H je pozitivně definitní. Pak vektor $s_+ = -H_+ g_+$ můžeme spočítat podle vzorce*

$$\frac{1}{\gamma} s_+ = \frac{\gamma}{\rho}\frac{\delta}{\alpha}\left(\frac{a}{b}d - Hy\right) - \frac{\rho}{\gamma}\frac{d^T g_+}{b} d = \frac{\gamma}{\rho}\frac{\delta}{\alpha}\left(\frac{a}{b}d - Hy\right) + \frac{\rho}{\gamma}\left(\frac{1}{\alpha}\frac{c}{b} - 1\right) d, \tag{452}$$

kde δ je číslo definované vztahem (300).

Důkaz Platí

$$d^T g_+ = d^T(y + g) = d^T y - d^T H^{-1} s = b - \frac{c}{\alpha}, \tag{453}$$

$$y^T H g_+ = y^T H(y + g) = y^T H y - y^T s = a - \frac{b}{\alpha}, \tag{454}$$

takže po dosazení do (286) dostaneme

$$\begin{aligned}
\frac{1}{\gamma} s_+ &= -\frac{1}{\gamma} H_+ g_+ = -Hy + \frac{1}{\alpha} d - \frac{\rho}{\gamma}\frac{d^T g_+}{b} d + \frac{1}{a}\left(a - \frac{b}{\alpha}\right) Hy \\
&\quad - \frac{\eta}{a}\left[\frac{a}{b}\left(b - \frac{c}{\alpha}\right) - \left(a - \frac{b}{\alpha}\right)\right]\left(\frac{a}{b}d - Hy\right) \\
&= \frac{1}{\alpha}\frac{b}{a}\left(\frac{a}{b}d - Hy\right) + \frac{\eta}{\alpha}\frac{b}{a}\left(\frac{ac - b^2}{b^2}\right)\left(\frac{a}{b}d - Hy\right) - \frac{\rho}{\gamma}\frac{d^T g_+}{b} d \\
&= \frac{1}{\alpha}\frac{b}{a}\left(1 + \frac{\eta(ac - b^2)}{b^2}\right)\left(\frac{a}{b}d - Hy\right) - \frac{\rho}{\gamma}\frac{d^T g_+}{b} d,
\end{aligned}$$

což spolu s (300) a (453) dává (452). □

Poznámka 177. Pro metodu DFP, kdy $\eta = 0$, dostaneme

$$s_+^{DFP} = \left[\frac{\gamma}{\alpha}\left(1 + \frac{\rho}{\gamma}\frac{c}{b}\right) - \rho\right] d - \frac{\gamma}{\alpha}\frac{b}{a} Hy. \tag{455}$$

Pro metodu BFGS, kdy $\eta = 1$, dostaneme

$$s_+^{BFGS} = \left[\frac{\gamma}{\alpha}\frac{c}{b}\left(\frac{\rho}{\gamma} + \frac{a}{b}\right) - \rho\right] d - \frac{\gamma}{\alpha}\frac{ac}{b^2} Hy. \tag{456}$$

Poznámka 178. Provádíme-li přesný výběr délky kroku, vymizí v (452) poslední člen, takže všechny aktualizace z Broydenovy třídy aplikované na matici H dávají rovnoběžné směrové vektory s_+. To je v souladu s tvrzením věty 77.

Vzorec (452) lze použít pouze tehdy, když $s = -Hg$. Pokud tento předpoklad neplatí (například u metod s lokálně omezeným krokem), nemusí být splněna podmínka spádovosti $s_+^T g_+ < 0$. Odvodíme ještě jeden vzorec, uvedený v práci [165], který má příznivější vlastnosti.

Věta 108. *Označme*

$$p = HVg_+, \qquad V = I - \frac{1}{b} y d^T. \tag{457}$$

Nechť H_+ je matice určená vztahem (286), kde $d = -\alpha Hg$, $\alpha > 0$ je délka kroku a matice H je pozitivně definitní. Pak vektor $s_+ = -H_+ g_+$ můžeme spočítat podle vzorce

$$-\frac{1}{\gamma} s_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \frac{b + \eta\alpha y^T p}{b + \alpha y^T p} V^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \delta \frac{\gamma b}{\rho c} V^T p, \tag{458}$$

kde $\alpha y^T p \geq 0$ a δ je číslo určené vztahem (300).

Důkaz (a) Použijeme-li rovnosti (453) a (457), dostaneme

$$\begin{aligned} p &= HVg_+ = H\left(I - \frac{1}{b} y d^T\right) g_+ = Hg + Hy - \frac{d^T g_+}{b} Hy \\ &= -\frac{1}{\alpha} d + Hy - \left(1 - \frac{c}{\alpha b}\right) Hy = \frac{1}{\alpha}\left(\frac{c}{b} Hy - d\right), \end{aligned} \tag{459}$$

takže $Hy = (\alpha p + d) b/c$ a jelikož $V^T d = 0$, platí

$$V^T Hy = \alpha \frac{b}{c} V^T p. \tag{460}$$

(b) Použijeme-li vyjádření (292), můžeme psát

$$\begin{aligned} -\frac{1}{\gamma} s_+ &= \frac{1}{\gamma} H_+ g_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + V^T \left(H + \frac{\eta - 1}{a} H y y^T H\right) V g_+ \\ &= \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + V^T p + \frac{\eta - 1}{a} V^T H y y^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \left(1 + \alpha \frac{\eta - 1}{a} \frac{b}{c} y^T p\right) V^T p. \end{aligned}$$

Ale $a = y^T Hy = y^T(\alpha p + d) b/c = (b + \alpha y^T p) b/c$, takže

$$-\frac{1}{\gamma} s_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \left(1 + \frac{\eta - 1}{b + \alpha y^T p} \alpha y^T p\right) V^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \frac{b + \eta\alpha y^T p}{b + \alpha y^T p} V^T p.$$

(c) Použijeme-li vztah (459) a nerovnost $ac - b^2 \geq 0$ (která plyne ze Schwarzovy nerovnosti), dostaneme

$$\alpha y^T p = b\left(\frac{ac - b^2}{b^2}\right) \geq 0.$$

Dosadíme-li výraz $\alpha y^T p$ do (180) a použijeme-li (300), dostaneme

$$\frac{b + \eta\alpha y^T p}{b + \alpha y^T p} = \frac{b^2 + \eta(ac - b^2)}{ac} = \delta \frac{\gamma b}{\rho c}.$$

□

Poznámka 179. Jelikož $V^T p = V^T HVg_+$, můžeme vzorec (458) zapsat ve tvaru

$$-\frac{1}{\gamma} s_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + V^T (\gamma^{BFGS} H) V g_+,$$

kde

$$\gamma^{BFGS} = \frac{b + \eta\alpha y^T p}{b + \alpha y^T p} = \delta \frac{\gamma b}{\rho c},$$

takže libovolná aktualizace z Broydenovy třídy dává stejný směrový vektor jako aktualizace BFGS škálovaná koeficientem γ^{BFGS}.

Poznámka 180. Neplatí-li $s = -Hg$, můžeme vzorec (458) upravit tak, že místo čísla $\alpha y^T p$ použijeme číslo $\tau = \max(\alpha y^T p, 0)$. Pak podle (458) platí

$$g_+^T s_+ = -\rho \frac{(d^T g_+)^2}{b} - \frac{b + \eta\tau}{b + \tau} g_+^T V^T H V g_+.$$

Pokud $d^T g_+ \neq 0$, dostaneme $g_+^T s_+ \leq -\rho (d^T g_+)^2 / b < 0$. Pokud $d^T g_+ = 0$, platí $V g_+ = g_+$, takže $g_+^T s_+ = -g_+^T H g_+ (b + \eta\tau)/(b + \tau) < 0$, neboť matice H je pozitivně definitní a $\tau \geq 0$.

Nyní popíšeme modifikaci metod s proměnnou metrikou v součinovém tvaru (poznámka 125), která používá ortogonální transformace umožňující značně zjednodušit použité aktualizace (tato modifikace je studována v pracech [143] a [144]). Nechť $H = SS^T$ a $\bar{S} = SQ^T$, kde Q je čtvercová ortogonální matice (takže $Q^T Q = QQ^T = I$). Pak

$$H = SS^T = SQ^T QS^T = \bar{S}\bar{S}^T,$$

takže matici S_+ lze získat aktualizacací matice $\bar{S}$. Matici Q volíme tak, aby tato aktualizace byla co nejjednodušší.

Věta 109. *Nechť $H = SS^T$, $d = S\tilde{d}$, $\tilde{y} = S^T y$, $a = \tilde{y}^T \tilde{y} > 0$, $b = \tilde{y}^T \tilde{d} > 0$, $c = \tilde{d}^T \tilde{d} > 0$. Nechť $\bar{S} = SQ^T$, kde Q je ortogonální matice taková, že vektor $Q\tilde{d}$ má pouze první prvek nenulový a vektor $Q\tilde{y}$ má pouze první dva prvky nenulové (tuto matici lze získat jako součin Givensových rotací sloužících k vynulování prvků uvedených vektorů). Nechť*

$$\begin{aligned}
\frac{1}{\sqrt{\gamma}} S_+ e_1 &= \sqrt{\frac{\rho}{\gamma b}}\, d \\
\frac{1}{\sqrt{\gamma}} S_+ e_2 &= \sqrt{\frac{(y^T \bar{s}_1)^2 + \eta (y^T \bar{s}_2)^2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2}} \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T \bar{s}_1} \bar{s}_1 \right) \\
\frac{1}{\sqrt{\gamma}} S_+ e_j &= \bar{s}_j, \qquad 3 \leq j \leq n,
\end{aligned}$$

kde $S_+ e_j$ je j-tý sloupec matice S_+ a $\bar{s}_j = \bar{S} e_j$, $1 \leq j \leq n$. Pak položíme-li $H_+ = S_+ S_+^T$, platí (286).

Důkaz Jelikož podle předpokladu platí $Q\tilde{d} = \lambda e_1$ a $Q\tilde{y} = \lambda_1 e_1 + \lambda_2 e_2$, kde λ, λ_1, λ_2 jsou vhodné koeficienty, můžeme psát

$$\begin{aligned}
d &= S\tilde{d} = SQ^T Q\tilde{d} = \lambda \bar{s}_1, \\
\bar{S}^T y &= QS^T y = Q\tilde{y} = \lambda_1 e_1 + \lambda_2 e_2,
\end{aligned}$$

takže $y^T \bar{s}_1 = \lambda_1$, $y^T \bar{s}_2 = \lambda_2$ a $y^T \bar{s}_j = 0$, $3 \leq j \leq n$. Platí tedy

$$\begin{aligned}
\left(I - \frac{dy^T}{y^T d} \right) \bar{s}_1 &= 0, \\
\left(I - \frac{dy^T}{y^T d} \right) \bar{s}_2 &= \bar{s}_2 - \frac{y^T \bar{s}_2}{y^T \bar{s}_1} \bar{s}_1, \\
\left(I - \frac{dy^T}{y^T d} \right) \bar{s}_j &= \bar{s}_j, \quad 3 \leq j \leq n
\end{aligned}$$

a

$$Hy = \bar{S}\bar{S}^T y = \lambda_1 \bar{s}_1 + \lambda_2 \bar{s}_2 = y^T \bar{s}_1 \, \bar{s}_1 + y^T \bar{s}_2 \, \bar{s}_2,$$

což po dosazení dává

$$\begin{aligned}\frac{Hy}{y^THy} - \frac{d}{y^Td} &= \frac{y^T\bar{s}_1\,\bar{s}_1 + y^T\bar{s}_2\,\bar{s}_2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2} - \frac{\bar{s}_1}{y^T\bar{s}_1} \\ &= \frac{y^T\bar{s}_2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}\bar{s}_2 + \frac{y^T\bar{s}_1}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}\bar{s}_1 - \frac{1}{y^T\bar{s}_1}\bar{s}_1 \\ &= \frac{y^T\bar{s}_2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right).\end{aligned}$$

Nyní použijeme vztah (293), podle kterého platí

$$\begin{aligned}\frac{1}{\gamma}H_+ &= \left(I - \frac{dy^T}{y^Td}\right)\bar{S}\bar{S}^T\left(I - \frac{dy^T}{y^Td}\right)^T + \frac{\rho}{\gamma}\frac{dd^T}{y^Td} \\ &\quad + y^THy(\eta-1)\left(\frac{d}{y^Td} - \frac{Hy}{y^THy}\right)\left(\frac{d}{y^Td} - \frac{Hy}{y^THy}\right)^T \\ &= \left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)^T + \sum_{j=3}^{n}\bar{s}_j\bar{s}_j^T + \frac{\rho}{\gamma}\frac{dd^T}{y^Td} \\ &\quad + (\eta-1)\frac{(y^T\bar{s}_2)^2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)^T \\ &= \sum_{j=3}^{n}\bar{s}_j\bar{s}_j^T + \frac{(y^T\bar{s}_1)^2 + \eta(y^T\bar{s}_2)^2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)\left(\bar{s}_2 - \frac{y^T\bar{s}_2}{y^T\bar{s}_1}\bar{s}_1\right)^T + \frac{\rho}{\gamma}\frac{dd^T}{y^Td} \\ &= \frac{1}{\gamma}\sum_{j=1}^{n}S_+e_j(S_+e_j)^T = \frac{1}{\gamma}S_+S_+^T.\end{aligned}$$

□

Poznámka 181. Z věty 109 plyne, že stačí aktualizovat pouze dva sloupce matice S. Většina operací se tedy spotřebovává na výpočet matice $\bar{S}$. To však jsou ortogonální transformace, které jsou velmi stabilní. Poznamenejme, že ve vzorci pro $(1/\gamma)S_+e_2$ se vyskytuje odmocnina z výrazu, který je kladný pokud $\eta \geq 0$ (pro metodu BFGS je tento výraz jednotkový).

Další modifikace metod s proměnnou metrikou používají různá zobecnění kvazinewtonowské podmínky. Jednu takovou možnost jsme již popsali v souvislosti s korekcí kvadratického modelu, kdy se kvazinewtonovská podmínka $H_+y = d$ nahradila podmínkou $H_+y = \rho d$. Nyní budeme vyšetřovat metody splňující zobecněnou kvazinewtonovskou podmínku $H_+\tilde{y} = d$ (nebo $B_+d = \tilde{y}$), kde $\tilde{y} = y + \tau d$. Tento princip lze použít k zajištění globální konvergence (poznámka 171) a také ke korekci kvadratického modelu, jak je ukázáno v práci [172] a v následující poznámce.

Poznámka 182. Nechť $B_+d = \tilde{y}$, kde $\tilde{y} = y + \tau d$. Pak $d^TB_+d = d^T\tilde{y} = d^Ty + \tau d^Td$, takže podmínka (397) je splněna právě tehdy, když

$$\tau = \left(\frac{1}{\rho} - 1\right)\frac{d^Ty}{d^Td}.$$

Abychom dostali korekci jako v poznámce 162, stačí položit $\tau\|d\|^2 = 2(F - F_+) + d^Tg_+ + d^Tg$. Abychom dostali korekci jako v poznámce 163, stačí položit $\tau\|d\|^2 = 6(F - F_+) + 3(d^Tg_+ + d^Tg)$ (což je trojnásobek předchozí korekce). Také se používá střední hodnota $\tau\|d\|^2 = 4(F - F_+) + 2(d^Tg_+ + d^Tg)$ Poznamenejme, že je třeba aby platilo $d^T\tilde{y} > 0$, což je nutné k tomu aby matice B_+ byla pozitivné definitní. Tato podmínka je splněna pokud $\tau > -d^Ty/d^Td$, což odpovídá hodnotě $\rho > 0$.

Kvazinewtonovská podmínka $H_+y = d$ se odvozuje z vlastností sečny. Tuto podmínku splňuje matice (375), nikoliv Hessova matice G_+. Abychom dostali lepší aproximaci matice G_+, je třeba vyjít z vlastností

tečny [59]. Tato možnost spočívá v použití gradientů minimalizované funkce vypočtených ve více bodech, například v bodech x_-, x, x_+, kde x_- je vektor z předchozího iteračního kroku. Těmito body se pomocí kvadratické interpolace proloží křivka

$$x(t) = at^2 + bt + c \tag{461}$$

taková, že $x(t_-) = x_-$, $x(t) = x$, $x(t_+) = x_+$. Podobná křivka se proloží gradienty g_-, g, g_+. Podle věty o střední hodnotě platí

$$g(x(t+\delta)) - g(x(t)) = \int_0^1 G((x(t+\lambda\delta)\mathrm{d}\lambda\ (x(t+\delta)) - x(t)),$$

což po vydělení číslem δ a přechodem k limitě dává $g'(x(t)) = G(x(t))x'(t)$, kde čárka označuje derivování. Odtud plyne, že vhodná tečná kvazinewtonovská podmínka má tvar

$$H_+ g'(x(t_+)) = x'(t_+). \tag{462}$$

Lemma 53. *Uvažujme křivku (461), kde vektory a, b, c jsou vybrány tak, že $x(t_-) = x_-$, $x(t) = x$, $x(t_+) = x_+$. Nechť $d = x_+ - x$ a $d_- = x - x_-$. Pak platí*

$$\frac{\tau+1}{\tau+2}(t_+ - t)x'(t_+) = d - \frac{1}{\tau(\tau+2)}d_- \tag{463}$$

kde $\tau = (t - t_-)/(t_+ - t)$.

Důkaz Podle (461) platí

$$\frac{d}{t_+ - t} = \frac{x_+ - x}{t_+ - t} = a(t_+ + t) + b, \qquad \frac{d_-}{t - t_-} = \frac{x - x_-}{t - t_-} = a(t + t_-) + b.$$

Odečtením těchto rovností dostaneme

$$\frac{d}{t_+ - t} - \frac{d_-}{t - t_-} = a(t_+ - t_-) = a(t_+ - t)(1 + \tau). \tag{464}$$

Jelikož $x'(t) = 2at + b$, můžeme psát

$$x'(t_+) = 2at_+ + b = a(t_+ + t) + b + a(t_+ - t) = \frac{d}{t_+ - t} + a(t_+ - t),$$

což spolu s (464) dává

$$(1+\tau)x'(t_+) = (1+\tau)\frac{d}{t_+ - t} + \left(\frac{d}{t_+ - t} - \frac{d_-}{t - t_-}\right) = (2+\tau)\frac{d}{t_+ - t} - \frac{d_-}{t - t_-},$$

odkud již snadnou úpravou dostaneme (463). □

Poznámka 183. Použijeme-li stejný interpolační polynom pro gradientní křivku $g(t)$, dostaneme vzorec (463), kde místo vektorů d, d_- a $x'(t_+)$ vystupují vektory y, y_- a $g'(t_+)$. Jelikož rovnost (462) lze vynásobit libovolným nenulovým číslem, můžeme tečnou kvazinewtonovskou podmínku zapsat ve tvaru

$$H_+ (y - \lambda y_-) = d - \lambda d_-, \qquad \lambda = \frac{1}{\tau(\tau+2)}. \tag{465}$$

Jelikož číslo τ se zachovává při lineárních transformacích parametru t, lze volit $t_+ = 1$, $t = 0$, $t_- = -\tau$. Pak platí (465), kde nyní $\tau = -t_-$.

Jednotlivé metody používající kvazinewtonovskou podmínku (465) se liší pouze výběrem parametru τ. Jednou z možností je volit ekvidistantní rozložení interpolačních uzlů. Pak $\tau = 1$, což vede na kvazinewtonovskou podmínku (465), kde $\lambda = 1/3$. Tato volba není příliš vhodná, neboť nebere v úvahu délky jednotlivých kroků. Výhodnější je volit parametr τ jako podíl norem vektorů d_- a d. Použití eukleidovských norem poněkud zlepšuje účinnost metod s proměnnou metrikou s počátečním škálováním, ale lepší výsledky dává hodnota $\tau = d_-^T B d_- / d^T B d$, kterou lze aproximovat výrazem $\tau = b_-/c$, kde $b_- = y_-^T d_-$ a $c = d^T B d = -\alpha d^T g$. V literatuře existuje celá řada dalších heuristických předpisů jak volit parametr τ, ale ty již účinnost metod s proměnnou metrikou příliš nezvyšují. Poznamenejme, že metody používající kvazinewtonovskou podmínku (465) mohou mít problém se zajištěním pozitivní definitnosti matice H_+, neboť nutná podmínka $(d - \lambda d_-)^T (y - \lambda y_-) > 0$ nemusí být splněna a její platnost je třeba ověřovat. Pokud není tato podmínka splněna, pokládáme $\lambda = 0$.

Závěrem popíšeme základní algoritmus, kterým se realizují metody s proměnnou metrikou z Broydenovy třídy a jejich modifikace. Nejprve uvedeme několik poznámek k implementaci jednotlivých kroků tohoto algoritmu.

(1) Výběr délky kroku: Metody s proměnnou metrikou nejsou citlivé na výběr délky kroku. Je možné použít algoritmus 1 beze změny. Volí se počáteční odhad $\alpha = 1$ nebo (zejména v počátečních iteracích) $\alpha = \min\left(1, 4(\underline{F} - F_i)/s_i^T g_i\right)$.

(2) Korekce (parametr ρ): Metody s proměnnou metrikou nejsou citlivé na volbu korekce, obvykle stačí pokládat $\rho = 1$. V případě, že provádíme škálování, se vyplácí volit některou z hodnot (398), (399), (401)) upravenou tak, aby platilo $\underline{\rho} \le \rho \le \underline{\rho}$.

(3) Škálování (parametr γ): Vhodné škálování značně zvyšuje účinnost metod s proměnnou metrikou, pokud volíme parametr γ tak, aby platilo $b/c \le \rho/\gamma \le a/b$. Škálování v každé iteraci však není účelné, je třeba používat nějakou strategii, která omezuje použití hodnoty $\gamma \neq 1$ v těch iteracích, kde je to nevhodné. Nejvíce se osvědčilo řízené nebo intervalové škálování popsané v poznámce 167.

(4) Výběr konkrétníi metody (parametr η): Praktické zkušenosti ukázují, že z jednoduchých metod je nejúčinnější metoda BFGS a že metoda DFP je velmi špatná. Ačkoliv metodu BFGS lze překonat některými složitějšími metodami, například metodou, která používá hodnotu $\eta = \eta^{VL+}$, kde

$$\eta^{VL+} = \frac{\max(0, \sqrt{c/a} - b^2/(ac))}{1 - b^2/(ac)}, \qquad b^2/(ac) < 1, \tag{466}$$

$$\eta^{VL+} = 1, \qquad b^2/(ac) \ge 1 \tag{467}$$

(poznámka 161), korekce a škálování rozdíly mezi nimi stírají (s celkovým zlepšením účinnosti), takže lze doporučit korigovanou a škálovanou metodu BFGS.

(5) Volba modifikace. Pokud neškálujeme nebo provádíme pouze počáteční škálování, zvyšují některé modifikace (například modifikace popsaná v poznámce 183) účinnost metod s proměnnou metrikou. Při použití řízeného škálování se tyto rozdíly stírají a naopak základní metody s proměnnou metrikou dávají lepší výsledky. Výjimku tvoří modifikace použitá ve větě 107, která šetří numerické operace a snižuje tak režiji metod s proměnnou metrikou z Broydenovy třídy.

Algoritmus metody s proměnnou metrikou lze popsat zhruba takto:

Algoritmus 8. Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\underline{\varepsilon} > 0$, $\underline{\rho} = 0.01$, $\overline{\rho} = 100$, $\underline{\gamma} = 0.7$, $\overline{\gamma} = 6$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme $F_1 := F(x_1)$, $g_1 := g(x_1)$. Zvolíme počáteční symetrickou pozitivně definití matici H_1 (obvykle $H_1 := I$) a položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \le \underline{\varepsilon}$ ukončíme výpočet. V opačném případě položíme $s_i := -H_i g_i$ a určíme délku kroku α_i použitím algoritmu 1. Položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$.

Krok 3 Položíme $d_i := x_{i+1} - x_i$ a $y_i := g_{i+1} - g_i$. Určíme parametr ρ_i tak, že položíme $\rho_i = 1$ nebo použijeme některou z hodnot (398), (399), (401). Jestliže $\rho_i < \underline{\rho}$ nebo $\rho_i > \overline{\rho}$ položíme $\rho_i := 1$. Použijeme řízené nebo intervalové škálování (poznámka 167) s hodnotou γ_i takovou, že $b_i/a_i \leq \gamma_i/\rho_i \leq c_i/b_i$ a mezemi $\underline{\gamma} \leq \gamma \leq \overline{\gamma}$ (pro metodu BFGS volíme $\gamma_i/\rho_i = b_i/a_i$). Zvolíme parametr $\eta_i > 0$ a určíme matici H_{i+1} podle (286) (pro metodu BFGS volíme $\eta_i = 1$).

Krok 4 Zvětšíme i o 1 a přejdeme na krok 2.

4.9 Davidonova třída metod s proměnnou metrikou

Zatím jsme se zabývali metodami s proměnnou metrikou patřícími do Broydenovy třídy. Nyní popíšeme Davidonovu třídu metod s proměnnou metrikou studovanou v pracech [36] a [92]. Tato třída používá směrové vektory $s_i = -H_i g_i$, kde H_i, $i \in N$, jsou symetrické pozitivně definitní matice konstruované podle rekurentního vztahu

$$H_{i+1} = H_i + \tilde{U}_i \tilde{M}_i \tilde{U}_i^T, \tag{468}$$

kde $\tilde{U}_i = [u_i, d_i - H_i y_i]$ a $\tilde{M}_i \in R^{2\times 2}$, a vyhovující podmínce

$$H_{i+1} y_i = d_i, \tag{469}$$

kde $y_i = g_{i+1} - g_i$ a $d_i = x_{i+1} - x_i$. Vektory $u_i \in R^n$, $i \in N$, se konstruují rekurentně tak, aby platilo

$$u_{i+1} \in \mathcal{L}(\tilde{U}_i), \qquad u_{i+1}^T y_i = 0 \tag{470}$$

(vektor u_{i+1} je tedy lineární kombinací vektorů u_i a $d_i - H_i y_i$ a je kolmý na vektor y_i). Tuto podmínku splňuje vektor

$$u_{i+1} = y_i^T (d_i - H_i y_i) u_i - y_i^T u_i (d_i - H_i y_i) \tag{471}$$

(který budeme v dalším výkladu používat) a každý jeho násobek.

Věta 110. *(Kvadratické ukončení) Nechť x_i, $i \in N$, je posloupnost generovaná metodou s proměnnou metrikou z Davidonovy třídy s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0\ \forall i \in N$) aplikovaná na ryze konvexní kvadratickou funkci*

$$Q(x) = \frac{1}{2}(x - x^*)^T G (x - x^*).$$

Nechť $g_i \neq 0$, $1 \leq i \leq n$. Pak $g_{n+1} = 0$ a $x_{n+1} = x^*$.

Důkaz Důkaz této věty je velmi podobný důkazu věty (74). Opět se indukcí pro $1 \leq i \leq n$ dokazují vztahy (273)–(276) a navíc vztah

$$u_{i+1}^T y_j = 0, \quad 1 \leq j \leq i. \tag{472}$$

Indukční předpoklad $u_i^T y_j = 0$, $1 \leq j < i$, se používá v části (a) k důkazu toho, že $\tilde{U}_i^T y_j = 0$, $1 \leq j < i$. Indukční krok pro (472) je velmi jednoduchý. Jelikož $u_{i+1} \in \mathcal{L}(\tilde{U}_i)$ a $\tilde{U}_i^T y_j = 0$, $1 \leq j < i$, platí $u_{i+1}^T y_j = 0$, $1 \leq j < i$. Protože podle (470) platí $u_{i+1}^T y_i = 0$, dostaneme $u_{i+1}^T y_j = 0$, $1 \leq j \leq i$. □

Věta 110 neposkytuje nic nového, co by nesplňovala i jednodušší Broydenova třída metod s proměnnou metrikou. Následující věta, uvedená v [138], však ukazuje, že za jistých předpokladů má Davidonova třída vlastnost kvadratického ukončení i bez přesného výběru délky kroku.

Věta 111. *Nechť x_i, $i \in N$, je posloupnost generovaná metodou s proměnnou metrikou z Davidonovy třídy aplikovaná na ryze konvexní kvadratickou funkci $Q(x)$. Pokud $y_i^T u_i \neq 0$ pro $1 \leq i \leq n$, platí $H_{n+1} = G^{-1}$.*

Důkaz Označme $\mathcal{Z}_i$, $i \in N$, podprostor vektorů $z \in R^n$ splňujících podmínky

$$GH_i z = z, \qquad u_i^T z = 0.$$

Dokážeme indukcí, že $\dim\mathcal{Z}_{n+1} = n$, takže $GH_{n+1} = I$ neboli $H_{n+1} = G^{-1}$. Předpokládejme, že pro nějaký index $1 < i \leq n$ platí $\dim\mathcal{Z}_i \geq i-1$, (platí to pro i = 2, neboť s použitím (468), (469) a (470) dostaneme $GH_2 y_1 = Gd_1 = y_1$ a $u_2^T y_1 = 0$, takže $y_1 \in \mathcal{Z}_2$ a tedy $\dim\mathcal{Z}_2 \geq 1$). Nechť $z \in \mathcal{Z}_i$. Jelikož $GH_i z = z$, můžeme psát

$$(d_i - H_i y_i)^T z = (d_i - H_i G d_i)^T z = d_i^T (z - GH_i z) = 0,$$

což spolu s $u_i^T z = 0$ dává $\tilde{U}_i^T z = 0$. Podle (468) a (470) pak platí $GH_{i+1} z = GH_i z = z$ a $u_{i+1}^T z = 0$, takže $z \in \mathcal{Z}_{i+1}$ a tedy $\mathcal{Z}_i \subset \mathcal{Z}_{i+1}$. Dále s použitím (469) a (470) dostaneme $GH_{i+1} y_i = Gd_i = y_i$ a $u_{i+1}^T y_i = 0$, takže $y_i \in \mathcal{Z}_{i+1}$. Jelikož předpokládáme, že $u_i^T y_i \neq 0$, nemůže platit $y_i \in \mathcal{Z}_i$, takže $\dim\mathcal{Z}_{i+1} \geq \dim\mathcal{Z}_i + 1 \geq i$. □

Poznámka 184. Ve větě 111 předpokládáme, že $y_i^T u_i \neq 0$, $1 \leq i \leq n$. Pokud $y_i^T u_i = 0$, platí pro tento index pouze $\dim\mathcal{Z}_{i+1} \geq \dim\mathcal{Z}_i$. Nicméně dimenze podprostoru $\mathcal{Z}_{i+1}$ se nemůže snížit a po n krocích splňujících podmínku $y_i^T u_i \neq 0$ platí $H_{i+1} = G^{-1}$.

Při vyšetřování aktualizací metod s proměnnou metrikou z Davidonovy třídy budeme index i vynechávat a index $i+1$ nahradíme symbolem $+$. Budeme přitom používat označení

$$\tilde{\alpha} = y^T u, \qquad \tilde{\beta} = y^T (d - Hy) = b - a, \tag{473}$$

takže vztah (471) zapíšeme ve tvaru

$$u_+ = \tilde{\beta} u - \tilde{\alpha}(d - Hy). \tag{474}$$

Poznamenejme, že čísla $\tilde{\alpha}$ a $\tilde{\beta}$ nemají nic společného s parametrem délky kroku α a parametrem β ve vzorci (306). Nejprve budeme vyšetřovat aktualizace z Davidonovy třídy zapsané ve tvaru použitém v práci [92].

Věta 112. *Nechť $H_+ = H + \tilde{U}\tilde{M}\tilde{U}^T$, kde H je symetrická pozitivně definitní matice a $\tilde{U} = [u, d - Hy]$, přičemž $y^T(d - Hy) \neq 0$. Pak rovnost $H_+ y = d$ platí právě tehdy, když*

$$\tilde{M} = \begin{bmatrix} -\varphi\tilde{\beta}, & \varphi\tilde{\alpha} \\ \varphi\tilde{\alpha}, & \frac{1}{\tilde{\beta}}(1 - \varphi\tilde{\alpha}^2) \end{bmatrix}, \tag{475}$$

kde $\varphi = -\det\tilde{M}$ je volný parametr.

Důkaz Označme $\tilde{m}_1$, $\tilde{m}_2$, $\tilde{m}_3$ prvky matice $\tilde{M}$. Platí-li (469), můžeme podle (468) psát

$$\begin{aligned} H_+ y - d &= -(d - Hy) + [u, d - Hy] \begin{bmatrix} \tilde{m}_1, & \tilde{m}_2 \\ \tilde{m}_2, & \tilde{m}_3 \end{bmatrix} \begin{bmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{bmatrix} \\ &= -(d - Hy) + (\tilde{m}_1\tilde{\alpha} + \tilde{m}_2\tilde{\beta})u + (\tilde{m}_2\tilde{\alpha} + \tilde{m}_3\tilde{\beta})(d - Hy) = 0, \end{aligned}$$

takže nutně

$$\begin{aligned} \tilde{m}_1\tilde{\alpha} + \tilde{m}_2\tilde{\beta} &= 0, \\ \tilde{m}_2\tilde{\alpha} + \tilde{m}_3\tilde{\beta} &= 1. \end{aligned}$$

Jeden parametr je nadbytečný. Zvolíme $\tilde{m}_2 = \varphi\tilde{\alpha}$ a zbylé prvky $\tilde{m}_1, \tilde{m}_3$ určíme řešením uvedených rovnic, takže

$$\tilde{m}_1 = -\varphi\tilde{\beta}, \qquad \tilde{m}_2 = \varphi\tilde{\alpha}, \qquad \tilde{m}_3 = \frac{1}{\tilde{\beta}}(1 - \varphi\tilde{\alpha}^2). \tag{476}$$

Tím dostaneme matici $\tilde{M}$ uvedenou ve větě 112. Z druhé strany, vynásobíme-li (468), kde matice $\tilde{M}$ je dána vztahem (475), vektorem y, dostaneme (469). □

Poznámka 185. Vztah $H_+ = H + \tilde{U}\tilde{M}\tilde{U}^T$ můžeme roznásobit. Pak platí

$$H_+ = H + \frac{1}{\tilde{\beta}}(d - Hy)(d - Hy)^T - \frac{\varphi}{\tilde{\beta}}\, u_+ u_+^T, \tag{477}$$

kde $u_+ = \tilde{\beta}u - \tilde{\alpha}(d - Hy)$. Zvolíme-li $\varphi = 0$, dostaneme metodu hodnosti 1 (vzorec (289) s $\gamma = 1$ a $\rho = 1$), která patří do Davidonovy třídy. Davidonovu třídu lze tedy chápat jako zobecnění metody hodnosti 1. Podstatné je, že Davidonova třída obsahuje aktualizace, které zajišťují pozitivní definitnost matice H_+ lépe než metoda R1.

Vyšetříme nyní, pro které hodnoty parametru φ je matice H_+ pozitivně definitní. Budeme přitom používat označení

$$\tilde{U}^T H^{-1} \tilde{U} = \begin{bmatrix} \tilde{a} & \tilde{b} \\ \tilde{b} & \tilde{c} \end{bmatrix}, \tag{478}$$

takže $\tilde{a} = u^T H^{-1} u,\ \tilde{b} = u^T H^{-1}(d - Hy),\ \tilde{c} = (d - Hy)^T H^{-1}(d - Hy) = a - 2b + c$. Dále označíme

$$\begin{aligned} A &= \tilde{\beta}^2(\tilde{a}\tilde{c} - \tilde{b}^2), \\ B &= \tilde{\beta}(\tilde{\beta} + \tilde{c})(\tilde{a}\tilde{c} - \tilde{b}^2), \\ C &= (\tilde{\beta} + \tilde{c})^2(\tilde{a}\tilde{c} - \tilde{b}^2), \\ D &= (\tilde{\beta}\tilde{b} - \tilde{\alpha}\tilde{c})^2, \end{aligned} \tag{479}$$

takže

$$D = \det{}^2(\tilde{U}^T H^{-1} U) = \left(\det \begin{bmatrix} \tilde{\alpha} + \tilde{b} & \tilde{\beta} + \tilde{c} \\ \tilde{b} & \tilde{c} \end{bmatrix}\right)^2, \tag{480}$$

kde $U = [d, Hy]$, a

$$B = \frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}} A, \qquad C = \frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}} B, \qquad A - 2B + C = \tilde{c}^2(\tilde{a}\tilde{c} - \tilde{b}^2), \qquad AC - B^2 = 0. \tag{481}$$

Lemma 54. *Nechť $H_+ = H + \tilde{U}\tilde{M}\tilde{U}^T$, kde H je symetrická pozitivně definitní matice, $\tilde{U} = [u, d - Hy]$ a $\tilde{M}$ je matice určená vztahem (475), přičemž $\tilde{\beta} \neq 0$ a $\tilde{a}\tilde{c} - \tilde{b}^2 > 0$. Pak matice $H^{-1/2} H_+ H^{-1/2}$ má $n - 2$ jednotkových vlastních čísel a zbylá dvě vlastní čísla $\underline{\lambda} \leq \overline{\lambda}$ jsou řešením kvadratické rovnice*

$$\lambda^2 - \sigma\lambda + \delta = 0, \tag{482}$$

kde

$$\sigma = \frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}} - \frac{\varphi}{\tilde{\beta}\tilde{c}}(A + D) + 1, \tag{483}$$

$$\delta = \frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}} - \frac{\varphi}{\tilde{\beta}\tilde{c}}(B + D). \tag{484}$$

Navíc platí $\underline{\lambda}'/\sigma' \geq 0$, $\overline{\lambda}'/\sigma' \geq 0$, kde $\underline{\lambda}'$, $\overline{\lambda}'$ jsou derivace kořenů $\underline{\lambda}$, $\overline{\lambda}$ rovnice (482) podle parametru φ a σ' je derivace čísla (483) podle parametru φ.

Důkaz. (a) Stejným způsobem jako v důkazu lemmatu 32 se ukáže, že matice

$$H^{-1/2} H_+ H^{-1/2} = I + H^{-1/2}\tilde{U}\tilde{M}\tilde{U}^T H^{-1/2}$$

má $n - 2$ jednotkových vlastních čísel a zbylá dvě vlastní čísla jsou řešením rovnice

$$\det((1 - \lambda)I + \tilde{M}\tilde{U}^T H^{-1}\tilde{U}) = 0.$$

Dosadíme-li do této rovnice vyjádření (475) a (478), dostaneme po úpravě

$$\det\begin{bmatrix} 1-\varphi(\tilde\beta\tilde a-\tilde\alpha\tilde b)-\lambda, & -\varphi(\tilde\beta\tilde b-\tilde\alpha\tilde c) \\ \dfrac{\varphi\tilde\alpha}{\tilde\beta}(\tilde\beta\tilde a-\tilde\alpha\tilde b)+\dfrac{\tilde b}{\tilde\beta}, & \dfrac{\varphi\tilde\alpha}{\tilde\beta}(\tilde\beta\tilde b-\tilde\alpha\tilde c)+\dfrac{\tilde\beta+\tilde c}{\tilde\beta}-\lambda \end{bmatrix}=0\,,$$

což dává rovnici (482) s koeficienty

$$\begin{aligned}\sigma &= \frac{\tilde\beta+\tilde c}{\tilde\beta}-\frac{\varphi}{\tilde\beta}\,(\tilde\alpha^2\tilde c-2\tilde\alpha\tilde\beta\tilde b+\tilde\beta^2\tilde a)+1\\ \delta &= \frac{\tilde\beta+\tilde c}{\tilde\beta}-\frac{\varphi}{\tilde\beta}\,(\tilde\alpha^2\tilde c-2\tilde\alpha\tilde\beta\tilde b+\tilde\beta^2\tilde a+\tilde\beta(\tilde a\tilde c-\tilde b^2)).\end{aligned}$$

Stejný výsledek dostaneme, dosadíme-li (479) do výrazů (483) a (484).

(b) Zbývá dokázat, že $\underline{\lambda}'/\sigma'\ge 0$ a $\overline{\lambda}'/\sigma'\ge 0$. Kořeny rovnice (482) lze (podobně jako v důkazu lemmatu 44) vyjádřit ve tvaru

$$\underline{\lambda}=\frac{\sigma}{2}-\sqrt{\Big(\frac{\sigma}{2}\Big)^2-\delta},\qquad \overline{\lambda}=\frac{\sigma}{2}+\sqrt{\Big(\frac{\sigma}{2}\Big)^2-\delta}.$$

Nechť $\underline{\lambda}\ne\overline{\lambda}$. Derivováním (482) podle parametru φ dostaneme

$$2\lambda\lambda'-\sigma'\lambda-\sigma\lambda'+\delta'=0,$$

což po úpravě dává

$$\frac{\lambda'}{\sigma'}=\frac{1}{2\lambda-\sigma}\left(\lambda-\frac{\delta'}{\sigma'}\right)=\frac{1}{2}\left(1+\frac{\frac{\sigma}{2}-\frac{\delta'}{\sigma'}}{\lambda-\frac{\sigma}{2}}\right).$$

Jelikož pro $\lambda=\underline{\lambda}$ a $\lambda=\overline{\lambda}$ platí $|\lambda-\sigma/2|=\sqrt{(\sigma/2)^2-\delta}$, je zřejmé, že $\underline{\lambda}'/\sigma'\ge 0$ a $\overline{\lambda}'/\sigma'\ge 0$, pokud je splněna nerovnost

$$\left|\frac{\sigma}{2}-\frac{\delta'}{\sigma'}\right|\le\left|\lambda-\frac{\sigma}{2}\right|=\sqrt{\Big(\frac{\sigma}{2}\Big)^2-\delta},$$

což po úpravě dává

$$\delta(\sigma')^2-\sigma\sigma'\delta'+(\delta')^2\le 0. \tag{485}$$

Použijeme-li vztahy (483) a (484), můžeme psát

$$\begin{aligned}\sigma &= \varphi\sigma'+\frac{\tilde\beta+\tilde c}{\tilde\beta}+1,\\ \delta &= \varphi\delta'+\frac{\tilde\beta+\tilde c}{\tilde\beta},\end{aligned}$$

což po dosazení do (485) a po úpravě dává

$$\left(\frac{\tilde\beta+\tilde c}{\tilde\beta}\sigma'-\delta'\right)(\sigma'-\delta')\le 0.$$

Podle (483) a (484) můžeme psát $\sigma'-\delta'=\tilde a\tilde c-\tilde b^2>0$, takže nerovnost (485) je ekvivalentní nerovnosti $((\tilde\beta+\tilde c)/\tilde\beta)\sigma'-\delta'\le 0$. Použijeme-li (483), (484) a (479), dostaneme

$$\begin{aligned}\frac{\tilde\beta+\tilde c}{\tilde\beta}\sigma'-\delta' &= \frac{\tilde\beta(B+D)}{\tilde\beta^2\tilde c}-\frac{(\tilde\beta+\tilde c)(A+D)}{\tilde\beta^2\tilde c}=\\ &= \frac{\tilde\beta B-(\tilde\beta+\tilde c)A-\tilde cD}{\tilde\beta^2\tilde c}=-\frac{D}{\tilde\beta^2}\le 0,\end{aligned} \tag{486}$$

takže platí (485) a tedy i $\underline{\lambda}'/\sigma' \geq 0$ a $\overline{\lambda}'/\sigma' \geq 0$. Podle (483), (484) jsou σ, δ a tedy i kořeny rovnice (482) spojitými funkcemi parametru φ, takže platí $\underline{\lambda}'/\sigma' \geq 0$, $\overline{\lambda}'/\sigma' \geq 0$ i když $\underline{\lambda} = \overline{\lambda}$. □

Věta 113. *Nechť jsou splněny předpoklady lemmatu 54. Pak matice H_+ je pozitivně definitní právě tehdy, když $B + D > 0$ a $\delta > 0$.*

Důkaz. Stejným způsobem jako v důkazu věty 79 se ukáže, že matice H_+ je pozitivně definitní právě tehdy, když $\sigma > 0$ a $\delta > 0$.

(a) předpokládejme, že $\sigma > 0$ a $\delta > 0$. Ukážeme, že v tomto případě musí platit $B + D > 0$. Nechť $B + D = 0$. Pak $\delta > 0$ implikuje $\tilde{\beta}(\tilde{\beta} + \tilde{c}) > 0$ (vzorec (484)), takže podle (479) platí $B + D \geq B = \tilde{\beta}(\tilde{\beta} + \tilde{c})(\tilde{a}\tilde{c} - \tilde{b}^2) > 0$, což je ve sporu s předpokladem $B + D = 0$. Nechť $B + D < 0$. Jelikož $A + D > 0$, mají derivace σ' a δ' opačná znaménka. Předpokládejme, že $\sigma' > 0$ a $\delta' < 0$, (důkaz pro $\sigma' < 0$ a $\delta' > 0$ je úplně stejný). Pak pro dostatečně velké hodnoty parametru φ platí $\delta < 0$, takže

$$\underline{\lambda} = \frac{\sigma}{2} - \sqrt{\left(\frac{\sigma}{2}\right)^2 - \delta} < \frac{\sigma}{2} - \frac{\sigma}{2} = 0.$$

Podle lemmatu 54 však platí $\underline{\lambda}' = (\underline{\lambda}'/\sigma')\sigma' \geq 0$, takže $\underline{\lambda}$ je neklesající funkcí parametru φ a tedy $\underline{\lambda} < 0$ pro všechny hodnoty parametru φ. Odtud plyne, že matice H_+ nemůže být pozitivně definitní pro žádnou hodnotu parametru φ.

(b) Nyní dokážeme, že z $B + D > 0$ a $\delta > 0$ plyne $\sigma > 0$, takže jsou splněny nutné a postačující podmínky pro pozitivní definitnost matice H_+. Jelikož $B + D > 0$, je podmínka $\sigma > 0$ ekvivalentní podmínce $\sigma(B + D) > 0$. Použijeme-li vztahy (479) až (484), dostaneme

$$\begin{aligned} \sigma(B + D) - \delta(A + D) &= \frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}}(B - A) + B + D = \\ &= C - B + B + D = C + D \geq 0. \end{aligned}$$

Jelikož $\delta(A + D) > 0$, platí také $\sigma(B + D) > 0$. □

Podle věty 113 je matice H_+ pozitivně definitní právě tehdy, když $B + D > 0$ a $\delta > 0$. Podmínka $B + D > 0$ nezávisí na výběru parametru φ ale pouze na hodnotách čísel $\tilde{\alpha}$, $\tilde{\beta}$ a prvků matice (478) (odpovídá v jistém smyslu podmínce $b > 0$ použité v oddlu 4.1). Jestliže $B + D > 0$, omezuje podmínka $\delta > 0$ volbu parametru φ. Položíme-li například $\delta = 1$ v (484), dostaneme

$$\varphi = \frac{\tilde{c}^2}{B + D}. \tag{487}$$

Další hodnotu parametru φ získáme minimalizací spektrálního čísla podmíněnosti matice $H^{-1/2}H_+H^{-1/2}$.

Lemma 55. *Nechť jsou splněny předpoklady lemmatu 54. Nechť $\underline{\lambda} \leq \overline{\lambda}$ jsou kořeny kvadratické rovnice (482). Pak*

a) Podmínka $\underline{\lambda} \leq 1 \leq \overline{\lambda}$ je splněna právě tehdy, když $\varphi \geq 0$.
b) Podíl $\overline{\lambda}/\underline{\lambda}$ nabývá svého minima (v oblasti, kde $\underline{\lambda} > 0$ a $\overline{\lambda} > 0$) právě tehdy, když

$$\varphi = \frac{\tilde{c}^2(D - B)}{(A + D)(B + D)}. \tag{488}$$

Důkaz. (a) Stejným způsobem jako v části (b) důkazu lemmatu 44 se ukáže, že $\underline{\lambda} \leq 1 \leq \overline{\lambda}$ platí právě tehdy, když $\det \tilde{M} \leq 0$, což dává $\varphi = -\det \tilde{M} \geq 0$.

[b) Stejným způsobem jako v části (c) důkazu lemmatu 44 se ukáže, že podíl $\overline{\lambda}/\underline{\lambda}$ nabývá svého minima, pokud $2\sigma'\delta - \sigma\delta' = 0$. Ale

$$\begin{aligned} 2\sigma'\delta - \sigma\delta' &= \varphi\sigma'\delta' + \frac{\tilde{\beta}+\tilde{c}}{\tilde{\beta}}(\sigma'-\delta') + \frac{\tilde{\beta}+\tilde{c}}{\tilde{\beta}}\sigma' - \delta' \\ &= \frac{\varphi}{\tilde{\beta}^2\tilde{c}^2}(A+D)(B+D) + \frac{B-D}{\tilde{\beta}^2}, \end{aligned}$$

takže podíl $\overline{\lambda}/\underline{\lambda}$ nabývá svého minima platí-li (488). □

Věta 114. *Nechť jsou splněny předpoklady lemmatu 54. Pak spektrální číslo podmíněnosti matice $H^{-1/2}H_+H^{-1/2}$ je minimální, pokud*

$$\varphi = \max\left(0, \frac{\tilde{c}^2(D-B)}{(A+D)(B+D)}\right). \tag{489}$$

Důkaz. Podle lemmatu 54 platí

$$\kappa(H^{-1/2}H_+H^{-1/2}) = \frac{\max(1,\overline{\lambda})}{\min(1,\underline{\lambda})}, \tag{490}$$

kde $\underline{\lambda} \le \overline{\lambda}$ jsou kořeny kvadratické rovnice (482). Je-li splněna podmínka $\varphi \ge 0$, platí $\underline{\lambda} \le 1 \le \overline{\lambda}$, neboli $\max(1,\overline{\lambda})/\min(1,\underline{\lambda}) = \overline{\lambda}/\underline{\lambda}$. Podle lemmatu 55 je podíl $\overline{\lambda}/\underline{\lambda}$ minimální (v oblasti, kde $H_+ > 0$) právě tehdy, platí-li (488), takže vyhovuje-li hodnota (488) podmínce $\varphi \ge 0$, dává tato hodnota minimum výrazu (490). V opačném příp není podíl λ_2/λ_1 ekvivalentní číslu (490). Jelikož podle lemmatu 54 mají derivace $\underline{\lambda}'$ a $\overline{\lambda}'$ stejná znaménka, můžeme hodnotu čísla (490) dále snižovat a minimální hodnotu dostaneme právě tehdy, platí-li buď $\underline{\lambda} = 1$, nebo $\overline{\lambda} = 1$, což implikuje podmínku $\varphi = 0$. Spojíme-li oba případy, dostaneme (489). □

Poznámka 186. Inverzí vztahu $H_+ = H + \tilde{U}\tilde{M}\tilde{U}^T$ podle důsledku 12 dostaneme (tak jako v důkazu věty 80) $B_+ = B + B\tilde{U}\tilde{K}\tilde{U}^TB$. Položíme-li $v = -Bu$ a $\tilde{V} = [v, y - Bd]$, můžeme psát $B_+ = B + \tilde{V}\tilde{K}\tilde{V}^T$. Dá se snadno ukázat (podobně jako v důkazu věty 112), že tato aktualizace splňuje kvazinewtonovskou podmínku $B_+d = y$, pokud

$$\tilde{K} = \begin{bmatrix} -\psi\tilde{\delta}, & \psi\tilde{\gamma} \\ \psi\tilde{\gamma}, & \frac{1}{\tilde{\delta}}(1-\psi\tilde{\gamma}^2) \end{bmatrix}, \tag{491}$$

kde $\psi = -\det K = \varphi/\delta$ je volný parametr a

$$\tilde{\gamma} = d^Tv, \qquad \tilde{\delta} = d^T(y - Bd) = b - c. \tag{492}$$

Poznamenejme, že čísla $\tilde{\gamma}$ a $\tilde{\delta}$ nemají nic společného se škálovacím parametrem γ a číslem δ v (484).

Poznámka 187. Vztah $B_+ = B + \tilde{V}\tilde{K}\tilde{V}^T$ můžeme roznásobit. Pak platí

$$B_+ = B + \frac{1}{\tilde{\delta}}(y - Bd)(y - Bd)^T - \frac{\psi}{\tilde{\delta}}v_+v_+^T, \tag{493}$$

kde

$$v_+ = \tilde{\delta}v - \tilde{\gamma}(y - Bd). \tag{494}$$

Vztah (493) je duální ke vztahu (477) a lze ho z (477) dostat záměnou, $d \to y$, $y \to d$, $u \to v$, $H \to B$, $\tilde{\alpha} \to \tilde{\gamma}$, $\tilde{\beta} \to \tilde{\delta}$, $\varphi \to \psi$.

Poznámka 188. Použijeme-li (473) a (492), dostaneme po jednoduchých úpravách

$$\tilde{\gamma} = -(\tilde{\alpha} + \tilde{b}) \qquad \tilde{\delta} = -(\tilde{\beta} + \tilde{c}). \tag{495}$$

Můžeme tedy psát

$$\begin{aligned} A &= (\tilde{\delta}+\tilde{c})^2(\tilde{a}\tilde{c}-\tilde{b}^2), \\ B &= \tilde{\delta}(\tilde{\delta}+\tilde{c})(\tilde{a}\tilde{c}-\tilde{b}^2), \\ C &= \tilde{\delta}^2(\tilde{a}\tilde{c}-\tilde{b}^2), \\ D &= (\tilde{\delta}\tilde{b}-\tilde{\gamma}\tilde{c})^2. \end{aligned} \tag{496}$$

Odtud je vidět, že záměna $\tilde{\alpha} \to \tilde{\gamma}$, $\tilde{\beta} \to \tilde{\delta}$ ponechá výrazy B a D beze změny ale způsobí záměnu $A \to C$, $C \to A$. Použijeme-li dualitu a vztah $\det H_+ / \det H = \det B / \det B_+$, dostaneme

$$\frac{1}{\delta} = \frac{\tilde{\delta}+\tilde{c}}{\tilde{\delta}} - \frac{\psi}{\tilde{\delta}\tilde{c}}\,(B+D). \tag{497}$$

Vektor v_+ vystupující v aktualizaci (493) a určený vztahem (494) se nerovná vektoru $-B_+u_+$. Platí však tato věta.

Věta 115. *Nechť $v = -Bu$ a v_+ je vektor určený vztahem (494). Pak platí*

$$-B_+u_+ = \frac{v_+}{\delta}. \tag{498}$$

Důkaz. Použijeme-li vztahy (493) a (494) spolu s (474), můžeme psát

$$\begin{aligned} B_+u_+ &= \tilde{\beta}Bu - \tilde{\alpha}B(d-Hy) + \frac{1}{\tilde{\delta}}B(d-Hy)(\tilde{\beta}\tilde{b}-\tilde{\alpha}\tilde{c}) \\ &\quad -\frac{\psi}{\tilde{\delta}}(\tilde{\delta}Bu - \tilde{\gamma}B(d-Hy))(\tilde{\beta}\tilde{\delta}\tilde{a} - \tilde{\alpha}\tilde{\delta}\tilde{b} - \tilde{\beta}\tilde{\gamma}\tilde{b} + \tilde{\alpha}\tilde{\gamma}\tilde{c}) \\ &= \frac{\tilde{\beta}}{\tilde{\delta}}(\tilde{\delta}Bu - \tilde{\gamma}B(d-Hy)) + \frac{\psi}{\tilde{\delta}}(\tilde{\delta}Bu - \tilde{\gamma}B(d-Hy))(\tilde{\alpha}^2\tilde{c} - 2\tilde{\alpha}\tilde{\beta}\tilde{b} + \tilde{\beta}^2\tilde{a} + \tilde{\beta}(\tilde{a}\tilde{c}-\tilde{b}^2)). \end{aligned}$$

Ale $\tilde{\beta} = -(\tilde{\delta}+\tilde{c})$ podle (495) a

$$\tilde{\alpha}^2\tilde{c} - 2\tilde{\alpha}\tilde{\beta}\tilde{b} + \tilde{\beta}^2\tilde{a} + \tilde{\beta}(\tilde{a}\tilde{c}-\tilde{b}^2) = \frac{B+D}{\tilde{c}},$$

jako v části (a) důkazu lemmatu 54, takže

$$B_+u_+ = -(\tilde{\delta}Bu - \tilde{\gamma}B(d-Hy))\left(\frac{\tilde{\delta}+\tilde{c}}{\tilde{\delta}} - \frac{\psi}{\tilde{\delta}\tilde{c}}\,(B+D)\right),$$

což spolu s (497) dává (498). □

Dosadíme-li hodnotu (488) nebo hodnotu (489) do vztahu (477), dostaneme pozitivně definitní matici H_+ (je-li matice H pozitivně definitní a platí-li $B+D > 0$). Další vhodné hodnoty parametru φ lze získat přetransformováním rekurentního vztahu (468) na $H_+ = H + \bar{U}\bar{M}\bar{U}^T$, kde $\bar{U} = PU = P[d, Hy]$ a $P = \tilde{U}(\tilde{U}^TH^{-1}\tilde{U})^{-1}\tilde{U}^TH^{-1}$ (P je matice projekce do $\mathcal{L}(\tilde{U})$). Dostaneme tak aktualizace z Davidonovy třídy zapsané ve tvaru použitém v práci [36]. V tomto případě platí

$$H + \bar{U}\bar{M}\bar{U}^T = H + \tilde{U}(\tilde{U}^TH^{-1}\tilde{U})^{-1}\tilde{U}^TH^{-1}U\bar{M}U^TH^{-1}\tilde{U}(\tilde{U}^TH^{-1}\tilde{U})^{-1}\tilde{U}^T,$$

takže $H + \tilde{U}\tilde{M}\tilde{U}^T = H + \bar{U}\bar{M}\bar{U}^T$, pokud

$$\tilde{M} = (\tilde{U}^TH^{-1}\tilde{U})^{-1}\tilde{U}^TH^{-1}U\bar{M}U^TH^{-1}\tilde{U}(\tilde{U}^TH^{-1}\tilde{U})^{-1}. \tag{499}$$

Z vyjádření (499) je patrné, že matice $\tilde{M}$ je jednoznačně určena maticí $\bar{M}$. Jsou-li matice $\tilde{U}^TH^{-1}\tilde{U}$ a $\tilde{U}^TH^{-1}U$ regulární (platí-li $\tilde{a}\tilde{c} - \tilde{b}^2 > 0$ a $D > 0$), je matice $\bar{M}$ jednoznačně určena maticí $\tilde{M}$. Výhoda transformovaných aktualizací spočívá v tom, že mají podobné vlastnosti jako aktualizace patřící do Broydenovy třídy.

Věta 116. *Nechť $H_+ = H + \bar{U}\bar{M}\bar{U}^T$, kde H je symetrická pozitivně definitní matice a $\bar{U} = PU = P[d, Hy]$, kde $P = \tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T H^{-1}$ a $\tilde{U} = [u, d - Hy]$. Pak $H_+ y = d$ platí právě tehdy, když*

$$\bar{M} = \begin{bmatrix} \frac{1}{\bar{b}}\left(\bar{\eta}\frac{\bar{a}}{\bar{b}} + 1\right), & -\frac{\bar{\eta}}{\bar{b}} \\ -\frac{\bar{\eta}}{\bar{b}}, & \frac{\bar{\eta} - 1}{\bar{a}} \end{bmatrix}, \tag{500}$$

kde $\bar{\eta}$ je volný parametr a

$$\begin{bmatrix} \bar{c} & \bar{b} \\ \bar{b} & \bar{a} \end{bmatrix} = (PU)^T H^{-1} PU = U^T H^{-1}\tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T H^{-1} U, \tag{501}$$

takže $\bar{a} = y^T PHy$, $\bar{b} = y^T Pd$, $\bar{c} = d^T H^{-1} Pd$) (čísla $\bar{a}$, $\bar{b}$, $\bar{c}$, která mají podobný význam jako čísla a, b, c v (303), se liší od čísel $\tilde{a}$, $\tilde{b}$, $\tilde{c}$ v (478)).

Důkaz Jelikož $P(d - Hy) = d - Hy$ a $P\bar{U} = \bar{U}$ (neboť $d - Hy \in \mathcal{L}(U)$ a $P^2 = P$), můžeme kvazinewtonovskou podmínku zapsat ve tvaru

$$\bar{U}\bar{M}\bar{U}^T P^T y = Pd - PHy.$$

Ale

$$\begin{aligned} P^T H^{-1} P &= H^{-1}\tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T H^{-1}\tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T H^{-1} \\ &= H^{-1}\tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T H^{-1} = H^{-1}P = P^T H^{-1} \end{aligned} \tag{502}$$

takže $\bar{U}^T P^T y = \bar{U}^T P^T H^{-1} Hy = \bar{U}^T P^T H^{-1} PHy = \bar{U}^T H^{-1} PHy$, což po dosazení do kvazinewtonovské podmínky dává

$$[Pd, PHy]\begin{bmatrix} \bar{m}_1, & \bar{m}_2, \\ \bar{m}_2, & \bar{m}_3 \end{bmatrix}\begin{bmatrix} \bar{b} \\ \bar{a} \end{bmatrix} = Pd - PHy,$$

kde

$$\begin{bmatrix} \bar{c}, & \bar{b} \\ \bar{b}, & \bar{a} \end{bmatrix} = \bar{U}^T H^{-1}\bar{U} = [d, Hy]^T P^T H^{-1} P[d, Hy] = [d, Hy]^T H^{-1}[Pd, PHy].$$

Porovnáme-li koeficienty u Pd a PHy, dostaneme

$$\begin{aligned} \bar{m}_1\bar{b} + \bar{m}_2\bar{a} &= 1, \\ \bar{m}_2\bar{b} + \bar{m}_3\bar{a} &= -1. \end{aligned}$$

Jeden parametr je nadbytečný. Zvolíme $m_2 = -\bar{\eta}/b$ a zbylé prvky $\bar{m}_1, \bar{m}_3$ určíme řešením uvedených rovnic. Tím dostaneme matici $\bar{M}$ uvedenou ve větě 116. □

Poznámka 189. Vztah $H_+ = H + \bar{U}\bar{M}\bar{U}^T$ můžeme roznásobit. Platí

$$H_+ = H + \frac{1}{\bar{b}}Pd(Pd)^T - \frac{1}{\bar{a}}PHy(PHy)^T + \frac{\bar{\eta}}{\bar{a}}\left(\frac{\bar{a}}{\bar{b}}Pd - PHy\right)\left(\frac{\bar{a}}{\bar{b}}Pd - PHy\right)^T. \tag{503}$$

Pro tuto aktualizaci platí stejné úvahy jako pro aktualizaci (286). Matice H_+ je pozitivně definitní, pokud $\bar{\eta} > \bar{\eta}^*$, kde $\bar{\eta}^* = -\bar{b}^2/(\bar{a}\bar{c} - \bar{b}^2)$. Pro $\bar{\eta} = 0$ dostaneme analogii metody DFP a pro $\bar{\eta} = 1$ dostaneme analogii metody BFGS.

Poznámka 190. Aktualizaci (503) můžeme vyjádřit v inverzním tvaru. Použije se k tomu stejný postup jako v důkazu věty 80 a rovnost $BP = P^T B$, která plyne z (502). Platí

$$B_+ = B + \frac{1}{\bar{b}}P^T y(P^T y)^T - \frac{1}{\bar{c}}P^T Bd(P^T Bd)^T + \frac{\bar{\beta}}{\bar{c}}\left(\frac{\bar{c}}{\bar{b}}P^T y - P^T Bd\right)\left(\frac{\bar{c}}{\bar{b}}P^T y - P^T Bd\right)^T, \tag{504}$$

kde

$$\bar{\beta}\bar{\eta}(\bar{a}\bar{c} - \bar{b}^2) + (\bar{\beta} + \bar{\eta})\bar{b}^2 = \bar{b}^2.$$

Pro tuto aktualizaci platí stejné úvahy jako pro aktualizaci (306). Matice B_+ je pozitivně definitní, pokud $\bar{b} > 0$ a $\bar{\beta} > \bar{\beta}^*$, kde $\bar{\beta}^* = -\bar{b}^2/(\bar{a}\bar{c} - \bar{b}^2)$. Pro $\bar{\beta} = 0$ dostaneme analogii metody BFGS a pro $\bar{\beta} = 1$ dostaneme analogii metody DFP.

Poznámka 191. Při realizaci rekurentního vztahu (503) matici P obvykle nekonstruujeme. Místo toho určujeme přímo vektory Pd a PHy. Platí

$$P = \frac{1}{\tilde{a}\tilde{c} - \tilde{b}^2}[u, d - Hy] \begin{bmatrix} \tilde{c} & -\tilde{b} \\ -\tilde{b} & \tilde{a} \end{bmatrix} \begin{bmatrix} u^T H^{-1} \\ (d - Hy)^T H^{-1} \end{bmatrix},$$

což spolu s (473) dává

$$PHy = \frac{\tilde{\beta}\tilde{a} - \tilde{\alpha}\tilde{b}}{\tilde{a}\tilde{c} - \tilde{b}^2}(d - Hy) - \frac{\tilde{\beta}\tilde{b} - \tilde{\alpha}\tilde{c}}{\tilde{a}\tilde{c} - \tilde{b}^2}u. \tag{505}$$

Dále platí

$$Pd = d - Hy + PHy, \tag{506}$$

neboť $d - Hy \in \mathcal{L}(\tilde{U})$, takže $Pd - PHy = P(d - Hy) = d - Hy$. Vztah (505) můžeme použít pouze tehdy, když $\tilde{a}\tilde{c} - \tilde{b}^2 \neq 0$. V opačném případě je vektor u rovnoběžný s vektorem $d - Hy$, takže matice projekce má tvar

$$P = \frac{(d - Hy)(d - Hy)^T H^{-1}}{(d - Hy)^T H^{-1}(d - Hy)}$$

a tedy $PHy = (\tilde{\beta}/\tilde{c})(d - Hy)$. Podobným způsobem postupujeme při realizaci rekurentního vztahu (504). Protože $P^T = H^{-1}\tilde{U}(\tilde{U}^T H^{-1}\tilde{U})^{-1}\tilde{U}^T = \tilde{V}(\tilde{V}^T B^{-1}\tilde{V})^{-1}\tilde{V}^T B^{-1}$, můžeme použít dualitu, jejíž pomocí dostaneme

$$P^T Bd = \frac{\tilde{\delta}\tilde{a} - \tilde{\gamma}\tilde{b}}{\tilde{a}\tilde{c} - \tilde{b}^2}(y - By) - \frac{\tilde{\delta}\tilde{b} - \tilde{\gamma}\tilde{c}}{\tilde{a}\tilde{c} - \tilde{b}^2}v \tag{507}$$

a

$$P^T y = y - Bd + P^T Bd. \tag{508}$$

Jsou-li vektory v a $y - Bd$ lineárně závislé, platí $P^T Bd = (\tilde{\delta}/\tilde{c})(y - Bd)$.

Nyní odvodíme vztah mezi parametry φ a $\bar{\eta}$ v (477) a (503).

Lemma 56. *Nechť jsou splněny předpoklady lemmatu 54 a nechť $\bar{a}$, $\bar{b}$, $\bar{c}$ jsou hodnoty definované vztahem (501). Pak platí*

$$\bar{a} = \frac{A + D}{\tilde{c}(\tilde{a}\tilde{c} - \tilde{b}^2)} \qquad \bar{b} = \frac{B + D}{\tilde{c}(\tilde{a}\tilde{c} - \tilde{b}^2)} \qquad \bar{c} = \frac{C + D}{\tilde{c}(\tilde{a}\tilde{c} - \tilde{b}^2)} \tag{509}$$

a

$$\bar{a}\bar{c} - \bar{b}^2 = \frac{D}{\tilde{a}\tilde{c} - \tilde{b}^2}. \tag{510}$$

Důkaz Použijeme-li vztahy (478), (480) a (501), můžeme psát

$$\begin{aligned} \begin{bmatrix} \bar{c} & \bar{b} \\ \bar{b} & \bar{a} \end{bmatrix} &= \frac{1}{\tilde{a}\tilde{c} - \tilde{b}^2} \begin{bmatrix} \tilde{\alpha} + \tilde{b} & \tilde{\beta} + \tilde{c} \\ \tilde{\alpha} & \tilde{\beta} \end{bmatrix} \begin{bmatrix} \tilde{c} & -\tilde{b} \\ -\tilde{b} & \tilde{a} \end{bmatrix} \begin{bmatrix} \tilde{\alpha} + \tilde{b} & \tilde{\alpha} \\ \tilde{\beta} + \tilde{c} & \tilde{\beta} \end{bmatrix} \\ &= \frac{1}{\tilde{a}\tilde{c} - \tilde{b}^2} \begin{bmatrix} \tilde{\alpha}\tilde{c} - \tilde{\beta}\tilde{b} & \tilde{\beta}\tilde{a} - \tilde{\alpha}\tilde{b} + \tilde{a}\tilde{c} - \tilde{b}^2 \\ \tilde{\alpha}\tilde{c} - \tilde{\beta}\tilde{b} & \tilde{\beta}\tilde{a} - \tilde{\alpha}\tilde{b} \end{bmatrix} \begin{bmatrix} \tilde{\alpha} + \tilde{b} & \tilde{\alpha} \\ \tilde{\beta} + \tilde{c} & \tilde{\beta} \end{bmatrix}, \end{aligned}$$

takže

$$\begin{aligned}\bar{a} &= \frac{\tilde{\alpha}^2\tilde{c} - 2\tilde{\alpha}\tilde{\beta}\tilde{b} + \tilde{\beta}^2\tilde{a}}{\tilde{a}\tilde{c} - \tilde{b}^2},\\ \bar{b} &= \frac{\tilde{\alpha}^2\tilde{c} - 2\tilde{\alpha}\tilde{\beta}\tilde{b} + \tilde{\beta}^2\tilde{a} + \tilde{\beta}(\tilde{a}\tilde{c} - \tilde{b}^2)}{\tilde{a}\tilde{c} - \tilde{b}^2},\\ \bar{c} &= \frac{\tilde{\alpha}^2\tilde{c} - 2\tilde{\alpha}\tilde{\beta}\tilde{b} + \tilde{\beta}^2\tilde{a} + (2\tilde{\beta} + \tilde{c})(\tilde{a}\tilde{c} - \tilde{b}^2)}{\tilde{a}\tilde{c} - \tilde{b}^2}.\end{aligned}$$

Stejný výsledek dostaneme (podobně jako v části (a) důkazu lemmatu 54), dosadíme-li (479) do výrazů (509). Vztah (510) plyne bezprostředně z (478), (480) a (501). □

Věta 117. *Nechť jsou splněny předpoklady lemmatu 54 a $B+D \neq 0$. Pak aktualizace (477) je ekvivalentní aktualizaci (503) právě tehdy, když*

$$\varphi = \frac{\tilde{c}^2 D}{(A+D)(B+D)} \left(1 - \tilde{\beta}\,\frac{\bar{\eta}}{\bar{b}}\right), \tag{511}$$

kde $\bar{b}$ je číslo uvedené v (509).

Důkaz Podobně jako v oddílu 4.1 lze psát $\det H_+ / \det H = \bar{\delta}$ a z (482) plyne, že $\det H_+ / \det H = \delta$, což dohromady dává $\delta = \bar{\delta}$. Použijeme-li (484) a (300) (s pruhovanými veličinami) spolu s (509) a (510), dostaneme

$$\frac{\tilde{\beta} + \tilde{c}}{\tilde{\beta}} - \frac{\varphi}{\tilde{\beta}\tilde{c}}\,(B+D) = \bar{\eta}\,\frac{\bar{a}\bar{c} - \bar{b}^2}{\bar{a}\bar{b}} + \frac{\bar{b}}{\bar{a}} = \bar{\eta}\,\frac{\tilde{c}^2(\tilde{a}\tilde{c} - \tilde{b}^2)D}{(A+D)(B+D)} + \frac{B+D}{A+D}.$$

Platí tedy

$$\varphi\,\frac{B+D}{\tilde{\beta}\tilde{c}} = \frac{B}{A} - \frac{B+D}{A+D} - \bar{\eta}\,\frac{\tilde{c}^2(\tilde{a}\tilde{c} - \tilde{b}^2)D}{(A+D)(B+D)}$$

(neboť $(\tilde{\beta} + \tilde{c})/\tilde{\beta} = B/A$ podle (479)). Použijeme-li znovu (479), dostaneme

$$\frac{B}{A} - \frac{B+D}{A+D} = \frac{(B-A)D}{\tilde{\beta}(A+D)} = \frac{\tilde{c}D}{\tilde{\beta}(A+D)},$$

takže

$$\varphi = \frac{\tilde{c}^2}{(A+D)(B+D)} \left(1 - \bar{\eta}\,\frac{\tilde{\beta}\tilde{c}(\tilde{a}\tilde{c} - \tilde{b}^2)}{B+D}\right),$$

což po úpravě dává (511). □

Poznámka 192. Použitím duality dostaneme

$$\psi = \frac{\tilde{c}^2 D}{(A+D)(B+D)} \left(1 - \tilde{\delta}\,\frac{\bar{\beta}}{\bar{b}}\right). \tag{512}$$

Poznámka 193. Položíme-li $\bar{\eta} = 0$ nebo $\bar{\beta} = 1$ (analogie metody DFP), dostaneme

$$\varphi = \frac{\tilde{c}^2 D}{(A+D)(B+D)}, \qquad \psi = \frac{\tilde{c}^2 D}{(B+D)^2}, \qquad \delta = \frac{\bar{b}}{\bar{a}} = \frac{B+D}{A+D}.$$

Položíme-li $\bar{\eta} = 1$ nebo $\bar{\beta} = 0$ (analogie metody BFGS), dostaneme

$$\varphi = \frac{\tilde{c}^2 D}{(B+D)^2}, \qquad \psi = \frac{\tilde{c}^2 D}{(B+D)(C+D)}, \qquad \delta = \frac{\bar{c}}{\bar{b}} = \frac{C+D}{B+D}.$$

Další hodnoty jsou uvedeny v následující tabulce.

$\bar\eta$	$\bar\beta$	φ	ψ	δ
0	1	$\dfrac{\tilde c^2 D}{(A+D)(B+D)}$	$\dfrac{\tilde c^2 D}{(B+D)^2}$	$\dfrac{B+D}{A+D}$
1	0	$\dfrac{\tilde c^2 D}{(B+D)^2}$	$\dfrac{\tilde c^2 D}{(B+D)(C+D)}$	$\dfrac{C+D}{B+D}$
$\dfrac{\bar b}{\bar b+\bar a}$	$\dfrac{\bar b}{\bar b+\bar c}$	$\dfrac{2\tilde c^2 D}{(A+B+2D)(B+D)}$	$\dfrac{2\tilde c^2 D}{(B+D)(B+C+2D)}$	$\dfrac{B+C+2D}{A+B+2D}$
$\dfrac{\bar b}{\bar b-\bar a}$	$\dfrac{\bar b}{\bar b-\bar c}$	0	0	$\dfrac{\tilde\beta+\tilde c}{\tilde\beta}$
$\dfrac{\bar b(\bar c-\bar b)}{\bar a\bar c-\bar b^2}$	$\dfrac{\bar b(\bar a-\bar b)}{\bar a\bar c-\bar b^2}$	$\dfrac{\tilde c^2(D-B)}{(A+D)(B+D)}$	$\dfrac{\tilde c^2(D-B)}{(B+D)(C+D)}$	$\dfrac{C+D}{A+D}$
$\dfrac{\bar b(\bar a-\bar b)}{\bar a\bar c-\bar b^2}$	$\dfrac{\bar b(\bar c-\bar b)}{\bar a\bar c-\bar b^2}$	$\dfrac{\tilde c^2}{B+D}$	$\dfrac{\tilde c^2}{B+D}$	1

Poznámka 194. Při výpočtu směrového vektoru s_+ můžeme ušetřit aritmetické operace, nahradíme-li vztah (451) vzorcem obsahujícím vektor Hy získaný při výpočtu matice H_+. Podobným způsobem jako v důkazu věty 107 lze použitím (477) snadno ukázat, že platí

$$s_+ = -H_+ g_+ = s - Hy - \frac{(d-Hy)^T g_+}{\tilde\beta}(d-Hy) + \varphi\,\frac{u_+^T g_+}{\tilde\beta}u_+ \tag{513}$$

Závěrem popíšeme základní algoritmus, kterým se realizují metody s proměnnou metrikou z Davidonovy třídy. Nejprve uvedeme několik poznámek k implementaci jednotlivých kroků tohoto algoritmu.

(1) Základní kroky algoritmu (výpočet směrového vektoru a výběr délky kroku) jsou stejné jako u algoritmu realizujícího metody z Broydenovy třídy. Vztah (451) můžeme nahradit vzorcem (513).

(2) K aktualizaci (477) potřebujeme hodnotu $\tilde a = u^T H^{-1} u$. Tuto hodnotu lze určit řešením soustavy lineárních rovnic, což vyžaduje $O(n^3)$ aritmetických opercí. Tomu se lze vyhnout, používáme-li vektor Bu, který získáváme rekurentně tak, že položíme $v = -Bu$, vypočteme vektor v_+ podle (494), a položíme $B_+u_+ = -v_+/\delta$ (věta 115).

(3) Jednoduché použití vztahu (474) může způsobit exponenciální nárůst normy vektoru u. Protože výsledek aktualizace (477) nezávisí na normě vektoru u, je výhodné tento vektor předem normalizovat. Děláme to tak, že před provedením aktualizace spočteme hodnoty $\tilde a$, $\tilde c$, položíme $\lambda = \sqrt{\tilde c/\tilde a}$ a vektory u, v vynásobníme číslem λ. Potom je třeba položit $\tilde a = \tilde c$.

(4) Je účelné aktualizaci (477) vhodně škálovat (používá se počáteční nebo řízené škálování). Děláme to tak, že před provedením aktualizace spočteme hodnoty a, c podle (78), položíme $\gamma = \sqrt{c/a}$ a vynásobíme matici H a vektor u číslem γ a vektor $v = -Bu$ číslem $1/\gamma$.

Algoritmus metody s proměnnou metrikou z Davidonovy třídy lze popsat zhruba takto:

Algoritmus 9. Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\underline{\varepsilon} > 0$, $\underline{\gamma} = 0.7$, $\overline{\gamma} = 6$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme $F_1 := F(x_1)$, $g_1 := g(x_1)$. Zvolíme počáteční symetrickou pozitivně definití matici H_1 (obvykle $H_1 := I$) a položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \le \underline{\varepsilon}$ ukončíme výpočet. V opačném případě položíme $s_i := -H_i g_i$. Použitím algoritmu 1 najdeme novou aproximaci řešení x_{i+1} a položíme $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$.

Krok 3 Položíme $d_i := x_{i+1} - x_i$, $y_i := g_{i+1} - g_i$ a jestliže $i = 1$, položíme $u_i := H_i y_i$ a $v_i := y_i$. Vypočteme vektory $H_i y_i$, $B_i d_i := -(\|d_i\|/\|s_i\|)g_i$ a čísla $a_i := y_i^T H_i y_i$, $b_i := y_i^T d_i$, $c_i := d_i^T B_i d_i$. Rozhodneme-li se v i-té iteraci škálovat, položíme $\gamma_i := \sqrt{c_i/a_i}$, $H_i := \gamma_i H_i$, $u_i := \gamma_i u_i$, $a_i := \gamma_i a_i$, $v_i := v_i/\gamma_i$ a $c_i := c_i/\gamma_i$.

Krok 4 Vypočteme čísla $\tilde{a}_i := -u_i^T v_i$, $\tilde{c}_i := a_i - 2b_i + c_i$ a $\lambda_i := \sqrt{\tilde{c}_i/\tilde{a}_i}$. Položíme $u_i := \lambda_i u_i$, $v_i := \lambda_i v_i$, $\tilde{a}_i = \tilde{c}_i$ a vypočteme čísla $\tilde{\alpha}_i := y_i^T u_i$, $\tilde{\beta}_i := b_i - a_i$, $\tilde{\gamma}_i := d_i^T v_i$, $\tilde{\delta}_i := b_i - c_i$ a $\tilde{b}_i := -(\tilde{\gamma}_i + \tilde{\alpha}_i)$. Vypočteme čísla A_i, B_i, C_i, D_i podle (479). Jestliže $B_i + D_i \leq 0$, přejdeme na krok 6.

Krok 5 Zvolíme hodnotu parametru φ_i a jí odpovídající hodnotu čísla δ_i (pro analogii metody BFGS volíme $\varphi_i = \tilde{c}_i^2 D_i/(B_i + D_i)^2$ a $\delta_i = (C_i + D_i)/(B_i + D_i)$). Vypočteme vektory $u_{i+1} := \tilde{\beta}_i u_i - \tilde{\alpha}_i(d_i - H_i y_i)$, $v_{i+1} := \tilde{\delta}_i v_i - \tilde{\gamma}_i(y_i - B_i d_i)$ a matici H_{i+1} podle (477). Položíme $v_{i+1} := v_{i+1}/\delta_i$, zvětšíme i o 1 a přejdeme na krok 2.

Krok 6 Zvolíme hodnotu parametru η_i a určíme matici H_{i+1} podle (286) (pro metodu BFGS volíme $\eta_i = 1$). Zvětšíme i o 1 a přejdeme na krok 2.

4.10 Numerické porovnání metod s proměnnou metrikou

V tomto oddílu porovnáme účinnost vybraných metod s proměnnou metrikou a jejich modifikací. Použijeme přitom soubor 91 testovacích úloh s 200 proměnnými (TEST28 z oddílu 1.5 bez úlohy 45). Nejprve porovnáme základní metody z Broydenovy třídy s různými škálovacími a korekčními strategiemi. V tabulce 3 jsou uvedeny výsledky získané těmito metodami:

`DFP` - metoda Davidona, Fletchera a Powella (287),
`BFGS` - metoda Broydena, Fletchera, Goldfarba a Shanna (288),
`H` - Hoshinova metoda (290),
`R1+` - modifikovaná metoda hodnosti 1 (353)–(354),
`OC` - optimálně podmíněná metoda používající vzorec (377) způsobem uvedeným ve větě 95,
`VS` - variačně odvozená metoda (385),
`VL+` - variačně odvozená metoda (466)–(467).

Jednotlivé metody jsou realizovány bez škálování `NS`, s počátečním škálováním `PS` a s řízeným škálováním `CS` (poznámka 167), a to bez korekce ($\rho = 1$) nebo s homogenní korekcí (401). Pro srovnání jsou též uvedeny výsledky získané metodou sdružených gradientů `CG`, což je algoritmus 5, kde používáme volbu `HST+` (vzorce (172), 173), (174)) a proceduru pro výběr délky kroku převzatou z programu `CG-DESCENT`. Metody s proměnnou metrikou (kromě metody `DFP`) byly realizovány jako metody spádových směrů používající slabou Wolfeho podmínku s $\varepsilon_1 = 0.0001$ a $\varepsilon_2 = 0.9$ (metoda `DFP` v tomto případě nespočetla třetinu úloh). Metody `DFP` a `CG` byly realizovány jako metody spádových směrů používající silnou Wolfeho podmínku s $\varepsilon_1 = 0.0001$ a $\varepsilon_2 = 0.1$.

	Bez korekce: $\rho = 1$				S korekcí: ρ podle (401)			
Metoda	NIT	NFV	selhání	čas	NIT	NFV	selhání	čas
DFP/NS	34068	77454	1	6.06	31419	71302	-	5.77
DFP/PS	44642	110644	5	9.61	41036	98867	4	9.87
DFP/CS	25441	58930	-	4.35	24534	57508	-	4.22
BFGS/NS	30177	55759	-	7.27	29857	54502	-	7.53
BFGS/PS	36268	39434	-	5.28	32800	35748	-	4.75
BFGS/CS	19428	22134	-	2.77	19292	22468	-	2.73
H/NS	34562	50457	-	6.77	35624	51457	-	7.00
H/PS	46641	48032	-	6.86	43061	44874	-	6.27
H/CS	22373	24112	-	3.13	22070	23975	-	3.07
R1+/NS	26433	52019	-	6.73	26345	51980	-	7.00
R1+/PS	24665	29378	-	3.38	24337	29265	-	3.22
R1+/CS	20653	23409	-	2.96	24913	28337	-	3.46
OC/NS	29308	45884	-	6.19	30841	47254	-	6.22
OC/PS	34766	36844	-	4.85	32573	34843	-	4.46
OC/CS	20263	22137	-	2.83	21007	23250	-	2.91
VS/NS	26608	54155	-	7.08	28096	55503	-	7.11
VS/PS	29097	33172	-	4.16	28252	32530	-	4.05
VS/CS	19821	22137	-	2.83	18950	23214	-	2.73
VL+/NS	27429	43773	-	6.17	27285	43244	-	6.05
VL+/PS	28835	31293	-	4.02	25951	28195	-	3.66
VL+/CS	18800	20678	-	2.71	17966	19923	-	2.60
CG	157391	315435	3	6.63	–			

Tabulka 3: TEST28 – 91 úloh

Tabulka 3 obsahuje celkový počet iterací NIT, celkový počet použitých funkčních hodnot NFV, počet selhání a celkový čas výpočtu. K selhání došlo, když nestačilo 8000 iterací nebo 8000 vyčíslení funkční hodnoty pro vyřešení dané úlohy. Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Metoda DFP je velmi neefektivní, používáme-li standardní výběr délky kroku založený na splnění slabé Wolfeho podmínky.
- Řízené škálování velmi zvyšuje efektivitu metod s proměnnou metrikou (podobnou vlastnost má i intervalové škálování).
- Metodu hodnosti 1 je třeba nahradit jinou metodou (například metodou BFGS) v případě, že hodnota η^{R1} nevychází kladná (dostaneme tak metodu s parametrem $\eta = \eta^{R1+}$).
- Efektivita metody BFGS může být překonána vhodnou volbou parametru η (například volbou $\eta = \eta^{VL+}$).
- Metody s proměnou metrikou jsou pro standardní (husté) úlohy menších rozměrů (do 250 proměnných) mnohem efektivnější než metoda CG. To samozřejmě neplatí pro rozsáhlé úlohy, pro které je buď nemožné nebo nevhodné pracovat s plnými maticemi.

V tabulce 4 jsou uvedeny výsledky získané testováním některých modifikací metod s proměnnou metrikou (jsou použity stejné testovací úlohy jako v předchozím případě). Byly testovány tyto metody:

MBFGS - dvoukroková metoda BFGS využívající kvazinewtonovskou podmínku (465),
DBFGS - analogie metody BFGS z Davidonovy třídy (poznámka 193),
BFGS - metoda Broydena, Fletchera, Goldfarba a Shanna (288) bez korekce (s $\rho = 1$),
VL+ - variačně odvozená metoda (466), (467) bez korekce (s $\rho = 1$),

Tyto metody (kromě metody MBFGS) používají buď základní vztah (451) nebo úsporné vzorce (452) a (513).

	Směrový vektor (451)				Směrový vektor (452) a (513)			
Metoda	NIT	NFV	selhání	čas	NIT	NFV	selhání	čas
MBFGS/NS	28015	54305	-	7.70		–		
MBFGS/PS	28640	32109	-	4.27		–		
MBFGS/CS	18248	24279	-	2.73		–		
DBFGS/NS	25117	51768	-	7.34	26150	53061	-	6.77
DBFGS/PS	29266	34218	-	4.21	30936	36104	-	3.33
DBFGS/CS	18313	23857	-	2.71	18401	23721	-	2.05
BFGS/NS	30177	55759	-	7.27	30187	55658	-	6.19
BFGS/PS	36268	39434	-	5.28	35206	37620	-	3.83
BFGS/CS	19428	22134	-	2.77	19919	22682	-	2.09
VL+/NS	27429	43773	-	6.17	27020	43996	-	5.07
VL+/PS	28835	31293	-	4.02	28412	30733	-	2.90
VL+/CS	18800	20678	-	2.71	19015	20801	-	1.99

Tabulka 4: TEST28 – 91 úloh

Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Modifikované metody MBFGS a DBFGS jsou efektivnější (měřeno počtem vyčíslených funkčních hodnot) než standardní metoda BFGS, používáme-li je bez škálování nebo s počátečním škálováním. Řízené škálování zvyšuje efektivitu všech uvedených metod, nejvíce se to však projeví u standardních metod s proměnnou metrikou, které pak dávají lepší výsledky.
- Použití úsporných vzorců (452) a (513) nezhoršuje rychlost konvergence metod s proměnnou metrikou, zvyšuje však jejich efektivitu měřenou dobou výpočtu.
- Metoda VL+ se zdá být opět nejvhodnější.

5 Metody s lokálně omezeným krokem

Newtonova netoda, realizovaná jako metoda spádových směrů, určuje směrový vektor s_i přesným nebo nepřesným řešením soustavy rovnic $B_i s_i + g_i = 0$, kde $B_i = G_i$. Je-li symetrická matice B_i pozitivně definitní, je vektor s_i globálním minimem kvadratické funkce (516) nebo jeho aproximací, čili přesným nebo nepřesným řešemím úlohy

$$s_i = \arg \min_{s \in R^n} Q_i(s). \tag{514}$$

Tento přístup nelze použít, je-li matice B_i indefinitní (takže funkce $Q_i(x)$ není zdola omezená a úloha (514) nemá konečné řešení). Abychom získali konečné řešení, je třeba připojit k úloze (514) dodatečné omezení $\|s_i\| \leq \Delta_i$, kde Δ_i je poloměr nadkoule, která se nazývá oblastí přijatelnosti. Metody s lokálně omezeným krokem určují směrový vektor s_i přesným nebo nepřesným řešemím úlohy

$$s_i = \arg \min_{\|s\| \leq \Delta_i} Q_i(s), \tag{515}$$

splňujícím podmínky (T1), uvedené v definici 38. Výběr délky koku podle (T2) je velmi jednoduchý. Důležitou součástí metod s lokálně omezeným krokem je adaptivní určování poloměru oblasti přijatelnosti podle pravidel (T3).

5.1 Základní vlastnosti metod s lokálně omezeným krokem

Při výkladu metod s lokálně omezeným krokem budeme používat označení

$$Q_i(s) = g_i^T s + \frac{1}{2} s^T B_i s \tag{516}$$

pro kvadratickou funkci, která lokálně aproximuje rozdíl $F(x_i + s) - F(x_i)$ a označení

$$\omega_i(s) = \frac{B_i s + g_i}{\|g_i\|} \tag{517}$$

pro přesnost určení směrového vektoru (předpokládáme, že $\|g_i\| \neq 0$, neboť v opačném případě je bod x_i stacionárním bodem funkce F). Dále budeme používat označení

$$\rho_i(s) = \frac{F(x_i + s) - F(x_i)}{Q_i(s)} \tag{518}$$

pro podíl skutečného a předpověděného poklesu funkce F.

Definice 38. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou s lokálně omezeným krokem, jestliže směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že*

$$\|s_i\| \leq \overline{\delta} \Delta_i, \tag{T1a}$$

$$\|s_i\| < \underline{\delta} \Delta_i \quad \Rightarrow \quad \|\omega_i(s_i)\| \leq \overline{\omega}_i \leq \overline{\omega}, \tag{T1b}$$

$$-Q_i(s_i) \geq \frac{\underline{\nu}}{2} \|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right), \tag{T1c}$$

kde $0 < \underline{\delta} \leq 1 \leq \overline{\delta}$, $0 < \underline{\nu} \leq 1$ a $0 \leq \overline{\omega} < 1$, délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq 0 \quad \Rightarrow \quad \alpha_i = 0, \tag{T2a}$$

$$\rho_i(s_i) > 0 \quad \Rightarrow \quad \alpha_i = 1, \tag{T2b}$$

a čísla $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se volí tak, že

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \underline{\beta} \|s_i\| \leq \Delta_{i+1} \leq \overline{\beta} \|s_i\|, \tag{T3a}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \Delta_i \leq \Delta_{i+1} \leq \min(\underline{\gamma}\Delta_i, \overline{\Delta}), \tag{T3b}$$

kde $0 < \underline{\rho} < 1$ *a* $0 < \underline{\beta} \leq \overline{\beta} < 1 < \underline{\gamma}$, *přičemž* $\overline{\beta}\,\overline{\delta} < 1$. *Řekneme, že metoda s lokálně omezeným krokem je striktní metodou s lokálně omezeným krokem, jsou-li podmínky (T2a) a (T2b) nahraženy podmínkami*

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \alpha_i = 0, \tag{T2c}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1. \tag{T2d}$$

Poznámka 195. Zvolíme-li $\overline{\omega} = 0$ nebo $\overline{\omega} > 0$, dostaneme přesné nebo nepřesné metody s lokálně omezeným krokem. Případ, kdy $\underline{\delta} < 1 < \overline{\delta}$, má význam při přibližném výpočtu optimálního lokálně omezeného kroku. V ostatních případech lze pokládat $\underline{\delta} = 1$ a $\overline{\delta} = 1$. Číslo $\underline{\nu}$ není vnějším parametrem metody. Jeho existence musí být zaručena, ale jeho velikost závisí na zvolené metodě (obvykle $\underline{\nu} = 1$). Podmínka (T3b) se obvykle realizuje tak, že

$$\underline{\rho} \leq \rho_i(s_i) \leq \overline{\rho} \quad \Rightarrow \quad \Delta_{i+1} = \Delta_i, \tag{T3c}$$

$$\rho_i(s_i) > \overline{\rho} \quad \Rightarrow \quad \Delta_{i+1} = \min(\underline{\gamma}\Delta_i, \overline{\Delta}), \tag{T3d}$$

kde $0 < \underline{\rho} < \overline{\rho} < 1$ a $\underline{\gamma} > 1$. Číslo $\overline{\Delta} > 0$ slouží k omezení délky kroku, abychom se nedostali mimo definiční obor funkce F (nebo mimo oblast kde tato funkce nabývá rozumných hodnot). Nerovnost na levé straně (T3b) lze zapsat ve tvaru

$$F(x_i) - F(x_{i+1}) \geq -\underline{\rho} Q_i(s_i). \tag{519}$$

Poznámka 196. Jelikož z (T2c) plyne (T2a), platí pro striktní metody s lokálně omezeným krokem stejná tvrzení jako pro obecné metody s lokálně omezeným krokem. Striktní metody s lokálně omezeným krokem mají poněkud výhodnější teoretické vlastnosti (platí pro ně věta 119 a věta 121). Co se týče numerické efektivity, oba typy metod s lokálně omezeným krokem se příliš neliší. Podmínka (T2c) vyžaduje menší hodnotu parametru $\underline{\rho}$ (například $\underline{\rho} = 0.01$) než podmínka (T2a) (kde stačí $\underline{\rho} = 0.1$). Striktní metody s lokálně omezeným krokem lze modifikovat tak, že použijeme dva parametry $0 < \underline{\rho}_1 < \underline{\rho}_2 < 1$, přičemž $\underline{\rho}_1$ vystupuje v (T2) a $\underline{\rho}_2$ vystupuje v (T3).

Poznámka 197. V podmínce (T1c) se někdy používá $\|\underline{s}_i\|$ místo Δ_i, což je možné, neboť podle (T1a) platí $\Delta_i \geq \|s_i\|/\overline{\delta}$, takže nahradíme-li číslo $\underline{\nu}$ podílem $\underline{\nu}/\overline{\delta}$, zůstane platnost podmínky (T1c) zachována. Navíc Δ_i se v podmínce (T1c) uplatňuje většinou tehdy, když $\|s_i\| \geq \underline{\delta}\Delta_i$ (viz důkaz věty 123).

Poznámka 198. Normy v (T1) a (T3) mohou být i jiné než eukleidovské. V tomto případě se využívá ekvivalence norem. Pro libovolnou vektorovou normu $\|s\|_*$ existují čísla $\underline{\xi}$ a $\overline{\xi}$ taková, že $\underline{\xi}\|s\| \leq \|s\|_* \leq \overline{\xi}\|s\|$ $\forall s \in R^n$. Podíl $\overline{\xi}/\underline{\xi}$ pak vystupuje v odpovídajících vzorcích.

Poznámka 199. Při vyšetřování metod s lokálně omezeným krokem budeme používat označení

$$\begin{aligned} N_1 &= \{i \in N : \|s_i\| < \underline{\delta}\Delta_i\}, \\ N_2 &= \{i \in N : \rho_i(s_i) \geq \underline{\rho}\}. \end{aligned}$$

Poznamenejme, že z $i \in N_2$ plyne $\Delta_{i+1} \geq \Delta_i$ a že $i \notin N_2$ implikuje $\Delta_{i+1} \leq \overline{\beta}\Delta_i < \Delta_i$.

Nejprve dokážeme globální konvergenci metod s lokálně omezeným krokem.

Lemma 57. *Aplikujeme-li metodu s lokálně omezeným krokem na funkci* $F : \mathcal{D} \to R$, *která splňuje předpoklad F3, existuje konstanta* $0 < \underline{c} < 1$ *taková, že*

$$\|s_i\| \geq \underline{c}\,\frac{m_i}{M_i}, \tag{520}$$

kde

$$m_i = \min_{1\le j\le i} \|g_j\|,$$
$$M_i = \max_{1\le j\le i} \|B_j\|.$$

Důkaz (a) Nechť $i \in N_1$. Pak podle (T1b) platí

$$|\|B_i s_i\| - \|g_i\|| \le \|B_i s_i + g_i\| = \|\omega_i(s_i)\|\|g_i\| \le \overline{\omega}\|g_i\|,$$

takže buď $\|B_i s_i\| \ge \|g_i\|$ nebo $\|B_i s_i\| < \|g_i\|$ a $\|B_i s_i\| \ge (1-\overline{\omega})\|g_i\|$. Spojením těchto nerovností dostaneme $\|B_i\|\|s_i\| \ge \|B_i s_i\| \ge (1-\overline{\omega})\|g_i\|$, což dává

$$\|s_i\| \ge (1-\overline{\omega})\frac{\|g_i\|}{\|B_i\|} \ge (1-\overline{\omega})\frac{m_i}{M_i}. \tag{521}$$

(b) Nechť $i \notin N_1$ a $i \notin N_2$. Pak podle definice množiny N_2 a funkce $Q_i(s)$ platí

$$F(x_i+s_i) - F(x_i) \ge \underline{\rho} Q_i(s_i) = \underline{\rho}\left(g_i^T s_i + \frac{1}{2}s_i^T B_i s_i\right) \ge \underline{\rho}\left(g_i^T s_i - \|B_i\|\|s_i\|^2\right).$$

Z druhé strany použitím (12) dostaneme

$$F(x_i+s_i) - F(x_i) \le g_i^T s_i + \overline{G}\|s_i\|^2,$$

což dohromady dává

$$(\overline{G} + \underline{\rho}\|B_i\|)\|s_i\|^2 \ge (\underline{\rho}-1)g_i^T s_i.$$

Podle (T1c) platí

$$-\frac{\underline{\nu}}{2}\|g_i\|\min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right) \ge Q_i(s_i) \ge g_i^T s_i - \|B_i\|\|s_i\|^2,$$

což spolu s předchozí nerovností dává

$$(\overline{G} + \underline{\rho}\|B_i\|)\|s_i\|^2 \ge (\underline{\rho}-1)g_i^T s_i \ge (\underline{\rho}-1)\|B_i\|\|s_i\|^2 - \frac{\underline{\nu}(\underline{\rho}-1)}{2}\|g_i\|\min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right),$$

neboli

$$(\overline{G} + \|B_i\|)\|s_i\|^2 \ge \frac{\underline{\nu}(1-\underline{\rho})}{2}\|g_i\|\min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right),$$

takže buď

$$\|s_i\| \ge \underline{\delta}\Delta_i \ge \underline{\delta}\frac{\|g_i\|}{\|B_i\|} \ge \underline{\delta}\frac{m_i}{M_i}, \tag{522}$$

nebo

$$(\overline{G} + \|B_1\|)\frac{M_i}{\|B_1\|}\|s_i\|^2 \ge (\overline{G}+M_i)\|s_i\|^2 \ge (\overline{G}+\|B_i\|)\|s_i\|^2 \ge \frac{\underline{\nu}(1-\underline{\rho})}{2}\|g_i\|\Delta_i \ge \frac{\underline{\nu}(1-\underline{\rho})}{2\overline{\delta}}\|g_i\|\|s_i\|,$$

což dává

$$\|s_i\| \ge \frac{\underline{\nu}(1-\underline{\rho})\|B_1\|}{2\overline{\delta}(\overline{G}+\|B_1\|)}\frac{\|g_i\|}{M_i} \ge \frac{\underline{\nu}(1-\underline{\rho})\|B_1\|}{2\overline{\delta}(\overline{G}+\|B_1\|)}\frac{m_i}{M_i}. \tag{523}$$

(c) Nechť $i = 1$. Pokud $\|g_1\| = 0$, platí zřejmě $\|s_1\| \ge \|g_1\|/\|B_1\| \ge m_1/M_1$. Pokud $\|g_1\| \ne 0$, můžeme psát

$$\|s_1\| = \frac{\|s_1\|\|B_1\|}{\|g_1\|}\frac{\|g_1\|}{\|B_1\|} \ge \frac{\|s_1\|\|B_1\|}{\|g_1\|}\frac{m_1}{M_1} \tag{524}$$

(d) Nechť $i \notin N_1$, $i \in N_2$ a $i \neq 1$. Nechť $k < i$ je největší index pro který neplatí současně $k \notin N_1$, $k \in N_2$ a $k \neq 1$. Pak podle (T1) a (T3) platí

$$\|s_i\| \geq \underline{\delta}\Delta_i \geq \underline{\delta}\Delta_{i-1} \geq \cdots \geq \underline{\delta}\Delta_{k+1} \geq \underline{\beta}\,\underline{\delta}\|s_k\|,$$

takže podle (521)–(524) platí

$$\|s_i\| \geq \underline{\beta}\,\underline{\delta}\|s_k\| \geq \underline{c}\,\frac{\|g_k\|}{M_k} \geq \underline{c}\,\frac{m_k}{M_k} \geq \underline{c}\,\frac{m_i}{M_i},$$

kde

$$\underline{c} = \underline{\beta}\,\underline{\delta}\min\left((1-\overline{\omega}),\, \underline{\delta},\, \frac{\underline{\nu}(1-\underline{\rho})\|B_1\|}{2\overline{\delta}(\overline{G}+\|B_1\|)},\, \frac{\|s_1\|\|B_1\|}{\|g_1\|}\right). \tag{525}$$

□

Poznámka 200. Z části (d) důkazu lemmatu 57 plyne, že pro libovolný index $i \in N_2$ existuje index $k \leq i$ a číslo $0 < \underline{c} < 1$ tak, že $\|s_i\| \geq \underline{c}\,\|g_k\|/M_k$.

Lemma 58. *(Powell) Nechť Δ_i, $i \in N$, a M_i, $i \in N$, jsou dvě posloupnosti kladných čísel a $N_2 \subset N$. Nechť*

$$\Delta_i \geq \frac{\mu}{M_i} > 0, \quad i \in N, \tag{526}$$

kde $\mu > 0$,

$$\Delta_{i+1} \leq \gamma\Delta_i, \quad i \in N_2, \tag{527}$$
$$\Delta_{i+1} \leq \beta\Delta_i, \quad i \notin N_2, \tag{528}$$
$$M_{i+1} \geq M_i, \quad i \in N, \tag{529}$$

kde $0 < \beta < 1 < \gamma$, a

$$\sum_{i\in N_2} \frac{1}{M_i} < \infty. \tag{530}$$

Pak

$$\sum_{i\in N} \frac{1}{M_i} < \infty. \tag{531}$$

Důkaz (a) Nechť $i \in N$, nechť r je přirozené číslo takové, že $\beta^{r-1}\gamma < 1$ (takové číslo existuje neboť $\beta < 1$ a $\gamma < \infty$) a nechť $p(i)$ je počet indexů z množiny $\{1, \ldots, i\}$, které jsou prvky množiny N_2 (čili $p(i)$ je mohutnost množiny $\{1, \ldots, i\} \cap N_2$). Nechť $i \in N_4$, kde

$$N_4 = \{i \in N : rp(i) < i\}.$$

Pak podle (527) a (528) pro $i \in N_4$ platí

$$\Delta_i \leq \gamma^{p(i-1)}\beta^{i-1-p(i-1)}\Delta_1 \leq \gamma^{(i-1)/r}\beta^{(r-1)(i-1)/r}\Delta_1 \leq \left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1.$$

Protože podle předpokladu je $\gamma\beta^{r-1} < 1$, můžeme psát

$$\sum_{i\in N_4}\Delta_i \leq \sum_{i\in N_4}\left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1 \leq \sum_{i=1}^{\infty}\left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1 = \frac{\Delta_1}{1-\left(\gamma\beta^{(r-1)}\right)^{1/r}} < \infty.$$

Použijeme-li nyní (526), dostaneme

$$\sum_{i\in N_4}\frac{1}{M_i} \leq \frac{1}{\mu}\sum_{i\in N_4}\Delta_i < \infty.$$

(b) Nyní stačí dokázat, že

$$\sum_{i\in N_5}\frac{1}{M_i}<\infty,$$

kde $N_5 = N \setminus N_4$, takže $N_5 = \{i \in N : rp(i) \geq i\}$. Označme

$$N_2 = \{i_1, i_2, i_3 \dots\}, \qquad N_5 = \{k_1, k_2, k_3 \dots\}$$

(předpokládáme uspořádání prvků podle velikosti) a sestrojme množinu

$$N_6 = \{l_1, l_2, l_3 \dots\} = \{\underbrace{i_1, \dots, i_1}_{r-\text{krát}}, \underbrace{i_2, \dots, i_2}_{r-\text{krát}}, \underbrace{i_3, \dots, i_3}_{r-\text{krát}}, \dots\}.$$

Z konstrukce množiny N_5 plyne, že

$$rp(k_j) \geq k_j \geq j \quad \forall j \in N,$$

takže podle definice množiny N_6 dostaneme

$$l_j \leq l_{rp(k_j)} = i_{p(k_j)} \leq k_j \quad \forall j \in N,$$

neboť $i_{p(k_j)}$ je poslední prvek množiny $\{1, \dots, k_j\} \cap N_2$. Podle (529) tedy platí $M_{l_j} \leq M_{k_j}$ $\forall j \in N$, takže podle (530) dostaneme

$$\sum_{i\in N_5}\frac{1}{M_i}=\sum_{j=1}^{\infty}\frac{1}{M_{k_j}}\leq\sum_{j=1}^{\infty}\frac{1}{M_{l_j}}=\sum_{i\in N_6}\frac{1}{M_i}=r\sum_{i\in N_2}\frac{1}{M_i}<\infty,$$

což spolu s (a) dává (531). □

Věta 118. *(globální konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem taková, že*

$$\sum_{i=1}^{\infty}\frac{1}{M_i}=\infty, \tag{532}$$

kde M_i, $i \in N$, jsou čísla definovaná v lemmatu 57. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak platí

$$\liminf_{i\to\infty}\|g_i\|=0. \tag{533}$$

Důkaz (a) Předpokládejme, že existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$. Pak podle (T1a) a (520) platí

$$\Delta_i \geq \frac{1}{\overline{\delta}}\|s_i\| \geq \frac{\underline{c}\,\underline{\varepsilon}}{\overline{\delta} M_i} \triangleq \frac{\mu}{M_i} \tag{534}$$

$\forall i \in N$. Použijeme-li (T2b), (T1c) a (534) můžeme psát

$$F_i - F_{i+1} = F(x_i) - F(x_i + s_i) \geq -\underline{\rho} Q_i(s_i) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{2}\min\left(\Delta_i, \frac{\underline{\varepsilon}}{M_i}\right) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}^2\underline{c}}{2\overline{\delta}}\frac{1}{M_i}$$

$\forall i \in N_2$, takže

$$F_1 - \underline{F} \geq \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i\in N_2}(F_i - F_{i+1}) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}^2\underline{c}}{2\overline{\delta}}\sum_{i\in N_2}\frac{1}{M_i}.$$

Platí tedy

$$\sum_{i\in N_2}\frac{1}{M_i}<\infty.$$

(b) Položíme-li $\beta = \overline{\beta}\,\overline{\delta} < 1$ a $\gamma = \underline{\gamma} > 1$, jsou splněny předpoklady lemmatu 58, takže platí (531) což je ve sporu s předpokladem (532). □

Poznámka 201. Předpoklady věty 118 jsou splněny například tehdy, jsou-li matice B_i, $i \in N$, stejnoměrně omezené, kdy platí

$$\|B_i\| \le \overline{B}, \qquad i \in N.$$

Důkaz tohoto dílčího tvrzení je velmi jednoduchý. Stačí část (a) důkazu věty 118 pozměnit tak, že

$$F_1 - \underline{F} \ge \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \ge \sum_{i\in N_2}(F_i - F_{i+1}) \ge \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}^2\underline{c}}{2\overline{\delta}} \sum_{i\in N_2}\frac{1}{\overline{B}}.$$

Je-li množina N_2 nekonečná, dojdeme ihned ke sporu. Je-li množina N_2 konečná, musí podle (T3a) platit $\Delta_i \to 0$, což je ve sporu s (534), neboť $M_i \le \overline{B}$.

Poznámka 202. Předpoklady věty 118 jsou splněny také tehdy, jsou-li matice B_i, $i \in N$, dostatečně omezené, čili tehdy, když platí

$$\|B_i\| \le C_i, \qquad i \in N$$

a čísla C_i vyhovují rekurentním nerovnostem

$$C_{i+1} \le C_i + \overline{C}\|s_i\| \le C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}\,i,$$

kde $C_1 > 0$ a $\overline{C} \ge 0$ jsou vhodné konstanty (pak též $M_i \le C_i$, $i \in N$). V tomto případě lze psát

$$\sum_{i=1}^{\infty}\frac{1}{M_i} \ge \frac{1}{C_1} + \sum_{i=1}^{\infty}\frac{1}{C_{i+1}} \ge \frac{1}{C_1} + \sum_{i=1}^{\infty}\frac{1}{C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}\,i} \ge \frac{1}{C_1} + \frac{1}{C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}}\sum_{i=1}^{\infty}\frac{1}{i} = \infty,$$

neboť harmonická řada je divergentní.

V poznámce 29 jsme ukázali, že pro metody stejnoměrně spádových směrů platí $\|g_i\| \to 0$. Nyní dokážeme, že totéž platí pro striktní metody s lokálně omezeným krokem, jsou-li matice B_i, $i \in N$, stejnoměrně omezené. Tato věta neplatí pro obecné metody s lokálně omezeným krokem, neboť vyžaduje platnost podmínky (T2c).

Věta 119. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná striktní metodou s lokálně omezeným krokem takovou, že $\|B_i\| \le \overline{B}$, $i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak platí*

$$\lim_{i\to\infty}\|g_i\| = 0.$$

Důkaz V poznámce 201 jsme ukázali, že

$$\liminf_{i\to\infty}\|g_i\| = 0. \tag{535}$$

Předpokládejme navíc, že

$$\limsup_{i\to\infty}\|g_i\| > \underline{\varepsilon} > 0.$$

Za tohoto předpokladu je množina N_2 nekonečná a obsahuje nekonečnou podmnožinu $\overline{N}_2 \subset N_2$ takovou, že $\|g_i\| \ge \underline{\varepsilon}$, $i \in \overline{N}_2$ (kdyby N_2 byla konečná, existoval by podle (T2c) index $k \in N$ takový, že $x_i = x_k$ $\forall i \ge k$, a podle (535) též $\|g_i\| = 0$ $\forall i \ge k$). Označme

$$\overline{N}_2 = \{k_1, k_2, k_3, \dots\}.$$

Jelikož posloupnost $F(x_{k_j})$, $j \in N$, je podle (T2) nerostoucí a podle předpokladu F1 zdola omezená, má tato posloupnost limitu. Existuje tedy index $m \in N$ takový, že

$$F(x_{k_j}) - F(x_{k_{j+1}}) < \underline{\rho}\,\frac{\underline{\nu}\,\underline{\varepsilon}^2}{8\overline{\delta}^2}\min\left(\frac{1}{\overline{B}},\ \frac{1}{\overline{G}}\right), \quad \forall j \ge m. \tag{536}$$

Nechť $j \geq m$ a l_j je nevětší index takový, že $k_j \leq l_j < k_{j+1}$, přičemž pro $k_j \leq l \leq l_j$ platí $\|g_l\| \geq \underline{\varepsilon}/(2\overline{\delta})$ (takový index existuje, neboť pro $l = k_j \in \overline{N}_2$ platí $\|g_l\| \geq \underline{\varepsilon} > \underline{\varepsilon}/(2\overline{\delta})$). Pak lze podle (519) a (T1c) psát

$$F(x_l) - F(x_{l+1}) \geq -\underline{\rho} Q_l(s_l) \geq \frac{\underline{\rho}\,\underline{\nu}}{2}\|g_l\| \min\left(\Delta_l, \frac{\|g_l\|}{\|B_l\|}\right) \geq \underline{\rho}\,\frac{\underline{\nu}\,\underline{\varepsilon}}{4\overline{\delta}^2} \min\left(\|s_l\|, \frac{\underline{\varepsilon}}{2\overline{B}}\right), \quad \forall k_j \leq l \leq l_j,$$

což spolu s (536) dává

$$\begin{aligned} \underline{\rho}\,\frac{\underline{\nu}\,\underline{\varepsilon}^2}{8\overline{\delta}^2} \min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{G}}\right) &> F(x_{k_j}) - F(x_{k_{j+1}}) \geq F(x_{k_j}) - F(x_{l_j+1}) \\ &= \sum_{l=k_j}^{l_j} (F(x_l) - F(x_{l+1})) \geq \underline{\rho}\,\frac{\underline{\nu}\,\underline{\varepsilon}}{4\overline{\delta}^2} \sum_{l=k_j}^{l_j} \min\left(\|s_l\|, \frac{\underline{\varepsilon}}{2\overline{B}}\right). \end{aligned}$$

Porovnáme-li obě strany této nerovnosti, vidíme, že případ, kdy $\|s_l\| \geq \underline{\varepsilon}/(2\overline{B})$ nemůže pro $k_j \leq l \leq l_j$ nastat (v opačném případě by pravá strana nebyla menší než levá). Můžeme tedy psát

$$\sum_{l=k_j}^{l_j} \|s_l\| < \frac{\underline{\varepsilon}}{2} \min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{G}}\right) \leq \frac{\underline{\varepsilon}}{2\overline{G}}.$$

Použijeme-li tuto nerovnost spolu s nerovností (15), dostaneme

$$\|g(x_{k_j}) - g(x_{l_j+1})\| \leq \overline{G}\|x_{k_j} - x_{l_j+1}\| \leq \overline{G} \sum_{l=k_j}^{l_j} \|s_l\| < \frac{\underline{\varepsilon}}{2}.$$

Jelikož posloupnost $\overline{N}_2$ je nekonečná a platí (535), musí existovat index $j \geq m$ takový, že $l_j + 1 < k_{j+1}$ (a tedy $\|g_{l_j+1}\| < \underline{\varepsilon}/2$). Pak podle toho co jsme dokázali platí

$$\|g(x_{k_j})\| \leq \|g(x_{l_j+1})\| + \|g(x_{k_j}) - g(x_{l_j+1})\| < \frac{\underline{\varepsilon}}{2} + \frac{\underline{\varepsilon}}{2} = \underline{\varepsilon},$$

což je ve sporu s předpokladem, že $\|g_{k_j}\| \geq \underline{\varepsilon}\ \forall k_j \in \overline{N}_2$. □

V další části tohoto oddílu budeme předpokládat, že $x_i \to x^*$ a že bod $x^* \in R^n$ vyhovuje postačujícím podmínkám pro extrém (věta 4). Abychom nemuseli stále ověřovat, zda pro daný index $i \in N$ již platí $x_i \in \mathcal{B}(x^*, \varepsilon)$, nahradíme předpoklady věty 4 silnějšími předpoklady (F4) a (F5). Tím se formálně zjednoduší a zpřehlední většina důkazů aniž by došlo k újmě na obecnosti.

Poznámka 203. V následujících dvou větách budeme potřebovat aby byla splněna podmínka

$$\Delta_{i+1} \leq \overline{\gamma}\|s_i\| \qquad \forall i \in N_2, \tag{537}$$

kde $\overline{\gamma}\,\underline{\delta} > 1$. Tuto podmínku splníme snadno tak, že položíme $\Delta_{i+1} = \overline{\gamma}\|s_i\|$, pokud v (T3c) vyjde $\Delta_{i+1} > \overline{\gamma}\|s_i\|$. Jelikož $\overline{\gamma}\,\underline{\delta} > 1$, může tento případ nastat pouze tehdy, když $i \in N_1$. Můžeme se snadno přesvědčit, že po této úpravě zůstane zachována platnost obou vět o globální konvergenci i platnost věty 122 o superlineární konvergenci. Pouze věta 126, týkající se Newtonovy metody s kvazioptimálním lokálně omezeným krokem, tuto úpravu nepovoluje. Podmínku (537) potřebujeme k odvození nerovnosti (538).

Nejprve ukážeme, že jsou-li matice B_i, $i \in N$, dostatečně omezené (poznámka 202) a platí-li (F4), (F5) a (537), jsou tyto matice stejnoměrně omezené. To má význam při vyšetřování superlineární konvergence metod s proměnnou metrikou pro řídké úlohy (věta 223)).

Věta 120. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem, pro kterou platí (537). Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která*

vyhovuje předpokladům F4 a F5. Pak, jsou-li matice B_i, $i \in N$, dostatečně omezené, jsou stejnoměrně omezené a platí

$$\sum_{i=1}^{\infty} \|s_i\| < \infty.$$

Důkaz (a) Nechť $k \in N_2$ a $l \in N_2$ jsou dva indexy takové že $j \notin N_2$ $\forall k < j < l$. Pak podle (T1), (T3) a (537) platí

$$\|s_j\| \le \overline{\delta}\Delta_j \le \overline{\beta}\,\overline{\delta}\|s_{j-1}\| \le \cdots \le (\overline{\beta}\,\overline{\delta})^{j-k-1}\|s_{k+1}\| \le \frac{1}{\underline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\Delta_{k+1} \le \frac{\overline{\gamma}}{\underline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\|s_k\| \tag{538}$$

pro $k < j < l$, neboli

$$\sum_{j=k}^{l-1}\|s_j\| \le \frac{\overline{\gamma}}{\underline{\beta}}\|s_k\|\sum_{j=k}^{l-1}(\overline{\beta}\,\overline{\delta})^{j-k} \le \frac{\overline{\gamma}}{\underline{\beta}}\|s_k\|\sum_{j=k}^{\infty}(\overline{\beta}\,\overline{\delta})^{j-k} = \frac{\overline{\gamma}}{\underline{\beta}(1-\overline{\beta}\,\overline{\delta})}\|s_k\| \triangleq \overline{D}\|s_k\|,$$

takže pro libovolný index $i \in N$ platí

$$C_1 + \sum_{j=1}^{i}\overline{C}\|s_j\| \le \overline{D}(C_1 + \sum_{j\in N_2}^{j\le i}\overline{C}\|s_j\|) \tag{539}$$

(předpokládáme bez újmy na obecnosti, že množina N_2 obsahuje index $i = 1$).

(b) Nyní můžeme postupovat podobně jako v důkazu věty 16. Použijeme-li (T1) a (T2), dostaneme

$$\frac{F_i - F_{i+1}}{\|g_i\|} \ge -\underline{\rho}\frac{Q_i(s_i)}{\|g_i\|} \ge \frac{\underline{\rho}\,\underline{\nu}}{2}\min\left(\Delta_i, \frac{\|g_i\|}{C_i}\right) \ge \frac{\underline{\rho}\,\underline{\nu}}{2\overline{\delta}}\min\left(\|s_i\|, \frac{\|g_i\|}{C_i}\right) \ge \frac{\underline{\rho}\,\underline{\nu}}{2\overline{\delta}}\frac{\|g_i\|\|s_i\|}{\|g_i\| + C_i\|s_i\|} \tag{540}$$

pro $i \in N_2$, neboť pro libovolná kladná čísla a, b platí $\min(a,b) \ge ab/(a+b)$. Dále podle (F4) platí

$$0 \ge F_{i+1} - F_i \ge s_i^T g_i + \frac{1}{2}\underline{G}\|s_i\|^2 \ge -\|s_i\|\|g_i\| + \frac{1}{2}\underline{G}\|s_i\|^2,$$

neboli

$$\|s_i\| \le \frac{2}{\underline{G}}\|g_i\| \tag{541}$$

pro $i \in N_2$ (bez újmy na obecnosti budeme předpokladat, že $\underline{G} \le C_1$, takže $\underline{G} \le C_i$ $\forall i \in N_2$). Dosadíme-li tento vztah do (540) a použijeme-li (24), dostaneme

$$\begin{aligned}\frac{F_i - F_{i+1}}{\|g_i\|} &\ge \frac{\underline{\rho}\,\underline{\nu}}{2\overline{\delta}}\frac{\underline{G}\|g_i\|\|s_i\|}{\underline{G}\|g_i\| + 2C_i\|g_i\|} \ge \frac{\underline{\rho}\,\underline{\nu}\,\underline{G}}{6\overline{\delta}}\frac{\|s_i\|}{C_i} \ge \frac{\underline{\rho}\,\underline{\nu}\,\underline{G}}{6\overline{\delta}\,\overline{C}}\frac{\overline{C}\|s_i\|}{C_1 + \sum_{j=1}^{i}\overline{C}\|s_j\|} \\ &\ge \frac{\underline{\rho}\,\underline{\nu}\,\underline{G}}{6\overline{\delta}\,\overline{C}\,\overline{D}}\frac{\overline{C}\|s_i\|}{C_1 + \sum_{j\in N_2}^{j\le i}\overline{C}\|s_j\|}\end{aligned}$$

pro $i \in N_2$, takže podobně jako v důkazu věty 16 platí

$$\frac{\underline{\rho}\,\underline{\nu}\,\underline{G}}{6\overline{\delta}\,\overline{C}\,\overline{D}}\sum_{i\in N_2}\frac{\overline{C}\|s_i\|}{C_1 + \sum_{j\in N_2}^{j\le i}\overline{C}\|s_j\|} \le \sum_{i\in N_2}\frac{F_i - F_{i+1}}{\|g_i\|} \le \sum_{i=1}^{\infty}\frac{F_i - F_{i+1}}{\|g_i\|} \le \frac{\sqrt{2\overline{G}}}{\underline{G}}\sqrt{F_1 - F^*},$$

takže součet na levé straně je konečný, a podobně jako v důkazu věty 12 existuje číslo $\underline{C}$ takové, že

$$C_k \le C_1 + \sum_{j=1}^{k}\overline{C}\|s_j\| \le \overline{D}(C_1 + \sum_{j\in N_2}^{j\le k}\overline{C}\|s_j\|) \le \frac{\overline{D}C_1}{\underline{C}}$$

pro $k \in N$, takže $\|B_k\| \leq C_k \leq \overline{B}$, kde $\overline{B} = \overline{D}C_1/\underline{C}$. Z toho, že $C_1 + \sum_{j=1}^{k} \overline{C}\|s_j\| \leq \overline{B}$ pro $k \in N$ plyne nerovnost

$$\sum_{i=1}^{\infty} \|s_i\| < \infty.$$

□

Pro striktní metody s lokálně omezeným krokem platí silnější tvrzení.

Věta 121. *(lineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná striktní metodou s lokálně omezeným krokem, pro kterou platí (537). Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje předpokladům F4 a F5. Pak, jsou-li matice B_i, $i \in N$, stejnoměrně omezené, platí*

$$\sum_{i=1}^{\infty} \|x_i - x^*\| < \infty.$$

Důkaz (a) Dokážeme nejprve, že posloupnost x_i, $i \in N_2$, je lineárně konvergentní. Důkaz tohoto dílčího tvrzení je velmi podobný důkazu věty 17. Nechť $i \in N_2$. Podle poznámky 200 existuje index $k \leq i$ takový, že $\|s_i\| \geq \underline{c}\,\|g_k\|/\overline{B}$. Jelikož posloupnost $F(x_i)$, $i \in N$, je nerostoucí, podle (43) a (49) platí

$$1 \geq \frac{F(x_i) - F(x^*)}{F(x_k) - F(x^*)} \geq \frac{\underline{G}^2}{2} \frac{\|x_i - x^*\|^2}{\|g_k\|^2} \geq \frac{\underline{G}^2}{2\overline{G}^2} \frac{\|g_i\|^2}{\|g_k\|^2},$$

takže

$$\|s_i\| \geq \frac{\underline{c}\,\underline{G}}{\sqrt{2}\,\overline{B}\,\overline{G}} \|g_i\|. \tag{542}$$

Jelikož $i \in N_2$, platí $\rho_i(s_i) \geq \underline{\rho}$, což spolu s (T1c) dává

$$F_i - F_{i+1} \geq \frac{\underline{\rho}\,\underline{\nu}}{2} \|g_i\|^2 \min\left(\frac{\underline{c}\,\underline{G}}{\sqrt{2}\,\overline{\delta}\,\overline{B}\,\overline{G}}, \frac{1}{\overline{B}} \right) = \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{G}}{2\sqrt{2}\,\overline{\delta}\,\overline{B}\,\overline{G}} \|g_i\|^2 \geq \frac{c}{\overline{G}} \|g_i\|^2.$$

kde

$$c = \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{G}}{2\sqrt{2}\,\overline{\delta}\,\overline{B}} < \frac{\underline{\rho}\,\underline{\nu}\,\underline{G}}{4\sqrt{2\delta}\,\overline{G}} < 1,$$

neboť podle (525) platí

$$\underline{c} < \frac{\underline{\nu}(1-\underline{\rho})\|B_1\|}{2\overline{\delta}(\overline{G} + \|B_1\|)} \leq \frac{\underline{\nu}(1-\underline{\rho})\overline{B}}{2\overline{\delta}\,\overline{G}} < \frac{\overline{B}}{2\overline{G}}.$$

Nechť $N_2 = \{k_1, k_2, k_3, \dots\}$. Použijeme-li vztah (49), dostaneme

$$F_{k_{i+1}} - F^* \leq F_{k_i+1} - F^* \leq \left(1 - c\frac{\underline{G}}{\overline{G}}\right)(F_{k_i} - F^*)$$

pro $i \in N$, což implikuje (podobně jako v důkazu věty 17), že posloupnost x_{k_i}, $i \in N$, konverguje k bodu x^* R-lineárně (s kvocientem $q = \sqrt{1 - c\,\underline{G}/\overline{G}}$). Podle poznámky 36 tedy platí

$$\sum_{i \in N_2} \|x_i - x^*\| < \infty.$$

(b) Ukážeme, že pokud

$$\|s_i\| \leq \frac{\underline{\nu}(1-\underline{\rho})}{\overline{\delta}(\overline{G} + \overline{B})} \|g_i\|, \tag{543}$$

platí $\rho_i(s_i) \geq \underline{\rho}$, takže $i \in N_2$. Podmínku $\rho_i(s_i) \geq \underline{\rho}$ lze podle (519) zapsat ve tvaru

$$F_{i+1} - F_i - Q_i(s_i) \leq (\underline{\rho} - 1)Q_i(s_i).$$

Ale

$$F_{i+1} - F_i - Q_i(s_i) \leq s_i^T g_i + \frac{1}{2}\overline{G}\|s_i\|^2 - s_i^T g_i + \frac{1}{2}\overline{B}\|s_i\|^2 = \frac{1}{2}(\overline{G} + \overline{B})\|s_i\|^2$$

a podle (T1c) platí

$$(\underline{\rho} - 1)Q_i(s_i) \geq (1 - \underline{\rho})\frac{\underline{\nu}}{2}\|g_i\| \min\left(\frac{\|s_i\|}{\overline{\delta}}, \frac{\|g_i\|}{\overline{B}}\right),$$

takže podmínka $\rho_i(s_i) \geq \underline{\rho}$ je splněna, pokud

$$\frac{1}{2}(\overline{G} + \overline{B})\|s_i\|^2 \leq (1 - \underline{\rho})\frac{\underline{\nu}}{2}\|g_i\| \min\left(\frac{\|s_i\|}{\overline{\delta}}, \frac{\|g_i\|}{\overline{B}}\right).$$

Z (543) plyne, že $\|s_i\|/\overline{\delta} \leq \|s_i\| \leq \|g_i\|/\overline{B}$ (neboť $0 < \underline{\rho} < 1$, $0 < \underline{\nu} \leq 1$, $\overline{\delta} \geq 1$ a $\overline{G} \geq 0$), takže podmínka $\rho_i(s_i) \geq \underline{\rho}$ je splněna, pokud

$$\frac{1}{2}(\overline{G} + \overline{B})\|s_i\|^2 \leq \frac{(1 - \overline{\rho})\underline{\nu}}{2\overline{\delta}}\|g_i\|\|s_i\|.$$

Tato nerovnost opět plyne z (543).

(c) Nechť $k \in N_2$ a $l \in N_2$ jsou dva indexy takové že $j \notin N_2$ pro $k < j < l$. Pak platí (538), což spolu s (541) dává

$$\|s_j\| \leq \frac{2\overline{\gamma}}{\underline{G}\overline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\|g_k\| = \frac{2\overline{\gamma}}{\underline{G}\overline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\|g_j\|,$$

neboť podle (T2c) pro $k < j < l$ platí $x_j = x_k$ a tedy $\|x_j - x^*\| = \|x_k - x^*\|$ a $\|g_j\| = \|g_k\|$. Jelikož $\overline{\beta}\,\overline{\delta} < 1$, existuje číslo $m \in N$ takové, že

$$\frac{2\overline{\gamma}}{\underline{G}\overline{\beta}}(\overline{\beta}\,\overline{\delta})^m \leq \frac{(1 - \underline{\rho})\underline{\nu}}{\overline{\delta}(\overline{G} + \overline{B})}.$$

Musí tedy být $j - k \leq m$, takže

$$\sum_{i=1}^{\infty} \|x_i - x^*\| \leq m \sum_{i \in N_2} \|x_i - x^*\| < \infty.$$

□

Poznámka 204. Ve větě 121 můžeme podmínku stejnoměrné omezenosti matic B_i nahradit podmínkou jejich dostatečné omezenosti (plyne to z věty 120).

Nyní dokážeme větu o superlineární konvergenci, která má podobný charakter jako věta 20.

Věta 122. *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem takovou, že $\|B_i\| \leq \overline{B}$ $\forall i \in N$. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, splňujícím postačující podmínky druhého řádu pro lokální minimum (matice $G^* = G(x^*)$ je pozitivně definitní). Nechť platí*

$$\lim_{i \to \infty} \overline{\omega}_i = 0, \qquad \lim_{i \to \infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0. \tag{544}$$

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz (a) Nechť $0 < \underline{G} < \underline{\lambda}(G^*) \leq \overline{\lambda}(G^*) < \overline{G}$. Ukážeme, že existuje index $k_1 \in N$ a číslo $\underline{C} > 0$ tak, že pro $i \geq k_1$ platí

$$\|g_i\| \geq \frac{1}{2}\underline{G}\|s_i\| \tag{545}$$

a

$$-Q_i(s_i) \geq \frac{1}{2}\underline{C}\|s_i\|. \tag{546}$$

Označme $\vartheta_i = (B_i - G_i)s_i/\|s_i\|$. Pak platí

$$B_i s_i = G_i s_i + \vartheta_i \|s_i\|,$$

takže

$$\begin{aligned} \|B_i s_i\| &\leq \overline{\lambda}(G_i)\|s_i\| + \|\vartheta_i\|\|s_i\|, \\ s_i^T B_i s_i &\geq \underline{\lambda}(G_i)\|s_i\|^2 - \|\vartheta_i\|\|s_i\|^2 \end{aligned}$$

a jelikož $\|\vartheta_i\| \to 0$ (podle (544)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, $\overline{\lambda}(G_i) \to \overline{\lambda}(G^*) < \overline{G}$, existuje index $k_1 \in N$ takový, že $\|B_i s_i\| \leq \overline{G}\|s_i\|$ a $s_i^T B_i s_i \geq \underline{G}\|s_i\|^2$ $\forall i \geq k_1$. Z definice $Q_i(s_i)$ pak plyne plyne

$$0 \geq Q_i(s_i) = g_i^T s_i + \frac{1}{2} s_i^T B_i s_i \geq \frac{1}{2}\underline{G}\|s_i\|^2 - \|g_i\|\|s_i\|,$$

což dává $\|g_i\| \geq (\underline{G}/2)\|s_i\|$ $\forall i \geq k_1$. Použijeme-li (T1c), můžeme psát

$$-Q_i(s_i) \geq \frac{\nu}{2}\|g_i\| \min\left(\frac{\|s_i\|}{\overline{\delta}}, \frac{\|g_i\|}{\overline{B}}\right) \geq \frac{\nu \underline{G}}{4\overline{\delta}} \min\left(1, \frac{\underline{G}}{2\overline{B}}\right)\|s_i\|^2 \triangleq \frac{1}{2}\underline{C}\|s_i\|^2.$$

(b) Ukážeme, že existuje index $k_2 \geq k_1$ takový, že $i \in N_2$ $\forall i \geq k_2$. Podle věty 5 platí

$$F(x_i + s_i) - F(x_i) = s_i^T g_i + \frac{1}{2} s_i^T G_i s_i + o(\|s_i\|^2) = Q_i(s_i) + \frac{1}{2} s_i^T (G_i - B_i) s_i + o(\|s_i\|^2),$$

takže

$$\rho_i(s_i) = \frac{F(x_i + s_i) - F(x_i)}{Q_i(s_i)} = 1 + \frac{s_i^T (G_i - B_i) s_i + o(\|s_i\|^2)}{2Q_i(s_i)}.$$

Podle (546) však platí

$$\left|\frac{s_i^T (G_i - B_i) s_i + o(\|s_i\|^2)}{2Q_i(s_i)}\right| \leq \frac{1}{\underline{C}} \frac{\|\vartheta_i\|\|s_i\|^2 + o(\|s_i\|^2)}{\|s_i\|^2} \to 0,$$

neboť $\|\vartheta_i\| \to 0$. Platí tedy $\rho_i(s_i) \to 1$ a jelikož $\underline{\rho} < 1$, existuje index $k_2 \geq k_1$ takový, že $\rho_i(s_i) \geq \underline{\rho}$ $\forall i \geq k_2$.

(c) Ukážeme, že existuje index $k \geq k_2$ takový, že $i \in N_1$ $\forall i \geq k$. Poznamenejme nejprve, že množina $N_1 \subset N$ je nekonečná. Kdyby tomu tak nebylo, existoval by index $k \geq k_2$ takový, že $i \notin N_1$ $\forall i \geq k$. Muselo by tedy platit $\|s_i\| \geq \underline{\delta}\Delta_i \geq \underline{\delta}\Delta_k$ $\forall i \geq k$, neboť z (b) plyne, že $i \in N_2$ $\forall i \geq k \geq k_2$. To je však spor, neboť podle (545) $\|g_i\| \to 0$ implikuje $\|s_i\| \to 0$. Omezme se nyní pouze na indexy $i \geq k_2$, $i \in N_1$, a označme $\omega_i = \omega_i(s_i)$. Podle (544) platí $\|\omega_i\| \overset{N_1}{\to} 0$ a $\|\vartheta_i\| \overset{N_1}{\to} 0$, takže stejným způsobem jako v důkazu věty 20 se dá ukázat, že existuje index $k_3 \geq k_2$, $k_3 \in N_1$, takový, že pro $i \geq k_3$, $i \in N_1$ platí (62). Použijeme-li větu 5, můžeme pro $i \geq k_3$ psát

$$g_{i+1} = g(x_i + s_i) = g_i + G_i s_i + o(\|s_i\|),$$

neboť podle (b) $i \in N_2$ pokud $i \geq k_3 \geq k_2$. Označme

$$\lambda_i = \frac{g_{i+1} - g_i - B_i s_i}{\|g_i\|} = -\frac{\vartheta_i \|s_i\| + o(\|s_i\|)}{\|g_i\|}$$

(druhá rovnost platí pro $i \geq k_3$). Pak z (62) pro $i \geq k_3$, $i \in N_1$, plyne, že $\|\lambda_i\| \leq \|\vartheta_i\| + o(1) \overset{N_1}{\to} 0$. Jelikož zároveň $\|\omega_i\| \leq \overline{\omega}_i \overset{N_1}{\to} 0$, existuje index $k \geq k_3$, $k \in N_1$ takový, že $\|\lambda_i\| < (\underline{G}/\overline{G})/(2\overline{\delta})$ a $\|\omega_i\| < (\underline{G}/\overline{G})/(2\overline{\delta})$, pokud $i \geq k$, $i \in N_1$. Pak pro $i \geq k$, $i \in N_1$ dostaneme

$$\begin{aligned} \|s_{i+1}\| &\leq \frac{1}{\underline{G}}\|g_{i+1}\| \leq \frac{1}{\underline{G}}(\|g_{i+1} - g_i - B_i s_i\| + \|B_i s_i + g_i\|) \leq \\ &\leq \frac{\overline{G}}{\underline{G}}(\|\lambda_i\| + \|\omega_i\|)\|s_i\| < \left(\frac{1}{2\overline{\delta}} + \frac{1}{2\overline{\delta}}\right)\|s_i\| = \frac{1}{\overline{\delta}}\|s_i\|. \end{aligned}$$

Jelikož podle (b) $i \in N_2$, pokud $i \geq k \geq k_2$, platí $\Delta_{i+1} \geq \Delta_i$, což dává

$$\|s_{i+1}\| < \|s_i\|/\overline{\delta} \leq \Delta_i \leq \Delta_{i+1},$$

takže $i+1 \in N_1$. Pokračujeme-li takto dále, dostaneme $i \in N_1$ $\forall i \geq k$.

(d) Platí

$$\frac{\|g_{i+1}\|}{\|g_i\|} \leq \frac{\|g_{i+1} - g_i - B_i s_i\| + \|B_i s_i + g_i\|}{\|g_i\|} \leq \|\lambda_i\| + \|\omega_i\|,$$

což spolu s $\|\lambda_i\| \to 0$ a $\|\omega_i\| \to 0$ dává

$$\lim_{i\to\infty} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \leq \frac{\overline{G}}{\underline{G}} \lim_{i\to\infty} \frac{\|g_{i+1}\|}{\|g_i\|} = 0.$$

□

Poznámka 205. Ve větě 122 předpokládáme, že $B_i \leq \overline{B}$ $\forall i \in N$. Je to nutné proto, že z (544) neplyne $\|B_i - G_i\| \to 0$, takže horní mez pro $\|B_i\|$ nelze odvodit z horní meze pro G_i. V případě Newtonovy metody, kdy $B_i = G_i$, nebo diferenční verze Newtonopvy metody, kdy $\|B_i - G_i\| \leq \overline{\vartheta}$, můžeme tento předpoklad vypustit.

Poznámka 206. Pokud neplatí (544), ale veličiny ϑ_i a ω_i, $i \in N$, jsou velmi malé, lze nalézt stejný odhad asymptotické rychlosti konvergence jako ve větě 22. Nechť $\|\omega_i\| \leq \overline{\omega}$ a $\|\vartheta_i\| \leq \overline{\vartheta}$ $\forall i \in N$. Z důkazu věty 122 plyne, že jsou-li hodnoty $\overline{\omega}$ a $\overline{\vartheta}$ dostatečně malé (příslušné nerovnosti nebudeme odvozovat), existuje index $k \in N$ takový, že $i \in N_1 \cap N_2$ pro $i \geq k$. Pak lze s použitím (70) psát

$$\begin{aligned} \lim_{i\to\infty} \frac{\|g_{i+1}\|}{\|g_i\|} &\leq \lim_{i\to\infty} \left(\frac{\|g(x_i + s_i) - g_i - G_i s_i\|}{\|g_i\|} + \frac{\|(B_i - G_i)s_i\|}{\|g_i\|} + \frac{\|B_i s_i + g_i\|}{\|g_i\|} \right) \\ &\leq \lim_{i\to\infty} \left(\frac{\|(B_i - G_i)s_i\|}{\|s_i\|}\frac{\|s_i\|}{\|g_i\|} + \frac{\|B_i s_i + g_i\|}{\|g_i\|} \right) \leq \overline{\vartheta}\frac{1+\overline{\omega}}{\underline{G} - \overline{\vartheta}} + \overline{\omega} = \frac{\overline{\omega}\,\overline{G} + \overline{\vartheta}}{\underline{G} - \overline{\vartheta}}, \end{aligned}$$

a stejným způsobem jako v části (d) důkazu věty 22 odvodíme, že posloupnost x_i, $i \in N$, konverguje lineárně k bodu $x^* \in R^n$ s asymptotickou rychlostí alespoň $(\overline{\omega}\,\underline{G} + \overline{\vartheta})/(\underline{G} - \overline{\vartheta})$.

5.2 Metody s optimálním lokálně omezeným krokem

Definice 39. *Metody s optimálním lokálně omezeným krokem používají směrový vektor*

$$s_i^* = \arg \min_{\|s\| \leq \Delta_i} Q_i(s), \tag{547}$$

přičemž $\|s_i^*\| = \Delta_i$, pokud toto minimum není jediné.

Vektor s_i^* určený podle (547) je nejlepším možným lokálně omezeným krokem, neboť je globálním minimem kvadratické funkce $Q_i(s)$ v oblasti určené nerovností $\|s\| \leq \Delta_i$. Abychom ukázali vlastnosti tohoto řešení, budeme se nejprve zabývat řešením jednorozměrné úlohy

$$s_i(\alpha^*) = \arg \min_{\|s_i(\alpha)\| \leq \Delta_i} Q_i(s_i(\alpha)), \tag{548}$$

kde $s_i(\alpha) = -\alpha g_i$.

Lemma 59. *Směrový vektor $s_i(\alpha^*) \in R^n$ určený podle (548), který lze vyjádřit ve tvaru*

$$s_i(\alpha^*) = -\frac{g_i^T g_i}{g_i^T B_i g_i} g_i, \quad g_i^T B_i g_i \geq \frac{\|g_i\|^3}{\Delta_i}, \tag{549}$$

$$s_i(\alpha^*) = -\frac{\Delta_i}{\|g_i\|} g_i, \qquad g_i^T B_i g_i < \frac{\|g_i\|^3}{\Delta_i}, \tag{550}$$

vyhovuje podmínce (T1c) s $\underline{\nu} = 1$.

Důkaz (a) Pokud $g_i^T B_i g_i \geq \|g_i\|^3/\Delta_i$, je funkce $Q_i(s(\alpha)) = (1/2)\alpha^2 g_i^T B_i g_i - \alpha g_i^T g_i$ ryze konvexní, nabývá svého minima pro $\alpha^* = g_i^T g_i / g_i^T B_i g_i$ a platí

$$\|s_i(\alpha^*)\| = \frac{\|g_i\|^3}{g_i^T B_i g_i} \leq \Delta_i,$$

takže vektor $s_i(\alpha^*)$ je řešením úlohy (548). Navíc platí

$$-Q_i(s_i(\alpha^*)) = \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} - \frac{1}{2}\frac{(g_i^T g_i)^2 g_i^T B_i g_i}{(g_i^T B_i g_i)^2} = \frac{1}{2}\frac{(g_i^T g_i)^2}{g_i^T B_i g_i} \geq \frac{1}{2}\frac{\|g_i\|^2}{\|B_i\|}.$$

(b) Pokud $g_i^T B_i g_i < \|g_i\|^3/\Delta_i$, je $Q_i'(s(\alpha)) = \alpha g_i^T B_i g_i - g_i^T g_i < 0$ pro $\alpha \leq \alpha^* = \Delta_i/\|g_i\|$, neboť buď $g_i^T B_i g_i \leq 0$ nebo $Q_i'(s(\alpha)) \leq Q_i'(s_i(\alpha^*)) = (\Delta_i/\|g_i\|) g_i^T B_i g_i - g_i^T g_i < 0$. Jelikož $\|s_i(\alpha^*)\| = \Delta_i$, je vektor $s_i(\alpha^*)$ řešením úlohy (548) a platí

$$-Q_i(s_i(\alpha^*) = \Delta_i\|g_i\| - \frac{1}{2}\frac{\Delta_i^2}{\|g_i\|^2} g_i^T B_i g_i > \Delta_i\|g_i\| - \frac{1}{2}\Delta_i\|g_i\| = \frac{1}{2}\Delta_i\|g_i\|.$$

□

Poznámka 207. Lemma 59 zdůvodňuje volbu podmínky (T1c), neboť ukazuje, že lze najít vektor s_i, který této podmínce vyhovuje. Vektor $s_i(\alpha^*)$, který je řešením úlohy (548), se nazývá Cauchyovým krokem. Podmínku (T1c) lze nahradit podmínkou $Q_i(s_i) \leq \underline{\nu} Q_i(s_i(\alpha^*))$. Poznamenejme, že vektor $s_i(\alpha^*)$ nesplňuje podmínku (T1b), takže ho nelze použít v metodách s lokálně omezeným krokem. Je však možné ho využít způsobem, který je popsán v oddílu 6.2.

Věta 123. *Směrový vektor $s_i^* \in R^n$ určený podle (547) vyhovuje podmínkám (T1) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\nu} = 1$.*

Důkaz (a) Podmínka (T1a) je přímo součástí podmínky (547). Předpokládejme, že $s_i^* \in R^n$ je řešením úlohy (547), přičemž $\|s_i^*\| < \Delta_i$. Pak nutně $Q_i(s)$ je ryze konvexní funkce a $B_i s_i^* + g_i = 0$, takže $\omega_i(s_i^*) = 0$ a

$$-Q_i(s_i^*) = g_i^T B_i^{-1} g_i - \frac{1}{2} g_i^T B_i^{-1} g_i = \frac{1}{2} g_i^T B_i^{-1} g_i \geq \frac{1}{2}\frac{\|g_i\|^2}{\|B_i\|}.$$

(b) Nechť $\|s_i^*\| = \Delta_i$. Podle (547) musí být $Q_i(s_i^*) \leq Q_i(s_i(\alpha^*))$, takže nutně

$$-Q_i(s_i^*) \geq -Q_i(s_i(\alpha^*)) \geq \frac{1}{2}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right).$$

□

Nyní uvedeme důležitou větu, která charakterizuje řešení úlohy (547).

Věta 124. *Vektor $s_i^* \in R^n$ je řešením úlohy (547) právě tehdy, když $\|s_i^*\| \leq \Delta_i$ a když existuje číslo $\lambda_i^* \geq 0$ takové, že matice $B_i + \lambda_i^* I$ je pozitivně semidefinitní a platí $(B_i + \lambda_i^* I)s_i^* + g_i = 0$ a $(\|s_i^*\| - \Delta_i)\lambda_i^* = 0$.*

Důkaz (a) Nejprve dokážeme nutnost. Jestliže $\|s_i^*\| < \Delta_i$, pak nutně $B_i s_i^* + g_i = 0$ a $(\|s_i^*\| - \Delta_i) \neq 0$ a funkce $Q_i(s)$ je konvexní, takže matice B_i je pozitivně semidefinitní. Jsou tedy splněny dokazované podmínky s $\lambda_i^* = 0$. Jestliže $\|s_i^*\| = \Delta_i$ musí být splněny Karushovy-Kuhnovy-Tuckerovy podmínky $(B_i + \lambda_i^* I)s_i^* + g_i = 0$ a $(\|s_i^*\| - \Delta_i)\lambda_i^* = 0$, kde $\lambda_i^* \geq 0$ (tvrzení 4). Zbývá dokázat pozitivní semidefinitnost matice $B_i + \lambda_i^* I$. Pro libovolný vektor $s \in R^n$ takový, že $\|s\| = \Delta_i$, platí

$$\begin{aligned} Q_i(s) - Q_i(s_i^*) &= g_i^T(s - s_i^*) + \frac{1}{2}s^T B_i s - \frac{1}{2}(s_i^*)^T B_i s_i^* \\ &= (s_i^*)^T(B_i + \lambda_i^* I)(s_i^* - s) + \frac{1}{2}s^T B_i s - \frac{1}{2}(s_i^*)^T B_i s_i^* \\ &= \frac{1}{2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) + \frac{1}{2}\lambda_i^*((s_i^*)^T s_i^* - s^T s) \\ &= \frac{1}{2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) \geq 0. \end{aligned}$$

Jelikož oba vektory s a s_i^* leží na kouli o poloměru Δ_i, může se vektor $v = \pm(s - s_i^*)/\|s - s_i^*\|$, kde $s \neq s_i^*$, rovnat libovolnému vektoru na jednotkové kouli, s výjimkou vektorů kolmých k s_i^*, a platí pro něj $v^T(B_i + \lambda_i^* I)v \geq 0$. Nechť $v \in R^n$, $\|v\| = 1$ a $v^T s_i^* = 0$. Pak existuje posloupnost $v_i \in R^n$, $\|v_i\| = 1$, $v_i^T s_i^* \neq 0$, $i \in N$ taková, že $v_i \to v$, takže $v^T(B_i + \lambda_i^* I)v = \lim_{i\to\infty} v_i^T(B_i + \lambda_i^* I)v_i \geq 0$. Platí tedy $v^T(B_i + \lambda_i^* I)v \geq 0 \; \forall v \in R^n$, takže matice $B_i + \lambda_i^* I$ je pozitivně semidefinitní.

(b) Nyní dokážeme postačitelnost. Jestliže $\|s_i^*\| < \Delta_i$, je funkce $Q_i(s)$ konvexní (matice $B_i + \lambda_i^* I$ je pro $\lambda_i^* = 0$ pozitivně semidefinitní), takže nutné podmínky jsou zároveň postačujícími podmínkami. Jestliže $\|s_i^*\| = \Delta_i$, pak dokazované podmínky implikují (tak jako v (a)), že

$$\begin{aligned} Q_i(s) - Q_i(s_i^*) &= g_i^T(s - s_i^*) + \frac{1}{2}s^T B_i s - \frac{1}{2}(s_i^*)^T B_i s_i^* \\ &= \frac{1}{2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) + \frac{1}{2}\lambda_i^*((s_i^*)^T s_i^* - s^T s) \geq \\ &\geq \frac{1}{2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) \geq 0 \end{aligned}$$

pro všechny vektory $s \in R^n$ takové, že $\|s\| \leq \|s_i\| = \Delta_i$. □

Některé dobré vlastnosti metod s optimálním lokálně omezeným krokem zůstanou zachovány i když řešíme úlohu (547) pouze přibližně. Proto zavádíme pojem kvazioptimálních metod s lokálně omezeným krokem.

Definice 40. *Metody s kvazioptimálním lokálně omezeným krokem používají místo podmínky (T1c) podmínku*

$$Q_i(s_i) \leq \underline{\nu} Q_i(s_i^*) \tag{551}$$

s $0 < \underline{\nu} \leq 1$, kde vektor s_i^* je řešením úlohy (547).

Poznámka 208. Podle definice 40 a věty 123 splňuje metoda s kvazioptimálním lokálně omezeným krokem podmínku (T1c) (s konstantou $\underline{\nu}$ vystupující v (551)).

5.3 Newtonova metoda s lokálně omezeným krokem

Newtonova metoda používá matice $B_i = G(x_i)$, $i \in N$, takže je-li splněn předpoklad F4, můžeme psát $\|B_i\| = \|G(x_i)\| \leq \overline{G}$, $i \in N$.

Věta 125. *Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F4. Pak Newtonova metoda realizovaná jako metoda s lokálně omezeným krokem je globálně konvergentní. Je-li navíc splněn předpoklad F5 a platí-li $x_i \to x^*$ a $\omega_i(s_i) \to 0$, je rychlost konvergence Q-superlineární.*

Důkaz Globální konvergence plyne bezprostředně z věty 118 (platí $\|B_i\| \leq \overline{G}$, $i \in N$). Superlineární konvergence plyne z toho, že $B_i = G_i$, takže

$$\frac{\|(B_i - G_i)s_i\|}{\|s_i\|} \leq \|B_i - G_i\| = 0,$$

což spolu s $\omega_i(s_i) \to 0$ implikuje Q-superlineární konvergenci (věta 122). □

Nejpoužívanější jsou tyto realizace Newtonovy metody.

(a) Nepřesná Newtonova metoda (kdy $\omega_i(s_i) > 0$). Jestliže platí (F4)–(F5) a $\omega_i(s_i) \to 0$, je tato realizace Q-superlineárně konvergentní. Soustava $B_i s_i + g_i = 0$ se řeší nepřesně metodou sdružených gradientů, což je výhodné zejména pro rozsáhlé řídké úlohy, neboť je obvykle zapotřebí méně než $O(n^3)$ operací na iteraci.

(b) Newtonova metoda s kvazioptimálním lokálně omezeným krokem. Pro tuto realizaci platí obzvláště silné tvrzení.

Věta 126. *Nechť x_i, $i \in N$, je posloupnost určená Newtonovou metodou s kvazioptimálním lokálně omezeným krokem. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1, F2 a F4. Pak existuje hromadný bod $x^* \in R^n$ posloupnosti x_i, $i \in N$, takový, že $g(x^*) = 0$ a $G(x^*) \succeq 0$. Jestliže bod $x^* \in R^n$ vyhovuje postačujícím podmínkám pro extrém ($g(x^*) = 0$ a $G(x^*) \succ 0$), je $x^* \in R^n$ jediným hromadným bodem posloupnosti x_i, $i \in N$, a posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.*

Důkaz (a) Nejprve dokážeme existenci hromadného bodu posloupnosti x_i, $i \in N$, splňujícího nutné podmínky pro extrém. Mohou nastat dva případy. Buď

$$\liminf_{i\to\infty} \Delta_i = 0$$

nebo

$$\liminf_{i\to\infty} \Delta_i > 0.$$

V prvním případě existuje podposloupnost x_i, $i \in M \subset N$, taková, že

$$\Delta_i \to 0 \quad \text{a} \quad i \notin N_2 \quad \forall i \in M \tag{552}$$

(proto nelze akceptovat podmínku (537)). Ve druhém případě existuje podposloupnost x_i, $i \in M \subset N$, taková, že

$$\Delta_i \geq \underline{\Delta} \quad \text{a} \quad i \in N_2 \quad \forall i \in M, \tag{553}$$

kde $\underline{\Delta} > 0$ (jelikož se tyto případy vylučují, budeme v obou případech používat stejnou indexovou množinu M). Vzhledem k tomu, že platí (F2), lze podposloupnost x_i, $i \in M$, vybrat tak, že $x_i \overset{M}{\to} x^*$ (existuje jediný hromadný bod posloupnosti x_i, $i \in M$). Z předpokladu $F \in C^2$ plyne, že $g_i \overset{M}{\to} g^* = g(x^*)$ a $G_i \overset{M}{\to} G^* = G(x^*)$.

(b) Předpokládejme, že platí (552) a $g^* \neq 0$. Pak existuje index $k_1 \in M$ takový, že $\|g_i\| \geq \|g^*\|/2$, pokud $i \in M$, $i \geq k_1$. Jelikož $\Delta_i \overset{M}{\to} 0$, existuje index $k_2 \in M$ takový, že $\Delta_i \leq \|g^*\|/(2\overline{G})$, pokud $i \in M$, $i \geq k_2$. Nechť $k = \max(k_1, k_2)$. Pak podle (T1a) a (T1c) platí

$$|Q_i(s_i)| \geq \frac{\nu}{2}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|G_i\|}\right) \geq \frac{\nu}{4}\|g^*\|\Delta_i \geq \frac{\nu\|g^*\|}{4\bar{\delta}}\|s_i\| \quad \forall i \in M,\ i \geq k,$$

což s použitím věty 5 dává

$$|\rho_i(s_i) - 1| = \left|\frac{F(x_i + s_i) - F(x_i)}{Q_i(s_i)} - 1\right| = \frac{o(\|s_i\|^2)}{|Q_i(s_i)|} = o(\|s_i\|) \to 0.$$

To je však ve sporu s předpokladem, že $i \notin N_2$.

(c) Předpokládejme, že platí (552) a matice G^* není pozitivně semidefinitní. Pak existuje index $k \in M$ takový, že $\underline{\lambda}_i \leq \underline{\lambda}^*/2 < 0$, pokud $i \in M$, $i \geq k$ (zde $\underline{\lambda}_i = \underline{\lambda}(G_i)$ a $\underline{\lambda}^* = \underline{\lambda}(G^*)$ jsou nejmenší vlastní čísla uvedených matic). Nechť v_i je vlastní vektor matice G_i příslušný vlastnímu číslu $\underline{\lambda}_i$ takový, že $v_i^T g_i \leq 0$ a $\|v_i\| = \Delta_i$. Pak podle (551) a (547) platí

$$|Q_i(s_i)| \geq \underline{\nu}|Q_i(s_i^*)| \geq \underline{\nu}|Q_i(v_i)| = -\underline{\nu}(v_i^T g_i + \frac{1}{2} v_i^T G_i v_i) \geq -\frac{\underline{\nu}}{2}\underline{\lambda}_i \Delta_i^2 \geq \frac{\underline{\nu}}{4\bar{\delta}^2}|\underline{\lambda}^*|\|s_i\|^2 \quad \forall i \in M,\ i \geq k,$$

takže podobně jako v části (b) dostaneme

$$|\rho_i(s_i) - 1| = \frac{o(\|s_i\|^2)}{|Q_i(s_i)|} = o(1) \to 0,$$

což odporuje předpokladu, že $i \notin N_2$.

(d) Předpokládejme, že platí (553). Použijeme-li (F1), dostaneme

$$F(x_1) - \underline{F} \geq \sum_{i=1}^{\infty}(F(x_i) - F(x_{i+1})) \geq \sum_{i \in M}(F(x_i) - F(x_i + s_i)),$$

takže $F(x_i) - F(x_i + s_i) \overset{M}{\to} 0$ a jelikož $M \subset N_2$, také $Q_i(s_i) \overset{M}{\to} 0$. Nechť

$$s^* = \arg \min_{\|s\| \leq \underline{\Delta}/2} Q^*(s), \tag{554}$$

kde

$$Q^*(s) = s^T g(x^*) + \frac{1}{2} s^T G(x^*) s.$$

Jelikož $x_i \overset{M}{\to} x^*$, existuje index $k \in M$ takový, že $\|x_i - x^*\| \leq \underline{\Delta}/2$, pokud $i \in M$, $i \geq k$. Platí tedy $\|x^* + s^* - x_i\| \leq \|x_i - x^*\| + \|s^*\| \leq \underline{\Delta}$, takže

$$Q_i(s_i) \leq \underline{\nu} Q_i(s_i^*) \leq \underline{\nu} Q_i(x^* + s^* - x_i) \quad \forall i \in M,\ i \geq k.$$

Jelikož $x_i \overset{M}{\to} x^*$, $g_i \overset{M}{\to} g^*$ a $G_i \overset{M}{\to} G^*$, platí $Q_i(x^* + s^* - x_i) \overset{M}{\to} Q^*(s^*)$, což spolu s $Q_i(s_i) \overset{M}{\to} 0$ a předchozí nerovností dává $Q^*(s^*) = 0$ (připomeňme, že všechny výrazy v teto nerovnosti jsou nekladne). Vektor $s^* = 0$ je tedy řešením úlohy (554), což je možné pouze tehdy, pokud $g(x^*) = 0$ a $G(x^*) \succeq 0$.

(e) Nechť $g(x^*) = 0$ a $G(x^*) \succ 0$. Pak podle poznámky 9 a předpokladu F4* existuje konstanta $\underline{G}$ a číslo ε, tak, že $v^T G(x) v \geq \underline{G}\|v\|^2$, kdykoliv $x \in \mathcal{B}(x^*, \varepsilon)$ (můžeme volit $\underline{G} = \underline{\lambda}^*/2$, kde $\underline{\lambda}^* > 0$ je nejmenší vlastní číslo matice $G(x^*)$). Nechť x_i, $i \in M$, je posloupnost definovaná v části (a) (buď (552) nebo (553)). Jelikož $x_i \to x^*$, musí od určitého indexu platit $x_i \in \mathcal{B}(x^*, \varepsilon)$. Abychom formálně zjednodušili některé úvahy, budeme bez újmy na obecnosti předpokládat, že to platí již od prvního indexu, čili že pro $i \in M$ je splněn předpoklad F4. Podle (F4) platí $s_i^T G_i s_i \geq \underline{G}\|s_i\|^2$ $\forall i \in M$, což dává

$$0 \geq Q_i(s_i) = s_i^T g_i + \frac{1}{2} s_i^T G_i s_i \geq -\|s_i\|\|g_i\| + \frac{1}{2}\underline{G}\|s_i\|^2,$$

takže $\|g_i\| \geq (\underline{G}/2)\|s_i\|$ $\forall i \in M$, a po dosazení do (T1c) dostaneme

$$|Q_i(s_i)| \geq \frac{\underline{\nu}\,\underline{G}^2}{8\bar{\delta}\,\overline{\overline{G}}}\|s_i\|^2.$$

Stejně jako v části (c) tedy platí

$$|\rho_i(s_i) - 1| = \frac{o(\|s_i\|^2)}{|Q_i(s_i)|} = o(1) \to 0,$$

takže existuje index $k_1 \in M$ takový, že $i \in N_2$, pokud $i \in M$, $i \geq k_1$. Tím jsme eliminovali případ (552).

(f) Nechť $g(x^*) = 0$ a $G(x^*) \succ 0$ a nechť platí (553). Jelikož $\|g_i\| \geq (\underline{G}/2)\|s_i\|$ a $\|g_i\| \to 0$, platí $\|s_i\| \to 0$. Existuje tedy index $k_2 \in M$ takový, že $\|s_i\| < \min(\varepsilon/2, \underline{\delta}\,\underline{\Delta})$, pokud $i \in M$, $i \geq k_2$. Pro $i \in M$, $i \geq \max(k_1, k_2)$, tedy platí $i \in N_1 \cap N_2$ a použijeme-li větu 5 a (T1b) s $\overline{\omega} = 0$, můžeme psát $g_{i+1} = g_i + G_i s_i + o(1)\|s_i\| = o(1)\|s_i\|$. Jelikož $o(1) \to 0$, existuje index $k \geq \max(k_1, k_2)$, takový, že

$$\|g_{i+1}\| < \frac{\underline{G}^2}{2\overline{G}}\|s_i\|,$$

pokud $i \in M$, $i \geq k$. Pro $i \in M$, $i \geq k$ tedy platí

$$\|s_{i+1}\| \leq \frac{2}{\underline{G}}\|g_{i+1}\| < \frac{\underline{G}}{\overline{G}}\|s_i\| \leq \|s_i\|$$

a

$$\|e_{i+1}\| \leq \frac{1}{\underline{G}}\|g_{i+1}\| < \frac{\underline{G}}{2\overline{G}}\|s_i\| \leq \frac{1}{\overline{G}}\|g_i\| \leq \|e_i\|,$$

(používáme vztahy (44) a (45) z důkazu věty 14) takže z $x_i \overset{M}{\to} x^*$ plyne $x_{i+1} \overset{M}{\to} x^*$ a přidáme-li $i+1$ do M, platí opět (553). Takto lze postupovat indukcí, čili lze předpokládat, že pro libovolný index $i \in N$, $i \geq k$ platí $i \in M$. Vektor $x^* \in R^n$ je tedy jediným hromadným bodem posloupnosti x_i, $i \in N$.

(g) Superlineární konvergence plyne ze vztahu

$$\|e_{i+1}\| \leq \frac{1}{\underline{G}}\|g_{i+1}\| = o(1)\|s_i\| = o(1)\|g_i\| = o(1)\|e_i\|,$$

který jsme poněkud podrobněji použili v části (f). □

Přestože Newtonova metoda, realizovaná jako metoda s optimálním lokálně omezeným krokem, má vynikající konvergenční vlastnosti, nelze ji doporučit pro řešení úloh s hustými Hessovými maticemi, kdy je zapotřebí příliš mnoho operací pro výpočet druhých derivací a pro opakované řešení soustavy lineárních rovnic. Newtonova metoda však velmi vhodná pro řešení rozsáhlých úloh s řídkými Hessovými maticemi jak je ukázáno v kapitole 10.

Jednou z nevýhod Newtonovy metody je nutnost použití druhých derivací minimalizované funkce. Druhé derivace se buď zadávají analyticky, nebo se určují pomocí automatického či numerického derivování. Při použití numerického derivování je Hessova matice určena nepřesně. Pro $i \in N$ neplatí $B_i = G_i$, ale pouze $\|B_i - G_i\| \leq \overline{\vartheta}$, kde horní odhad pro chybu $\overline{\vartheta} \geq 0$ udává následující věta.

Věta 127. *Nechť je splněn předpoklad F6 a nechť*

$$Be_j = \frac{g(x + \delta e_j) - g(x)}{\delta} \tag{555}$$

pro $1 \leq j \leq n$, kde e_j, $1 \leq j \leq n$, jsou sloupce jednotkové matice řádu n. Pak platí

$$\|B - G(x)\| \leq \frac{1}{2}\overline{L}\sqrt{n}\delta. \tag{556}$$

Důkaz Použijeme-li větu o střední hodnotě (tvrzení 3), dostaneme

$$g(x + \delta e_j) = g(x) + G(x)\delta e_j + \int_0^1 (G(x + \tau\delta e_j) - G(x))\delta e_j \mathrm{d}\tau,$$

takže podle F6 platí

$$\begin{aligned} \|(B - G(x))e_j\| &= \left\| \frac{g(x + \delta e_j) - g(x)}{\delta} - G(x)e_j \right\| \leq \frac{1}{\delta}\left\| \int_0^1 (G(x + \tau\delta e_j) - G(x))\delta e_j \mathrm{d}\tau \right\| \\ &\leq \frac{1}{2\delta}\overline{L}\delta^2\|e_j\|^2 = \frac{1}{2}\overline{L}\delta. \end{aligned}$$

Nechť $s \in R^n$ je libovolný vektor s jednotkovou normou. Pak lze psát

$$\begin{aligned}\|(B - G(x))s\| &= \left\| \sum_{j=1}^{n} (B - G(x)) e_j e_j^T s \right\| \leq \sum_{j=1}^{n} |e_j^T s| \|(B - G(x))e_j\| \leq \frac{1}{2}\overline{L}\delta \sum_{j=1}^{n} |e_j^T s| \\ &\leq \frac{1}{2}\overline{L}\sqrt{n}\delta\|s\| = \frac{1}{2}\overline{L}\sqrt{n}\delta\end{aligned}$$

a jelikož

$$\|B - G(x)\| = \max_{\|s\|=1} \|(B - G(x))s\|,$$

platí (556). □

Poznámka 209. Jsou-li splněny předpoklady F5, F6 a je-li matice B určena podle vzorce (555), kde

$$\delta < \frac{(1 - \overline{\omega})\underline{G}}{\overline{L}\sqrt{n}} \tag{557}$$

a $0 \leq \overline{\omega} < 1$, platí podle (556) $\|B - G(x)\| \leq \overline{\vartheta}$, kde

$$\overline{\vartheta} \leq \frac{1}{2}\overline{L}\sqrt{n}\delta < \frac{(1 - \overline{\omega})\underline{G}}{2}.$$

Navíc podle věty 21 $\|Bs + g\| \leq \overline{\omega}$ implikuje $\|Gs + g\| \leq \overline{\omega}'$, kde $\overline{\omega}' = (\overline{\omega}\,\underline{G} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$.

5.4 Nemonotonní metody s lokálně omezeným krokem

V některých případech, například při realizaci Newtonovy metody, je výhodné používat nemonotonní metody s lokálně omezeným krokem, kdy posloupnost F_i, $i \in N$, není nerostoucí. V definici nemonotonních metod s lokálně omezeným krokem budeme místo hodnot F_i, $i \in N$, používat čísla $\overline{F}_i \geq F_i$, $i \in N$, jejichž výběr je určen konkrétní metodou. Budeme předpokládat, že posloupnost $\overline{F}_i$, $i \in N$, je nerostoucí a $\overline{F}_1 = F_1$, takže $\overline{F}_i \leq \overline{F}$ $\forall i \in N$, kde $\overline{F}$ je číslo použité v předpokladu F2. Abychom dokázali globální konvergenci nemonotonních metod s lokálně omezeným krokem, budeme předpokládat, že tyto metody jsou striktní.

Definice 41. *Striktní nemonotonní metody s lokálně omezeným krokem se liší od striktních metod s lokálně omezeným krokem (definice 38) pouze tím, že podmínky (T2c), (T2d) nahradíme podmínkami*

$$\overline{\rho}_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \alpha_i = 0, \tag{T2e}$$

$$\overline{\rho}_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1, \tag{T2f}$$

kde

$$\overline{\rho}_i(s_i) = \frac{F(x_i + s) - \overline{F}_i}{Q_i(s)}$$

a $0 < \underline{\rho} < 1$ *(podmínky (T1) a (T3), ve kterých vystupuje podíl (518), zůstanou zachovány). Množinu indexů, pro které platí (T2f) označíme* $\overline{N}_2$.

Nejprve vyšetříme jednoduchou nemonotonní metodu s lokálně omezeným krokem, pro kterou platí

$$\overline{F}_i = \max\{F_j : i - \min(m, i) + 1 \leq j \leq i\}, \tag{558}$$

kde m je číslo udávající počet funkčních hodnot použitých k určení $\overline{F}_i$.

Věta 128. *(Globální konvergence metody (558)) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná striktní nemonotonní metodou s lokálně omezeným krokem (558) takovou, že $\|B_i\| \le \overline{B}$ $\forall i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak platí*

$$\liminf_{i\to\infty} \|g_i\| = 0. \tag{559}$$

Důkaz Jelikož používáme podmínky (T1) a (T3), můžeme použít lemma 57, podle kterého platí

$$\Delta_i \ge \frac{1}{\overline{\delta}}\|s_i\| \ge \frac{\underline{c}\, m_i}{\overline{\delta}\,\overline{B}}, \quad m_i = \min_{1\le j\le i} \|g_j\|,$$

kde $0 < \underline{c} < 1$.

(a) Předpokládejme, že není splněna podmínka (559). Pak musí existovat číslo $\underline{\varepsilon} > 0$ takové, že $m_i \ge \underline{\varepsilon}$ $\forall i \in N$, což spolu s předchozí nerovností dává

$$\Delta_i \ge \frac{\underline{c}\,\underline{\varepsilon}}{\overline{\delta}\,\overline{B}} > 0, \quad \forall i \in N. \tag{560}$$

Musí tedy existovat nekonečná podmnožina $N_2 \subset N$ indexů, pro které platí (T3b). Zřejmě $N_2 \subset \overline{N}_2$, neboť $\overline{\rho}_i(s_i) \ge \rho_i(s_i)$. Množina $\overline{N}_2$ je tedy také nekonečná a existuje její nekonečná podmnožina $N_4 = \{i_1, i_2, i_3, \dots\} \subset \overline{N}_2$ taková, že $i_{k+1} - i_k \ge m$ $\forall k \in N$.

(b) Ukážeme, že není-li splněna podmínka (559), platí

$$F_{i_k+j} \le \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}} \quad \forall j \in N, \tag{561}$$

kde $\underline{\varepsilon}$ je číslo použité v (560). Důkaz provedeme indukcí. Nechť $i_k \in N_4$ a $j \in N$. Budeme předpokládat, že buď $i_k + j - 1 \in \overline{N}_2$ (což je splněno například pro $j = 1$), nebo $i_k + j - 1 \notin \overline{N}_2$ a platí (561) s $j - 1$ místo j (indukční předpoklad). V prvním případě použitím (T1c), (T2f) a (560) dostaneme

$$\begin{aligned} F_{i_k+j} &\le \overline{F}_{i_k+j-1} + \underline{\rho} Q_{i_k+j-1} \le \overline{F}_{i_k+j-1} - \frac{\underline{\rho}\,\underline{\nu}}{2}\|g_{i_k+j-1}\| \min\left(\Delta_{i_k+j-1}, \frac{\|g_{i_k+j-1}\|}{\|B_{i_k+j-1}\|}\right) \\ &\le \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{2} \min\left(\frac{\underline{c}\,\underline{\varepsilon}}{\overline{\delta}\,\overline{B}}, \frac{\underline{\varepsilon}}{\overline{B}}\right) \le \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}}, \end{aligned}$$

neboť posloupnost $\overline{F}_i$, $i \in N$, je nerostoucí (je to ukázáno v důkazu věty 25). Ve druhém případě podle (T2e) a indukčního předpokladu platí

$$F_{i_k+j} = F_{i_k+j-1} \le \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}}.$$

Tím je indukční krok proveden.

(c) Není-li splněna podmínka (559), platí

$$\overline{F}_{i_{k+1}} \le \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}},$$

neboť $i_{k+1} - i_k \ge m$ a maximální hodnota $\overline{F}_{i_k}$ z množiny hodnot uvedených v (558) po m krocích vypadne (pro ty zbylé platí (561)). Můžeme tedy psát

$$\overline{F}_{i_1} - \underline{F} \ge \overline{F}_{i_1} - \lim_{k\to\infty} \overline{F}_{i_{k+1}} = \sum_{k=1}^{\infty} (\overline{F}_{i_k} - \overline{F}_{i_{k+1}}) \ge \sum_{k=1}^{\infty} \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}} = \infty,$$

což je spor, neboť výraz na levé straně této nerovnosti je podle předpokladu F1 konečný. □

Poznámka 210. Hodnotu (558) lze získat tak, že nejprve položíme $\mathcal{F}_1 = \{F_1\}$. Pro $i \in N$ vypočteme $\overline{F}_i = \max\{F_j : j \in \mathcal{F}_i\}$. Následně položíme $\tilde{\mathcal{F}}_{i+1} = \mathcal{F}_i \cup \{F_{i+1}\}$. Má-li $\tilde{\mathcal{F}}_{i+1}$ nanejvýš m prvků, položíme $\mathcal{F}_{i+1} = \tilde{\mathcal{F}}_{i+1}$. V opačném případě získáme $\mathcal{F}_{i+1}$ tak, že z $\tilde{\mathcal{F}}_{i+1}$ vyjmeme prvek s nejmenším indexem. Tento postup lze modifikovat tak aby se množina $\mathcal{F}_i$ měnila pouze po úspěšném kroku (kdy $i \in \overline{N}_2$). Opět položíme $\mathcal{F}_1 = \{F_1\}$. Pro $i \in N$ vypočteme $\overline{F}_i = \max\{F_j : j \in \mathcal{F}_i\}$. Následně položíme

$$\overline{\rho}_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \tilde{\mathcal{F}}_{i+1} = \mathcal{F}_i, \tag{562}$$

$$\overline{\rho}_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \tilde{\mathcal{F}}_{i+1} = \mathcal{F}_i \cup \{F_{i+1}\}. \tag{563}$$

Má-li $\tilde{\mathcal{F}}_{i+1}$ nanejvýš m prvků, položíme $\mathcal{F}_{i+1} = \tilde{\mathcal{F}}_{i+1}$. V opačném případě získáme $\mathcal{F}_{i+1}$ tak, že z $\tilde{\mathcal{F}}_{i+1}$ vyjmeme prvek s nejmenším indexem. Nemonotonní metoda s lokálně omezeným krokem (562)–(563) je také globálně konvergentní. Důkaz tohoto tvrzení je podobný důkazu věty 128.

Nyní vyšetříme nemonotonní metodu s lokálně omezeným krokem, kde se čísla $\overline{F}_i$, $i \in N$, určují rekurentně tak, že $\overline{n}_1 = 1$, $\overline{F}_1 = F_1$ a

$$\overline{\rho}_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \overline{n}_{i+1} = \overline{n}_i, \qquad \overline{F}_{i+1} = \overline{F}_i \tag{564}$$

$$\overline{\rho}_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \overline{n}_{i+1} = \lambda \overline{n}_i + 1, \quad \overline{F}_{i+1} = \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \tag{565}$$

pro $i \in N$, kde $0 \leq \lambda \leq 1$.

Věta 129. *(Globální konvergence metody (564)–(565)) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná nemonotonní metodou s lokálně omezeným krokem (564)–(565) taková, že $\|B_i\| \leq \overline{B}$ $\forall i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak platí (559).*

Důkaz Předpokládejme, že není splněna podmínka (559). Podobně jako v části (a) důkazu věty 128 z toho plyne, že množina $\overline{N}_2 = \{i_1, i_2, i_3, \dots\}$ je nekonečná a existuje číslo $\underline{\varepsilon} > 0$ takové, že platí $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$ a (560), a podobně jako v části (b) důkazu věty 128 dostaneme nerovnost

$$F_{i_{k+1}} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}}, \quad \forall k \in N,$$

která spolu s (T2e)–(T2f) a (564)–(565) dává

$$\overline{F}_{i_{k+1}} = \frac{\lambda \overline{n}_{i_k} \overline{F}_{i_k} + F_{i_{k+1}}}{\overline{n}_{i_{k+1}}} = \overline{F}_{i_k} + \frac{F_{i_{k+1}} - \overline{F}_{i_k}}{\overline{n}_{i_{k+1}}} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{2\overline{\delta}\,\overline{B}\overline{n}_{i_{k+1}}} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{4\overline{\delta}\,\overline{B}k},$$

neboť podle (565) a poznámky 50 platí $\overline{n}_{i_{k+1}} \leq k + 1 \leq 2k$. Můžeme tedy psát

$$\overline{F}_{i_1} - \underline{F} \geq \overline{F}_{i_1} - \lim_{k\to\infty} \overline{F}_{i_{k+1}} = \sum_{k=1}^{\infty} (\overline{F}_{i_k} - \overline{F}_{i_{k+1}}) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{c}\,\underline{\varepsilon}^2}{4\overline{\delta}\,\overline{B}} \sum_{k=1}^{\infty} \frac{1}{k} = \infty,$$

což je spor, neboť výraz na levé straně této nerovnosti je podle předpokladu F1 konečný. □

5.5 Kombinované metody s lokálně omezeným krokem

Metody s lokálně omezeným krokem se liší od metod spádových směrů určením směrového vektoru a výběrem délky kroku. Proto se nabízí dvě možnosti jak tyto metody kombinovat. První kombinovaná metoda používá směrový vektor $s_i = -\lambda_i H_i g_i$, kde $H_i = B_i^{-1}$ a

$$\lambda_i = 1, \qquad \|H_i g_i\| \leq \Delta_i, \tag{566}$$

$$\lambda_i = \frac{\Delta_i}{\|H_i g_i\|}, \quad \|H_i g_i\| > \Delta_i, \tag{567}$$

a délka kroku se vybírá podle (T2) a (T3). Je zřejmé, že tento směrový vektor splňuje podmímky (T1a) a (T1b). Podmínku (T1c) však splňovat nemusí. Ukážeme, že podmínka (T1c) je splněna, je-li matice H_i pozitivně definitní a existuje-li číslo $\overline{\kappa}$ takové, že $\kappa(H_i) \leq \overline{\kappa}$ $\forall i \in N$.

Lemma 60. *Nechť $s_i = -\lambda_i H_i g_i$, kde H_i je pozitivně definitní matice a $0 < \lambda_i \leq 1$ je číslo určené podle (566)–(567). Pak platí*

$$-Q_i(s_i) \geq \frac{1}{2}\|g_i\|\Delta_i \frac{g_i^T H_i g_i}{\|g_i\|\|H_i g_i\|} \geq \frac{1}{2\sqrt{\kappa(H_i)}}\|g_i\|\Delta_i \tag{568}$$

Jestliže $\kappa(H_i) \leq \overline{\kappa}$, je splněna podmínka (T1c) s $\underline{\nu} = 1/\sqrt{\overline{\kappa}}$.

Důkaz Jelikož $H_i = B_i^{-1}$, platí

$$Q_i(s_i) = \frac{1}{2}\lambda_i^2 g_i^T H_i g_i - \lambda_i g_i^T H_i g_i = \frac{1}{2}\lambda_i(\lambda_i - 2) g_i^T H_i g_i, \tag{569}$$

neboli

$$-Q_i(s_i) = \frac{1}{2}\lambda_i(2 - \lambda_i) g_i^T H_i g_i \geq \frac{1}{2}\lambda_i g_i^T H_i g_i \geq \frac{1}{2}\frac{\Delta_i}{\|H_i g_i\|} g_i^T H_i g_i,$$

což s použitím věty 10 dává (568). Zbytek tvrzení je zřejmý. □

Důsledek 19. *Uvažujme metodu s lokálně omezeným krokem, pro kterou platí $s_i = -\lambda_i H_i g_i$, (566)–(567) a (T2), (T3). Pak splňuje-li funkce $F : \mathcal{D} \to R$ předpoklady F1 a F3 a existují-li čísla $0 < \underline{H} \leq \overline{H}$ taková, že $\underline{H} \leq \underline{\lambda}(H_i) \leq \overline{\lambda}(H_i) \leq \overline{H}$ $\forall i \in N$, je tato metoda globálně konvergentní (platí (559)).*

Důkaz Existují-li čísla $0 < \underline{H} \leq \overline{H}$ taková, že $\underline{H} \leq \underline{\lambda}(H_i) \leq \overline{\lambda}(H_i) \leq \overline{H}$ $\forall i \in N$, platí $\kappa(H_i) \leq \overline{H}/\underline{H}$ $\forall i \in N$, takže je podle lemmatu 60 splněna podmínka (T1c). Jelikož také $\|B_i\| \leq 1/\underline{H}$ $\forall i \in N$, jsou splněny předpoklady věty 118, takže platí (559). □

Poznámka 211. Z toho co jsme zatím uvedli je patrné, že použitím vektoru $s_i = -\lambda_i H_i g_i$, který nemusí splňovat podmínku (T1c), ztrácíme hlavní výhodu metod s lokálně omezeným krokem (nezávislost na pozitivní definitnosti a podmíněnosti matic B_i, $i \in N$). Tuto úpravu však můžeme použít k realizaci metod s proměnnou metrikou, kdy matice H_i, $i \in N$, jsou pozitivně definitní a kdy je výhodné, že určení vektoru $s_i = -\lambda_i H_i g_i$ vyžaduje $O(n^2)$ operací, zatímco určení vektoru vyhovujícího podmínce (T1c), které je v jistém smyslu ekvivalentní řešení soustavy rovnic $B_i s_i + g_i = 0$, vyžaduje $O(n^3)$ operací. Poznamenejme, že hodnotu $Q_i(s_i)$ můžeme počítat podle (569) pomocí matice H_i.

Jiná kombinovaná metoda je založena na tom, že se směrový vektor určuje tak, aby byl spádový, a jednoduchý výběr délky kroku (T2c), (T2d) se nahradí složitější procedurou, vyžadující splnění podmínky

$$\rho_i(\alpha_i s_i) \geq \underline{\rho}. \tag{570}$$

Používá se přitom modifikovaný Armijův výběr délky kroku, kdy $\alpha_i > 0$ je první člen vyhovující podmínce (570) v posloupnosti α_i^j, $j \in N$, takové, že $\alpha_i^1 = 1$, a

$$\underline{\beta}\alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta}\alpha_i^j \quad \forall j \in N, \tag{571}$$

kde $0 < \underline{\beta} \leq \overline{\beta} < 1$ jsou konstanty nezávislé na indexu $i \in N$. Směrový vektor se určuje podle (T1) a (T3).

Definice 42. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je striktní kombinovanou metodou s lokálně omezeným krokem, jestliže směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že*

$$\|s_i\| \leq \overline{\delta}\Delta_i, \tag{T1a}$$

$$\|s_i\| < \underline{\delta}\Delta_i \quad \Rightarrow \quad \|\omega_i(s_i)\| \leq \overline{\omega}_i \leq \overline{\omega}, \tag{T1b}$$

$$-Q_i(s_i) \;\geq\; \frac{\underline{\nu}}{2}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right), \tag{T1c}$$

$$-g_i^T s_i \;\geq\; \underline{\nu}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right), \tag{T1d}$$

kde $0 < \underline{\delta} < 1 < \overline{\delta}$, $0 < \underline{\nu} < 1$ *a* $0 \leq \overline{\omega} < 1$, *délky kroku* $\alpha_i \geq 0$, $i \in N$, *se vybírají tak, že*

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \rho_i(\alpha_i s_i) \geq \underline{\rho}, \tag{572}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1, \tag{573}$$

kde $0 < \underline{\rho} < 1$, *a čísla* $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, *se volí tak, že*

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \max(\alpha_i, \underline{\beta})\|s_i\| \leq \Delta_{i+1} \leq \overline{\beta}\|s_i\|, \tag{574}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \Delta_i \leq \Delta_{i+1} \leq \min(\underline{\gamma}\Delta_i, \overline{\Delta}), \tag{575}$$

kde $0 < \underline{\beta} \leq \overline{\beta} < 1 < \underline{\gamma}$, *přičemž* $\overline{\beta}\,\overline{\delta} < 1$.

Poznámka 212. Striktní kombinovaná metoda s lokálně omezeným krokem se od striktní metody s lokálně omezeným krokem (definice 38) liší tím, že směrový vektor musí splňovat dodatečnou podmínku (T1d) (spádovost), že podmínka (T2c) je nahražena podmínkou (572) (Armijův výběr délky kroku) a že podmínka (T3a) je nahražena podmínkou (574), která se vztahuje k případu (572). Poznamenejme, že zvolíme-li číslo $\underline{\beta}$ dostatečně malé, platí $\Delta_{i+1} = \alpha_i\|s_i\| = \|x_{i+1} - x_i\|$ ve většině případů, kdy $\rho_i(s_i) < \underline{\rho}$.

Poznámka 213. Je-li splněn předpoklad F3, je definice 42 korektní (podmínky (572)–(575) lze splnit). Předně z $\|s_i\| > 0$, $\|g_i\| > 0$ a (T1d) plyne, že $g_i^T s_i < 0$, takže

$$\lim_{\alpha\to 0} \rho_i(\alpha s_i) \geq \lim_{\alpha\to 0} \frac{\alpha g_i^T s_i - \alpha^2 \overline{G}\|s_i\|^2}{\alpha g_i^T s_i + (\alpha^2/2) s_i^T B_i s_i} = \lim_{\alpha\to 0} \frac{g_i^T s_i - \alpha \overline{G}\|s_i\|^2}{g_i^T s_i + (\alpha/2) s_i^T B_i s_i} = 1,$$

a jelikož $\underline{\rho} < 1$, existuje číslo $\overline{\alpha}_i > 0$ takové, že $\rho_i(\alpha_i s_i) \geq \underline{\rho}$, pokud $0 < \alpha_i \leq \overline{\alpha}_i$. Dále z (571) plyne, že $\alpha_i \leq \overline{\beta}$, pokud $\rho_i(s_i) \leq \underline{\rho}$, takže lze splnit nerovnost v (574).

V dalších úvahách budeme používat indexové množiny N_1, N_2, které mají stejný význam jako v poznámce 199.

Věta 130. *Nechť* $x_i \in R^n$, $i \in N$, *je posloupnost generovaná kombinovanou metodou s lokálně omezeným krokem (definice 42) taková, že*

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty, \quad M_i = \max_{1\leq j\leq i} \|B_j\|.$$

Nechť funkce $F : \mathcal{D} \to R$ *splňuje předpoklady F1 a F3. Pak platí (559).*

Důkaz V případech, kdy $i \in N_1$, $i \notin N_2$, $i = 1$, můžeme postupovat stejně jako v částech (a), (b), (c) důkazu lemmatu 57. V případě, kdy $i \in N_2$, můžeme, tak jako v části (d) zmíněného důkazu, využít toho, že podle (574) platí $\Delta_{i+1} \geq \underline{\beta}\Delta_i$, pokud $i \notin N_2$. Platí tedy tvrzení analogické lemmatu 57, takže je splněna nerovnost (520). Jelikož podle (574) platí $\Delta_{i+1} \leq \overline{\beta}\|s_i\|$, pokud $i \notin N_2$, lze použít lemma 58. Zbytek důkazu je totožný s důkazem věty 118. □

Zbývá ukázat, jak lze nalézt směrový vektor vyhovující podmínkám (T1a)–(T1d). Metody, které budeme vyšetřovat splňují podmínky (T1a)–(T1c) (lemma 59, věta 123). Proto stačí ukázat, že platí i (T1d).

Lemma 61. *Směrový vektor* $s_i(\alpha^*) \in R^n$ *vyšetřovaný v lemmatu 59 vyhovuje podmínce (T1d) s* $\underline{\nu} = 1$.

Důkaz Pokud $g_i^T B_i g_i \geq \|g_i\|^3/\Delta_i$, platí

$$-g_i^T s_i(\alpha^*) = \frac{g_i^T g_i}{g_i^T B_i g_i} g_i^T g_i \geq \frac{\|g_i\|^2}{\|B_i\|}.$$

Pokud $g_i^T B_i g_i < \|g_i\|^3/\Delta_i$, platí

$$-g_i^T s_i(\alpha^*) = \frac{\Delta_i}{\|g_i\|} g_i^T g_i = \Delta_i \|g_i\|.$$

□

Věta 131. *Směrový vektor $s_i^* \in R^n$ určený řešením úlohy (547) vyhovuje podmínce (T1d) s $\underline{\nu} = 1/4$.*

Důkaz (a) Podle věty 124 platí $(B_i + \lambda_i^* I)s_i^* = -g_i$, kde buď $\|s_i^*\| < \Delta_i$ a $\lambda_i^* = 0$, nebo $\|s_i^*\| = \Delta_i$ a $\lambda_i^* \geq 0$. Pokud $\|s_i^*\| = \Delta_i$, můžeme psát

$$\|g_i\| \geq \underline{\lambda}(B_i + \lambda_i^* I)\Delta_i = (\underline{\lambda}(B_i) + \lambda_i^*)\Delta_i$$

neboť matice $B_i + \lambda_i^* I$ je pozitivně semidefinitní, takže její vlastní čísla jsou zároveň singulárními čísly. Dostaneme tedy odhad

$$0 \leq \lambda_i^* \leq \frac{\|g_i\|}{\Delta_i} - \underline{\lambda}(B_i) \leq \frac{\|g_i\|}{\Delta_i} + \|B_i\|,$$

který platí i v případě, že $\|s_i^*\| < \Delta_i$, kdy $\lambda_i^* = 0$.
(b) Je-li matice $B_i + \lambda_i^* I$ regulární, můžeme s použitím horní meze pro λ_i^* získané v (a) psát

$$\begin{aligned} -g_i^T s_i &= g_i^T (B_i + \lambda_i^* I)^{-1} g_i \geq \frac{\|g_i\|^2}{\overline{\lambda}(B_i + \lambda_i^* I)} \geq \frac{\|g_i\|^2}{\|B_i\| + \lambda_i^*} \geq \frac{\|g_i\|^2}{2\|B_i\| + \|g_i\|/\Delta_i} \\ &\geq \frac{\|g_i\|^2}{4\max(\|B_i\|, \|g_i\|/\Delta_i)} = \frac{1}{4}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right). \end{aligned}$$

Je-li matice $B_i + \lambda_i^* I$ singulární, využijeme toho, že $s_i = -(B_i + \lambda_i^* I)^\dagger g_i + v_i$, kde $g_i^T v_i = 0$ a $(B_i + \lambda_i^* I)^\dagger$ je pseudoinverzní matice, jejíž nenulová vlastní čísla jsou převrácenými hodnotami nenulových vlastních čísel matice $B_i + \lambda_i^* I$. Dostaneme tak stejné nerovnosti jako v regulárním případě. □

Další metody, které splňují podmínky (T1a)–(T1d), jsou popsány v oddílu 6 (věta 138, věta 139).

5.6 Metody kvadratické regularizace

Podle věty 124 je optimální lokálně omezený krok s_i^* řešením úlohy

$$s_i^* = \arg\min_{s \in R^n} Q_{\lambda_i^*}(s), \qquad Q_{\lambda_i^*}(s) = g_i^T s + \frac{1}{2} s^T B_i s + \frac{1}{2}\lambda_i^* \|s\|^2,$$

kde λ_i^* je Lagrangeův multiplikátor splňující podmínky věty 124. Člen $(1/2)\lambda_i^*\|s\|^2$ se nazývá kvadratickou regularizací. Z tohoto důvodu je možné definovat metody kvadratické regularizace, které jsou podobné metodám s lokálně omezeným krokem. Při jejich popisu budeme používat označení

$$Q_{\lambda_i}(s) = g_i^T s + \frac{1}{2} s^T B_i s + \frac{1}{2}\lambda_i \|s\|^2. \tag{576}$$

a budeme předpokládat, že $\lambda_i \geq \max(0, -\underline{\lambda}(B_i))$ (pak matice $B_i + \lambda_i I$ je pozitivně semidefinitní). Podíl skutečného a předpověděného poklesu funkce F je pak definován vztahem

$$\rho_i(s) = \frac{F(x_i + s) - F(x_i)}{Q_{\lambda_i}(s)}. \tag{577}$$

Definice 43. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou kvadratické regularizace, jestliže směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že*

$$Q_{\lambda_i}(s_i) \leq \underline{\nu} Q_{\lambda_i}(s_i(\alpha^*)), \qquad s_i(\alpha^*) = \arg\min_{s_i(\alpha) = -\alpha g_i} Q_{\lambda_i}(s_i(\alpha)), \tag{Q1}$$

kde $0 < \underline{\nu} \leq 1$, délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq 0 \quad \Rightarrow \quad \alpha_i = 0, \tag{Q2a}$$

$$\rho_i(s_i) > 0 \quad \Rightarrow \quad \alpha_i = 1, \tag{Q2b}$$

a čísla $\lambda_i \geq 0$, $i \in N$, se volí tak, že

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \underline{\gamma}\lambda_i \leq \lambda_{i+1} \leq \overline{\gamma}\lambda_i, \tag{Q3a}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad 0 < \lambda_{i+1} \leq \lambda_i, \tag{Q3b}$$

kde $0 < \underline{\rho} < 1$ a $1 < \underline{\gamma} < \overline{\gamma}$. Řekneme, že metoda kvadratické regularizace je striktní metodou kvadratické regularizace, jsou-li podmínky (Q2a) a (Q2b) nahraženy podmínkami

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \alpha_i = 0, \tag{Q2c}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1. \tag{Q2d}$$

Poznámka 214. Podmínka (Q3b) se obvykle realizuje tak, že

$$\underline{\rho} \leq \rho_i(s_i) \leq \overline{\rho} \quad \Rightarrow \quad \lambda_{i+1} = \lambda_i, \tag{Q3c}$$

$$\rho_i(s_i) > \overline{\rho} \quad \Rightarrow \quad \lambda_{i+1} = \min(\lambda_i, \max(\underline{\beta}\lambda_i, \|g_i\|)), \tag{Q3d}$$

kde $0 < \underline{\rho} < \overline{\rho} < 1$ a $0 < \beta < 1$.

Poznámka 215. Aby minimalizační úloha v (Q1) měla řešení, je třeba, aby platilo $\lambda_i \geq \max(0, -\underline{\lambda}(B_i))$. To lze algoritmicky zajistit tak, že při určování Choleského rozkladu matice $B_i + \lambda_i I$ adaptivně zvětšujeme hodnotu parametru λ_i. Nicméně, je-li matice B_i indefinitní, je účelnější používat metody s lokálně omezeným krokem. Metody kvadratické regularizace mají význam zejména tehdy, je-li matice B_i pozitivně semidefinitní, například při minimalizaci součtu čtverců [88], [117].

Poznámka 216. Při vyšetřování metod kvadratické regularizace se omezíme na případy, kdy je matice B_i pozitivně semidefinitní a budeme používat označení

$$\begin{aligned} N_1 &= \{i \in N : \rho_i(s_i) < \underline{\rho}\}, \\ N_2 &= \{i \in N : \rho_i(s_i) \geq \underline{\rho}\}, \\ N_3 &= \{i \in N : \rho_i(s_i) > \overline{\rho}\}. \end{aligned}$$

Jelikož $0 \leq \underline{\rho} < \overline{\rho}$, platí $N_3 \subset N_2$.

Lemma 62. *Nechť $\|g_i\| > 0$, $\|B_i\| > 0$, $B_i \succeq 0$ a $\lambda_i > 0$. Pak platí*

$$-Q_{\lambda_i}(s_i(\alpha^*)) \geq \frac{\|g_i\|^2}{2(\|B_i\| + \lambda_i)} \geq \frac{\|g_i\|^2}{4} \min\left(\frac{1}{\|B_i\|}, \frac{1}{\lambda_i}\right). \tag{578}$$

Důkaz Pro zjednodušení zápisu budeme index i vynechávat. Nechť $s(\alpha) = -\alpha g$, kde $\alpha > 0$. Pak můžeme psát

$$-Q_\lambda(s(\alpha)) = -g^T s(\alpha) - \frac{1}{2} s(\alpha)^T B s(\alpha) - \frac{1}{2}\lambda\|s(\alpha)\|^2 \geq \alpha\|g\|^2 \left(1 - \frac{1}{2}\alpha(\|B\| + \lambda)\right).$$

Označme $\tilde{\alpha}$ hodnotu, která maximalizuje výraz na pravé straně této nerovnosti. Tuto hodnotu získáme vynulováním první derivace, což dává

$$\|g\|^2 (1 - \tilde{\alpha}(\|B\| + \lambda)) = 0,$$

neboli

$$\tilde{\alpha} = \frac{1}{\|B\| + \lambda}.$$

Po dosazení dostaneme

$$-Q_\lambda(s(\alpha^*)) \geq -Q_\lambda(s(\tilde{\alpha})) \geq \tilde{\alpha}\|g\|^2 \left(1 - \frac{1}{2}\frac{\|B\| + \lambda}{\|B\| + \lambda}\right) = \frac{\|g\|^2}{2(\|B\| + \lambda)}.$$

□

Lemma 63. *Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklad F4 a s_i je vektor určený podle (Q1). Pak pokud $\|g_i\| > 0$, $\|B_i\| > 0$, $B_i \succeq 0$ a $\lambda_i \geq \overline{G}$, platí $\rho_i(s_i) \geq \underline{\rho}$, takže $i \in N_2$.*

Důkaz Podmínku $\rho(s) \geq \underline{\rho}$ můžeme zapsat ve tvaru $F_+ - F - Q_\lambda(s) \leq (\underline{\rho} - 1)Q_\lambda(s)$. Ale

$$F_+ - F - Q_\lambda(s) \leq s^T g + \frac{1}{2}\overline{G}\|s\|^2 - s^T g - \frac{1}{2}s^T B s - \frac{1}{2}\lambda\|s\|^2 = \frac{1}{2}(\overline{G} - \lambda)\|s\|^2 \leq 0,$$

neboť $B \succeq 0$ a $\lambda \geq \overline{G}$, a podle (578) platí

$$(\underline{\rho} - 1)Q_\lambda(s) \geq (1 - \underline{\rho})\underline{\nu}\frac{\|g\|^2}{4}\min\left(\frac{1}{\|B\|}, \frac{1}{\lambda}\right) \geq 0,$$

takže

$$F_+ - F - Q_\lambda(s) \leq 0 \leq (\underline{\rho} - 1)Q_\lambda(s),$$

což dává $\rho(s) \geq \underline{\rho}$. □

Věta 132. *(globální konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou kvadratické regularizace taková, že $\|B_i\| \leq \overline{B}$, $B_i \succeq 0$, $i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F4. Pak platí*

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Důkaz Předpokládejme, že existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$. Podle lemmatu 63 z $\lambda_i \geq \overline{G}$ plyne $i \in N_2$, takže λ_i se může zvětšovat pouze tehdy, když $\lambda_i < \overline{G}$. Platí tedy

$$\lambda_i \leq \max\left(\lambda_1, \overline{\gamma}\overline{G}\right) \stackrel{\Delta}{=} \overline{\lambda}, \qquad i \in N. \tag{579}$$

Z této nerovnosti plyne, že množina N_2 je nekonečná (pokud $i \in N_1$ $\forall i \geq k$, pak z (Q3a) plyne $\lambda_i \to \infty$). Použijeme-li (578), (579) a (Q1), můžeme pro $i \in N_2$ psát

$$F_i - F_{i+1} \geq -\underline{\rho}Q(s_i) \geq -\underline{\nu}\,\underline{\rho}Q(s_i(\alpha^*)) \geq \frac{\underline{\nu}\,\underline{\rho}\|g_i\|^2}{4}\min\left(\frac{1}{\|B_i\|}, \frac{1}{\lambda_i}\right) \geq \frac{\underline{\nu}\,\underline{\rho}\underline{\varepsilon}^2}{4}\min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{\lambda}}\right)$$

takže

$$F_1 - \underline{F} \geq F_1 - \lim_{i\to\infty} F_i = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i\in N_2}(F_i - F_{i+1}) \geq \sum_{i\in N_2}\frac{\underline{\nu}\,\underline{\rho}\underline{\varepsilon}^2}{4}\min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{\lambda}}\right)$$

což je spor, neboť množina N_2 je nekonečná a tudíž i výraz na pravé straně je nekonečný. □

Nejčastěji se používají optimální metody kvadratické regularizace, které určují směrový vektor podle vzorce

$$s_i(\lambda_i) = \arg\min_{s\in R^n} = Q_{\lambda_i}(s).$$

Jelikož kvadratická funkce $Q_{\lambda_i}(s)$ je pro $B_i \succeq 0$ a $\lambda_i > 0$ konvexní, určíme vektor $s_i(\lambda_i)$ jednoduše řešením soustavy rovnic $(B_i + \lambda_i I)s_i(\lambda_i) + g_i = 0$. Neoptimální metody kvadratické regularizace se používají například tehdy, je li úloha velmi rozsáhlá. Pak je třeba řešit soustavu rovnic $(B_i + \lambda_i I)s_i(\lambda_i) + g_i = 0$ iteračně a iterační proces ukončovat až když je splněna podmínka (Q1). Používáme-li k tomuto účelu metodu sdružených gradientů, je podmínka (Q1) splněna již v prvním iteračním kroku (věta 69 a důsledek 4).

Příbuznost optimálních metod kvadratické regularizace s metodami s optimálním lokálně omezeným krokem dokládá věta 124 a také skutečnost, že norma $\|s_i(\lambda_i)\|$ je klesající funkcí parametru λ_i.

Lemma 64. *Pro libovolný vektor $x \neq 0$ platí*

$$\nabla\|x\| = \frac{x}{\|x\|}, \qquad \nabla^2\|x\| = \nabla\left(\frac{x}{\|x\|}\right) = \frac{1}{\|x\|}\left(I - \frac{xx^T}{\|x\|^2}\right),$$

kde I je jednotková matice řádu n.

Důkaz Platí

$$\frac{\partial}{\partial x_k}\|x\| = \frac{\partial}{\partial x_k}\sqrt{\sum_{l=1}^{n} x_l^2} = \frac{1}{2}\frac{2x_k}{\sqrt{\sum_{l=1}^{n} x_l^2}} = \frac{x_k}{\|x\|},$$

$$\begin{aligned}\frac{\partial^2}{\partial x_k \partial x_l}\|x\| &= \frac{\partial}{\partial x_l}\left(\frac{x_k}{\|x\|}\right) = \frac{\delta_{kl}}{\|x\|} + \frac{x_k}{\|x\|^2}\frac{\partial}{\partial x_l}\|x\| \\ &= \frac{\delta_{kl}}{\|x\|} - \frac{x_k x_l}{\|x\|^3} = \frac{1}{\|x\|}\left(\delta_{kl} - \frac{x_k x_l}{\|x\|^2}\right),\end{aligned}$$

kde $\delta_{kl} = 1$, pokud $k = l$, a $\delta_{kl} = 0$, pokud $k \neq l$. □

Věta 133. *Nechť $B \succeq 0$, $\lambda > 0$ a nechť $s(\lambda) \neq 0$ je vektor, který je globálním minimem funkce $Q_\lambda(s)$. Pak $\|s(\lambda)\|$ je klesající funkcí parametru λ.*

Důkaz Podle lemmatu 64 platí

$$\|s(\lambda)\|' = \frac{s^T(\lambda)s'(\lambda)}{\|s(\lambda)\|}.$$

Jelikož matice $B + \lambda$ je podle předpokladu pozitivně definitní, je vektor $s(\lambda)$ globálním minimem funkce $Q_\lambda(s)$ právě tehdy, když $(B+\lambda I)s(\lambda) = -g$. Derivováním této rovnosti dostaneme $(B+\lambda I)s'(\lambda)+s(\lambda) = 0$, neboli

$$(s'(\lambda))^T(B + \lambda I)s'(\lambda) + s^T(\lambda)s'(\lambda) = 0,$$

což dává

$$\|s(\lambda)\|' = \frac{s^T(\lambda)s'(\lambda)}{\|s(\lambda)\|} = -\frac{(s'(\lambda))^T(B + \lambda I)s'(\lambda)}{\|s(\lambda)\|} < 0$$

. □

Závěrem dodejme, že metody kvadratické regularizace se používaly před vyvinutím metod s lokálně omezeným krokem a to zejména pro minimalizaci součtu čtverců. V současné době byly nahraženy účinnějšími metodami kubické regularizace, které jsou popsány v oddílu 7.

6 Výpočet lokálně omezeného kroku

6.1 Výpočet optimálního lokálně omezeného kroku

Optimální lokálně omezený krok s_i^* je řešením úlohy (547). Podle věty 124 platí $s_i^* = s_i(\lambda_i^*)$, kde vektor $s_i(\lambda_i^*)$ je řešením soustavy rovnic $(B_i + \lambda_i^* I)s_i(\lambda_i^*) + g_i = 0$ se symetrickou pozitivně semidefinitní maticí $B_i + \lambda_i^* I$ a $\lambda_i^* \geq 0$. Přitom

(a) $\lambda_i^* = 0$, pokud $B_i \succeq 0$ a $\|s_i(0)\| \leq \Delta_i$.
(b) $\lambda_i^* > 0$ a $\|s_i^*\| = \Delta_i$, pokud $B_i \not\succeq 0$ nebo $\|s_i(0)\| > \Delta_i$.

Pokud nenastane případ (a), řešíme přibližně úlohu (b) tak, že hledáme číslo $\lambda_i > 0$ takové, že matice $B_i + \lambda_i I$ je pozitivně semidefinitní a $\underline{\delta}\Delta_i \leq \|s_i(\lambda_i)\| \leq \overline{\delta}\Delta_i$, kde $(B_i + \lambda_i I)s_i(\lambda_i) + g_i = 0$. Protože se omezíme na jeden konkrétní iterační krok, budeme index i vynechávat.

Věta 134. *Nechť $\lambda \geq 0$, $B + \lambda I \succeq 0$ a $\|s\| \geq \underline{\delta}\Delta$, kde $(B + \lambda I)s + g = 0$. Pak vektor s splňuje podmínku (551) s $\underline{\nu} = \underline{\delta}^2$.*

Důkaz Zřejmě

$$\begin{aligned} Q(s) &= g^T s + \frac{1}{2}s^T Bs = -s^T(B + \lambda I)s + \frac{1}{2}s^T Bs \\ &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda s^T s\right) \leq -\frac{1}{2}\underline{\delta}^2\left(s^T(B + \lambda I)s + \lambda\Delta^2\right) \end{aligned}$$

a pro libovolný vektor $z \in R^n$ platí

$$\begin{aligned} Q(s+z) &= g^T(s+z) + \frac{1}{2}(s+z)^T B(s+z) = -s^T(B + \lambda I)(s+z) + \frac{1}{2}(s+z)^T B(s+z) \\ &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda(s+z)^T(s+z)\right) + \frac{1}{2}z^T(B + \lambda I)z. \end{aligned} \tag{580}$$

Nechť vektor $s^* = s + z^*$ je řešním úlohy (547). Pak $(s+z^*)^T(s+z^*) = (s^*)^T s^* \leq \Delta^2$ a $(z^*)^T(B+\lambda I)z^* \geq 0$, takže podle (580) dostaneme

$$\begin{aligned} Q(s^*) &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda(s+z^*)^T(s+z^*)\right) + \frac{1}{2}(z^*)^T(B + \lambda I)z^* \\ &\geq -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda\Delta^2\right), \end{aligned} \tag{581}$$

což po dosazení do úvodní nerovnosti dává $Q(s) \leq (1/\underline{\delta})^2 Q(s^*)$. □

Číslo $\lambda \geq 0$ vyhovující předpokladům věty 134 lze získat řešením nelineární rovnice ekvivalentní rovnici $\|s(\lambda)\| = \Delta$. Přímé použití rovnice $\|s(\lambda)\| = \Delta$ není vhodné, neboť funkce $\|s(\lambda)\|$ má póly v bodech, které odpovídají vlastním číslům matice B. Vhodnější (z hlediska omezenosti) je pro tento účel rovnice $\phi(\lambda) = 0$, kde $\phi(\lambda) = 1/\Delta - 1/\|s(\lambda)\|$. Tato rovnice se řeší pomocí Newtonovy metody.

Lemma 65. *Nechť $\phi(\lambda) = 1/\Delta - 1/\|s(\lambda)\|$, kde $(B+\lambda I)s(\lambda) + g = 0$, matice $B + \lambda I$ je pozitivně definitní a $g \neq 0$ (takže $s(\lambda) \neq 0$). Pak platí*

$$\phi'(\lambda) = -\frac{s(\lambda)^T(B + \lambda I)^{-1}s(\lambda)}{\|s(\lambda)\|^3} = -\frac{g^T(B + \lambda I)^{-3}g}{(g^T(B + \lambda I)^{-2}g)^{3/2}} < 0 \tag{582}$$

a $\phi''(\lambda) \geq 0$ (takže funkce $\phi(\lambda)$ je za daných předpokladů konvexní).

Důkaz Derivováním rovnosti $(B+\lambda I)s(\lambda)+g=0$ dostaneme $(B+\lambda I)s'(\lambda)+s(\lambda)=0$, což dává $s'(\lambda)=-(B+\lambda I)^{-1}s(\lambda)$. Podle definice funkce $\phi(\lambda)$ a lemmatu 64 pak platí

$$\phi'(\lambda)=\frac{\|s(\lambda)\|'}{\|s(\lambda)\|^2}=\frac{s(\lambda)^Ts'(\lambda)}{\|s(\lambda)\|^3}=-\frac{s(\lambda)^T(B+\lambda I)^{-1}s(\lambda)}{\|s(\lambda)\|^3}=-\frac{g^T(B+\lambda I)^{-3}g}{(g^T(B+\lambda I)^{-2}g)^{3/2}}.$$

Dalším derivováním dostaneme $(B+\lambda I)s''(\lambda)+2s'(\lambda)=0$, takže $s''(\lambda)=-2(B+\lambda I)^{-1}s'(\lambda)$ a

$$\phi''(\lambda)=\frac{s(\lambda)^Ts''(\lambda)+s'(\lambda)^Ts'(\lambda)}{\|s(\lambda)\|^3}-\frac{3(s(\lambda)^Ts'(\lambda))^2}{\|s(\lambda)\|^5}=3\frac{\|s(\lambda)\|^2\|s'(\lambda)\|^2-(s(\lambda)^Ts'(\lambda))^2}{\|s(\lambda)\|^5}$$

(neboť $s(\lambda)^Ts''(\lambda)=-2s(\lambda)^T(B+\lambda I)^{-1}s'(\lambda)=2s'(\lambda)^Ts'(\lambda)$) a podle Schwarzovy nerovnosti pak platí $\phi''(\lambda)\geq 0$. □

Důsledek 20. *Nechť jsou splněny předpoklady lemmatu 65. Nechť λ_i, $1\leq i\leq n$, jsou vlastní čísla matice B (seřazená vzestupně) a v_i, $1\leq i\leq n$, jim odpovídající ortonormální vlastní vektory. Nechť $g=\sum_{1=1}^n\gamma_iv_i$ (takže $\gamma_i=v_i^Tg$, $1\leq i\leq n$). Pak platí*

$$\phi(\lambda)=\frac{1}{\Delta}-\frac{1}{\left(\sum_{1=1}^n\frac{\gamma_i^2}{(\lambda_i+\lambda)^2}\right)^{1/2}},\qquad \phi'(\lambda)=-\frac{\sum_{1=1}^n\frac{\gamma_i^2}{(\lambda_i+\lambda)^3}}{\left(\sum_{1=1}^n\frac{\gamma_i^2}{(\lambda_i+\lambda)^2}\right)^{3/2}}. \tag{583}$$

Důkaz Jelikož matice B je symetrická, existuje rozklad $V^TBV=\Lambda$, kde $V=[v_1,\ldots,v_n]$ je ortogonální matice (takže $VV^T=V^TV=I$) a $\Lambda=\mathrm{diag}(\lambda_1,\ldots,\lambda_n)$. Pak rovnici $(B+\lambda I)s(\lambda)+g=0$ můžeme zapsat ve tvaru

$$(\Lambda+\lambda I)\tilde{s}(\lambda)+\tilde{g}=0, \tag{584}$$

kde $\tilde{s}(\lambda)=V^Ts(\lambda)$ a $\tilde{g}=V^Tg$, takže $\|\tilde{s}(\lambda)\|=\|s(\lambda)\|$ a $\|\tilde{g}\|=\|g\|$. Podle lemmatu 65 (aplikovaného na rovnici (584)) platí

$$\phi(\lambda)=\frac{1}{\Delta}-\frac{1}{\|\tilde{s}(\lambda)\|}=\frac{1}{\Delta}-\frac{1}{(\tilde{g}^T(\Lambda+\lambda I)^{-2}\tilde{g})^{1/2}},$$

$$\phi'(\lambda)=-\frac{\tilde{g}^T(\Lambda+\lambda I)^{-3}\tilde{g}}{(\tilde{g}^T(\Lambda+\lambda I)^{-2}\tilde{g})^{3/2}}.$$

Využijeme-li toho, že $\tilde{g}_i=v_i^Tg=\gamma_i$, $1\leq i\leq n$, a že matice $\Lambda+\lambda I$ je diagonální, dostaneme dokazované tvrzení. □

Poznámka 217. Aby matice $B+\lambda^*I$ byla positivně semidefinitní, musí platit $\lambda^*\geq-\lambda_1$, kde λ_1 je nejmenší vlastní číslo matice B. Abychom zjednodušili některé úvahy, budeme bez újmy na obecnosti předpokládat, že nejmenší vlastní číslo λ_1 je jednoduché. Budeme rozlišovat dva případy: regulární případ, kdy $\lambda^*>-\lambda_1$, a singulární případ, kdy $\lambda^*=-\lambda_1$. V regulárnım přıpadě existují dvě možnosti. Pokud $\max(0,-\lambda_1)<\lambda<\lambda^*$ (takže $\|s(\lambda)\|>\Delta$ a $\phi(\lambda)>0$), je krok Newtonovy metody

$$\lambda_+=\lambda+\frac{\|s(\lambda)\|^3}{s(\lambda)^T(B+\lambda I)^{-1}s(\lambda)}\left(\frac{1}{\Delta}-\frac{1}{\|s(\lambda)\|}\right)=\lambda+\frac{\|s(\lambda)\|^2}{s(\lambda)^T(B+\lambda I)^{-1}s(\lambda)}\left(\frac{\|s(\lambda)\|-\Delta}{\Delta}\right)$$

dobře definován a platí $\lambda<\lambda_+<\lambda^*$ (plyne to z konvexity funkce $\phi(\lambda)$). Pokud $\lambda^*<\lambda$ (takže $\|s(\lambda)\|<\Delta$ a $\phi(\lambda)<0$), platí $\lambda_+<\lambda^*$ a je třeba zajistit aby byla splněna podmínka $\max(0,-\lambda_1)<\lambda_+$. To lze provést použitím mezí $\underline{\lambda}<\lambda^*<\overline{\lambda}$ aktualizovaných v každém kroku algoritmu. (poznámka 224). Singulární případ nastane například tehdy, když $B\not\succeq 0$ a $g=0$, neboť rovnice $(B+\lambda^*I)s(\lambda^*)+g=0$, $\lambda^*\geq-\lambda_1$, má řešení $s(\lambda^*)$, $\|s(\lambda^*)\|=\Delta$, pouze tehdy, když $\lambda^*=\lambda_1$.

Poznámka 218. Jestliže $\gamma_1 = v_1^T g \neq 0$, lze se snadno přesvědčit (použitím vztahů (583)), že platí

$$\lim_{\lambda \downarrow -\lambda_1} \phi(\lambda) = \frac{1}{\Delta}, \qquad \lim_{\lambda \downarrow -\lambda_1} \phi'(\lambda) = -\frac{1}{|\gamma_1|}$$

a

$$\lim_{\lambda \to \infty} \phi(\lambda) = -\infty, \qquad \lim_{\lambda \to \infty} \phi'(\lambda) = -\frac{1}{\|g\|}$$

(jelikož funkce $\phi(\lambda)$ není v bodě $\lambda = \lambda_1$ diferencovatelná, používáme jednostrané limity $\lambda \downarrow -\lambda_1$, což znamená, že $\lambda \to -\lambda_1$ a $\lambda > 0$). Z těchto vztahů je patrné, že pro $\gamma_1 = v_1^T g \neq 0$ jsou funkce $\phi(\lambda)$ a $\phi'(\lambda)$ omezené v okolí bodu $\lambda = -\lambda_1$ a platí $\lambda^* > -\lambda_1$.

Poznámka 219. Singulární případ může nastat pouze tehdy, když $\gamma_1 = v_1^T g = 0$, neboť pro $\lambda^* = -\lambda_1$ platí

$$v_1^T g = -v_1^T (B + \lambda^* I) s(\lambda^*) = -v_1^T (B - \lambda_1 I) s(\lambda_1) = 0$$

(používáme vztah $Bv_1 = \lambda_1 v_1$). Označme $\phi_1 = \lim_{\lambda \downarrow -\lambda_1} \phi(\lambda)$. Jestliže $\gamma_1 = v_1^T g = 0$, vymizí v (583) člen $\gamma_1^2/(\lambda_1 + \lambda)^2$ a předpokládáme-li, že $\|g\| \neq 0$, můžeme psát $\phi_1 < 1/\Delta$. Pokud $\phi_1 > 0$, platí $\lambda^* > -\lambda_1$, neboť funkce $\phi(\lambda)$ je pro $\lambda > -\lambda_1$ spojitá a $\lim_{\lambda \to \infty} \phi(\lambda) = -\infty$. Pokud $\phi_1 \leq 0$, nastane singulární případ.

Poznámka 220. V singulárním případě nelze použít Newtonovu metodu. Abychom to ukázali, zapíšeme rovnici (584) (kde $v_1^T g = 0$) blokově ve tvaru

$$\begin{bmatrix} \lambda_1 + \lambda, & 0 \\ 0, & \Lambda_2 + \lambda I \end{bmatrix} \begin{bmatrix} v_1^T s(\lambda) \\ V_2^T s(\lambda) \end{bmatrix} = - \begin{bmatrix} 0 \\ V_2^T g \end{bmatrix}, \tag{585}$$

kde $\Lambda_2 = \mathrm{diag}(\lambda_2, \ldots, \lambda_n)$ a $V_2 = [v_2, \cdots v_n]$. Pokud $\lambda \neq -\lambda_1$, platí $v_1^T s(\lambda) = 0$, takže $\lim_{\lambda \downarrow -\lambda_1} v_1^T s(\lambda) = 0$. Pokud $\lambda = -\lambda_1$, může výraz $v_1^T s(\lambda)$ nabývat libovolných hodnot. Funkce $\phi(\lambda)$ je v tomto případě nejednoznačná. Její minimální hodnotu ϕ_1 dostaneme, položíme-li $v_1^T s(\lambda) = 0$ (neboť norma $\|s(\lambda)\| = \|V^T s(\lambda)\|$ je minimální, pokud $v_1^T s(\lambda) = 0$), takže $\phi(\lambda)$ nabývá všech hodnot v intervalu $\phi_1 \leq \phi(\lambda) \leq 1/\Delta$ a pokud $\phi_1 < 0$, nelze hodnotu $\lambda^* = -\lambda_1$ nalézt pomocí Newtonovy metody. Následující poznámka ukazuje jak lze tento poblém obejít.

Poznámka 221. Předpokládejme, že $\lambda^* = -\lambda_1$ a $\phi_1 < 0$. Nechť $\lambda > -\lambda_1$ a $\phi(\lambda) < 0$ (takže $\|s(\lambda)\| < \Delta$). Zvolíme-li vektor $z(\lambda) = \alpha(\lambda) v_1$ tak, aby platilo $\|s(\lambda) + z(\lambda)\| = \Delta$, můžeme psát

$$\begin{bmatrix} \lambda_1 + \lambda, & 0 \\ 0, & \Lambda_2 + \lambda I \end{bmatrix} \begin{bmatrix} v_1^T (s(\lambda) + z(\lambda)) \\ V_2^T (s(\lambda) + z(\lambda)) \end{bmatrix} = \begin{bmatrix} \alpha(\lambda)(\lambda_1 + \lambda) \\ 0 \end{bmatrix} - \begin{bmatrix} 0 \\ V_2^T g \end{bmatrix} \tag{586}$$

(neboť $v_1^T v_1 = 1$ a $v_1^T V_2 = 0$). Jelikož

$$\Delta = \|s(\lambda) + z(\lambda)\| \geq \|z(\lambda)\| - \|s(\lambda)\| = |\alpha(\lambda)| - \|s(\lambda)\|,$$

platí $|\alpha(\lambda)| \leq \Delta + \|d(\lambda)\| \leq 2\Delta$, a tedy i $|\alpha(\lambda)(\lambda_1 + \lambda)| \to 0$ pro $\lambda \to -\lambda_1$. Pokud $\lambda \approx -\lambda_1$, takže $|\alpha(\lambda)(\lambda_1 + \lambda)| \approx 0$, lze v rovnici (586) zanedbat první člen na pravé straně, takže vektor $s(\lambda) + z(\lambda)$ je dobrou aproximací řešení rovnice (585) a platí $\|s(\lambda) + z(\lambda)\| = \Delta$, což znamená, že vektor $s(\lambda) + z(\lambda)$ je (pro $\lambda \approx -\lambda_1$) dobrou aproximací řešení úlohy (547). Přesnější výsledek udává následující věta.

Věta 135. *Nechť $\lambda \geq 0$, $B + \lambda I \succeq 0$, $(B + \lambda I)s + g = 0$, $\|s + z\| = \Delta$ a*

$$z^T (B + \lambda I) z \leq (1 - \underline{\delta}^2) \left(s^T (B + \lambda I) s + \lambda \Delta^2 \right). \tag{587}$$

Pak vektor $s + z$ splňuje podmínku (551) s $\underline{\nu} = \underline{\delta}^2$.

Důkaz Podle (580) platí

$$Q(s+z) = -\frac{1}{2}\left(s^T(B+\lambda I)s + \lambda\Delta^2\right) + \frac{1}{2}z^T(B+\lambda I)z$$

a použijeme-li (587), dostaneme

$$Q(s+z) \leq -\frac{1}{2}\underline{\delta}^2\left(s^T(B+\lambda I)s + \lambda\Delta^2\right),$$

což spolu s (581) dává $Q(s+z) \leq (1/\underline{\delta})^2 Q(s^*)$. □

Poznámka 222. Podle poznámky 221 je výhodné volit $z = \alpha v$, kde vektor v je dobrou aproximací vlastního vektoru v_1 příslušného nejmenšímu vlastnímu číslu matice B. Tento vektor lze určit různými metodami, například pomocí Choleského rozkladu jako v programech knihovny LAPACK [2].

Poznámka 223. Nechť $s \in R^n$, $v \in R^n$ a $\|s\| < \Delta$. Číslo $\alpha \geq 0$, pro které platí $\|s + \alpha v\| = \Delta$, určujeme podle vzorců

$$\alpha = \frac{\sqrt{(v^Ts)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} - v^Ts}{\|v\|^2} = \frac{\Delta^2 - \|s\|^2}{\sqrt{(v^Ts)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} + v^Ts}.$$

První vzorec volíme pokud $v^Ts \leq 0$ a druhý v opačném případě. Oba vzorce se zjednoduší, pokud $\|v\| = 1$. Tyto vzorce lze snadno získat řešením kvadratické rovnice vzniklé roznásobením vztahu $\|s + \alpha v\|^2 = \Delta^2$.

Poznámka 224. Abychom zabránili selhání Newtonovy metody, je účelné používat a aktualizovat dolní odhad $\underline{\mu}$ pro číslo $-\lambda_1$ a meze $0 \leq \underline{\lambda} < \lambda^* < \overline{\lambda}$. V prvním iteračním kroku Newtonovy metody můžeme jako $\underline{\mu}$ zvolit maximální diagonální prvek matice $-B$. Počáteční meze $\underline{\lambda} < \lambda^* < \overline{\lambda}$ lze určit z vlastností čísla $\overline{\lambda}^*$. Jestliže $(B + \lambda^* I)s(\lambda^*) + g = 0$ a $\|s(\lambda^*)\| = \Delta$, platí

$$s(\lambda^*)^T(B+\lambda^* I)^2 s(\lambda^*) = \|g\|^2,$$

což s přihlédnutím k extremálním vlastnostem vlastních čísel matice $(B + \lambda^* I)$ dává

$$\underline{\lambda}(B) + \lambda^* \leq \frac{\|g\|}{\Delta} \leq \overline{\lambda}(B) + \lambda^*.$$

Jestliže $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$, můžeme definovat dolní mez $\underline{\lambda}$ a horní mez $\overline{\lambda}$ tak, že

$$\underline{\lambda} = \frac{\|g\|}{\Delta} - \overline{B} \leq \lambda^* \leq \frac{\|g\|}{\Delta} - \underline{B} = \overline{\lambda}$$

Lze položit $\underline{B} = -\|B\|$ a $\overline{B} = \|B\|$, nebo použít jiné odhady pro vlastní čísla matice B, například Gerschgorinovy kruhy. Dolní mez $\underline{\lambda}$ je třeba ještě upravit tak, aby platilo $\underline{\lambda} \geq 0$.

Poznámka 225. V počátečních krocích Newtonovy metody se může stát, že matice $B + \lambda I$ není pozitivně definitní. Proto je účelné použít místo Choleského rozkladu $B + \lambda I = R^T R$ Gillův-Murrayův rozklad $B + \lambda I + E = R^T R$ (definice 27). Pokud $E = 0$, je matice $B + \lambda I$ pozitivně definitní. V opačném případě poskytuje věta 29 odhad čísla, které můžeme přičíst k $\underline{\mu}$. Odhad $\underline{\mu}$ můžeme též upravit pomocí vektoru v jednotkové délky, který je dobrou aproximací vlastního vektoru v_1 příslušného nejmenšímu vlastnímu číslu matice B. Jelikož číslo $\lambda_1 + \lambda$ je nejmenším vlastním číslem matice $B + \lambda I$ a jelikož v_1 je vlastní vektor této matice příslušný vlastnímu výslu $\lambda_1 + \lambda$, můžeme psát

$$\lambda_1 + \lambda = v_1^T(B+\lambda I)v_1 \leq v^T(B+\lambda I)v$$

(předpokládáme, že $\|v\| = \|v_1\| = 1$), takže číslo $\underline{\mu} = \lambda - v^T(B+\lambda I)v \leq -\lambda_1$ je dobrým dolním odhadem čísla $-\lambda_1$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 10. Data $0 < \underline{\beta} < 1$ (obvykle $\underline{\beta} = 0.1$), $0 < \underline{\delta} < 1 < \overline{\delta}$ (obvykle $\underline{\delta} = 0.9$ a $\overline{\delta} = 1.1$), $\Delta > 0$.

Krok 1 Nechť $\underline{\mu}$ je maximální diagonální prvek matice $-B$ a $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$ (poznámka 224). Položíme $\underline{\lambda} := \|g\|/\Delta - \overline{B}$, $\overline{\lambda} := \|g\|/\Delta - \underline{B}$, $\lambda := \max(0,\ \underline{\mu}, \underline{\lambda})$ a $k := 0$.

Krok 2 Položíme $\underline{\lambda} := \max(0,\ \underline{\mu}, \underline{\lambda})$. Jestliže $k > 0$ a $\lambda \leq \underline{\mu}$, položíme $\lambda := \max(\sqrt{\underline{\lambda}\,\overline{\lambda}},\ \underline{\lambda} + \underline{\beta}(\overline{\lambda} - \underline{\lambda}))$.

Krok 3 Určíme Gillův-Murrayův rozklad $B + \lambda I + E = R^T R$ a položíme $k := k + 1$. Je-li $E = 0$ (takže $B + \lambda I \succ 0$), přejdeme na krok 4. V opačném případě určíme vektor $v \in R^n$ takový, že $\|v\| = 1$ a $v^T(B + \lambda I)v < 0$ (věta 29), položíme $\underline{\mu} := \max(\underline{\mu},\ \lambda - v^T(B + \lambda I)v)$ a přejdeme na krok 2.

Krok 4 Určíme vektor $s \in R^n$ řešením rovnice $R^T R s + g = 0$. Jestliže $\|s\| > \overline{\delta}\Delta$, položíme $\underline{\lambda} := \lambda$ a přejdeme na krok 6. Jestliže $\underline{\delta}\Delta \leq \|s\| \leq \overline{\delta}\Delta$ ukončíme výpočet. Jestliže $\|s\| < \underline{\delta}\Delta$ a $\lambda = 0$ ukončíme výpočet. Jestliže $\|s\| < \underline{\delta}\Delta$ a $\lambda > 0$ položíme $\overline{\lambda} := \lambda$ a přejdeme na krok 5.

Krok 5 Určíme vektor $v \in R^n$ tak, aby tento vektor byl dobrou aproximací vlastního vektoru matice B příslušného vlastnímu číslu $\underline{\lambda}(B)$ a aby platilo $\|v\| = 1$ a $v^T s \geq 0$ (tento vektor lze určit z rozkladu $R^T R$ způsobem, který používají programy knihovny LAPACK). Určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha v\| = \Delta$ (poznámka 223). Jestliže $\alpha^2\|Rv\|^2 \leq (1 - \underline{\delta}^2)(\|Rs\|^2 + \lambda\Delta^2)$, položíme $s := s + \alpha v$ a ukončíme výpočet. V opačném případě položíme $\underline{\mu} := \max(\underline{\mu},\ \lambda - \|Rv\|^2)$ a přejdeme na krok 6.

Krok 6 Určíme vektor $v \in R^n$ řešením rovnice $R^T v = s$, položíme

$$\lambda := \lambda + \frac{\|s\|^2}{\|v\|^2}\left(\frac{\|s\| - \Delta}{\Delta}\right)$$

a přejdeme na krok 2

6.2 Využití směru největšího spádu (metody psí nohy)

Nevýhodou metod s optimálním lokálně omezeným krokem je nutnost řešení úlohy (547), což vyžaduje opakované řešení soustavy $(B_i + \lambda I)s_i(\lambda) + g_i = 0$, která obsahuje n rovnic o n neznámých. V průměru se tato soustava řeší 2-3 krát v každém iteračním kroku, ale v singulárním případě může být tento počet mnohem vyšší. Proto se úloha (547) často nahražuje úlohou

$$s_i = s_i(\alpha^*, \beta^*) = \arg \min_{\|s(\alpha,\beta)\| \leq \Delta_i} Q_i(s(\alpha, \beta)), \tag{588}$$

kde

$$s(\alpha, \beta) = -(\alpha g_i + \beta B_i^{-1} g_i).$$

Věta 136. *Směrový vektor $s_i \in R^n$ určený podle (588) vyhovuje podmínkám (T1a)–(T1c) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\nu} = 1$.*

Důkaz (a) Podmínka (T1a) je přímo součástí podmínky (588). Předpokládejme, že $s_i(\alpha^*, \beta^*) \in R^n$ je řešením úlohy (588), přičemž $\|s_i(\alpha^*, \beta^*)\| < \Delta_i$. Pak

$$Q_i(s(\alpha, \beta)) = \frac{1}{2}\alpha^2 g_i^T B_i g_i + \alpha\beta g_i^T g_i + \frac{1}{2}\beta^2 g_i^T B_i^{-1} g_i - \alpha g_i^T g_i - \beta g_i^T B_i^{-1} g_i$$

je ryze konvexní kvadratická funkce a

$$\begin{aligned} \frac{\partial Q_i(s(\alpha^*, \beta^*))}{\partial \alpha} &= \alpha^* g_i^T B_i g_i + (\beta^* - 1) g_i^T g_i \ = 0, \\ \frac{\partial Q_i(s(\alpha^*, \beta^*))}{\partial \beta} &= \alpha^* g_i^T g_i + (\beta^* - 1) g_i^T B_i^{-1} g_i = 0, \end{aligned}$$

neboli $\alpha^* = 0$, $\beta^* = 1$, takže $\omega_i(s_i(\alpha^*, \beta^*)) = 0$ a

$$-Q_i(s_i(\alpha^*, \beta^*)) = g_i^T B_i^{-1} g_i - \frac{1}{2} g_i^T B_i^{-1} g_i = \frac{1}{2} g_i^T B_i^{-1} g_i \geq \frac{1}{2} \frac{\|g_i\|^2}{\|B_i\|}.$$

(b) Nechť $\|s_i(\alpha^*, \beta^*)\| = \Delta_i$. Podle (588) musí být $Q_i(s_i(\alpha^*, \beta^*)) \leq Q_i(s_i(\alpha^*, 0)) = Q_i(s_i(\alpha^*))$, kde $s_i(\alpha^*)$ je řešením úlohy (548), takže podle lemmatu 59 platí

$$-Q_i(s_i(\alpha^*, \beta^*)) \geq -Q_i(s_i(\alpha^*)) \geq \frac{1}{2} \|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|B_i\|}\right).$$

□

Úloha (588) má dimenzi 2 a soustava rovnic s maticí B_i se řeší pouze jednou (k určení vektoru $B_i^{-1} g_i$). Vektor s_i získaný řešením úlohy (588) vyhovuje podle věty 136 podmínkám (T1a)–(T1c) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\nu} = 1$ a jeho použitím dostaneme metody, které konvergují téměř stejně dobře jako metody s optimálním lokálně omezeným krokem. Ukazuje se že efektivita metod založených na promítání do podprostoru generovaného vektory g_i a $B_i^{-1} g_i$ se příliš nezmění nahradíme-li přesné řešení úlohy (588) speciálním přibližným výběrem koeficientů α a β, který se nazývá metodou psí nohy (název této metody pochází od jejího autora M.J.D.Powella).

Metoda psí nohy je založena na použití Cauchyova vektoru s_C a Newtonova vektoru s_N, kde

$$s_C = -\frac{g^T g}{g^T B g} g, \qquad s_N = -B^{-1} g.$$

Cauchyův vektor je spádovým směrem právě tehdy, platí-li $g^T B g > 0$. Proto budeme rozlišovat dva případy, buď $g^T B g > 0$ nebo $g^T B g \leq 0$. Jestliže $g^T B g \leq 0$, můžeme položit $s = -(\Delta/\|g\|)g$, neboť v tomto případě pro $\alpha \geq 0$ platí

$$s^T(g + \alpha B s) = -\frac{\Delta}{\|g\|}\left(g^T g - \frac{\alpha \Delta}{\|g\|} g^T B g\right) \leq -\frac{\Delta}{\|g\|} g^T g,$$

takže kvadratická funkce $Q(x + \alpha s)$ (funkce proměnné α, jejíž derivace je $s^T(g + \alpha B s)$) klesá pro $\alpha \geq 0$. Jestliže $g^T B g > 0$ a $\|s_C\| \geq \Delta$, můžeme opět položit $s = -(\Delta/\|g\|)g$. Platí totiž

$$s^T(g + \alpha B s) = -\frac{\Delta}{\|g\|}\left(g^T g - \frac{\alpha \Delta}{\|g\|} g^T B g\right) = -\Delta \|g\| \left(1 - \alpha \frac{\Delta}{\|s_C\|}\right),$$

takže funkce $Q(x + \alpha s)$ klesá pro $0 \leq \alpha < \|s_C\|/\Delta$ a nabývá svého minima pro $\alpha = \|s_C\|/\Delta \geq 1$.

Pokud $g^T B g > 0$ a $\|s_C\| < \Delta$, mohou opět nastat dva případy, platí buď $(s_N - s_C)^T s_C \geq 0$ nebo $(s_N - s_C)^T s_C < 0$.

Věta 137. *Nechť $g^T B g > 0$. Jestliže $(s_N - s_C)^T s_C \geq 0$, platí $0 < s_C^T s_C / s_C^T s_N \leq 1$ a pokud*

$$\frac{s_C^T s_C}{s_C^T s_N} \leq \tau \leq 1, \tag{589}$$

je kvadratická funkce $Q(s_C + \alpha(\tau s_N - s_C))$ nerostoucí pro $0 \leq \alpha \leq 1$ (jestliže $(s_N - s_C)^T s_C > 0$, je tato funkce klesající a nabývá svého minima pro $\alpha = 1$). Dále platí $\|\tau s_N\| \geq \|s_C\|$ (rovnost nastane právě tehdy, platí-li $\tau = s_C^T s_C / s_C^T s_N$ a jsou-li vektory s_C a s_N rovnoběžné). Jestliže $(s_N - s_C)^T s_C < 0$, kvadratická funkce $Q(s_C + \alpha(s_C - s_N))$ klesá pro $\alpha \geq 0$.

Důkaz (a) Nechť $g^T B g > 0$ a $(s_N - s_C)^T s_C \geq 0$. Prostým dosazením dostaneme

$$(s_N - s_C)^T s_C = \frac{g^T g}{(g^T B g)^2}\left(g^T B g g^T B^{-1} g - (g^T g)^2\right), \tag{590}$$

takže nerovnost $(s_N - s_C)^T s_C \geq 0$ je splněna právě tehdy, jestliže $g^T B g g^T B^{-1} g - (g^T g)^2 \geq 0$ (je-li matice B positivně definitní plyne tato nerovnost ze Schwarzovy nerovnosti). Musí tedy platit $g^T B g g^T B^{-1} g > 0$, neboli

$$\frac{s_C^T s_C}{s_N^T s_C} = \frac{(g^T g)^2}{g^T B^{-1} g g^T B g} > 0,$$

což spolu s nerovností $(s_N - s_C)^T s_C \geq 0$ dává $0 < s_C^T s_C / s_C^T s_N \leq 1$. Jestliže $g^T B g > 0$ a číslo τ je určeno podle (589) (takže $\tau \geq (g^T g)^2/(g^T B^{-1} g g^T B g)$), můžeme psát

$$\begin{aligned}
(\tau s_N - s_C)^T s_C &= \frac{g^T g}{(g^T B g)^2} \left(\tau g^T B g g^T B^{-1} g - (g^T g)^2\right) \geq 0, \qquad (591)\\
(\tau s_N - s_C)^T B (\tau s_N - s_C) &= \tau^2 g^T B^{-1} g - 2\tau \frac{(g^T g)^2}{g^T B g} + \frac{(g^T g)^2}{g^T B g}\\
&= \frac{\tau}{g^T B g} \left(\tau g^T B g g^T B^{-1} g - (g^T g)^2\right) + (1-\tau)\frac{(g^T g)^2}{g^T B g} \geq 0,
\end{aligned}$$

přičemž poslední nerovnost je rovností právě tehdy, když $(s_N - s_C)^T s_C = 0$ (kdy nutně $\tau = 1$). Dále platí

$$\begin{aligned}
(\tau s_N - s_C)^T (g + B s_C) &= (\tau s_N - s_C)^T (\tau g + B s_C) + (1-\tau)(\tau s_N - s_C)^T g\\
&= (\tau s_N - s_C)^T B (\tau B^{-1} g + s_C) + (1-\tau)(\tau s_N - s_C)^T g\\
&= -(\tau s_N - s_C)^T B (\tau s_N - s_C) - (1-\tau)\frac{g^T B g}{g^T g}(\tau s_N - s_C)^T s_C,
\end{aligned}$$

takže pro derivaci kvadratické funkce $Q(s_C + \alpha(\tau s_N - s_C))$ dostaneme

$$\begin{aligned}
(\tau s_N - s_C)^T (g + B(s_C + \alpha(\tau s_N - s_C)) = &-(\tau s_N - s_C)^T B (\tau s_N - s_C)(1-\alpha)\\
&-(1-\tau)\frac{g^T B g}{g^T g}(\tau s_N - s_C)^T s_C.
\end{aligned}$$

Pokud $0 \leq \alpha \leq 1$, je tato derivace nekladná, takže funkce $Q(s_C + \alpha(\tau s_N - s_C))$ je nerostoucí (jestliže $(s_N - s_C)^T s_C > 0$, je tato funkce klesající a nabývá svého minima pro $\alpha = 1$). Vztah $\|\tau s_N\| \geq \|s_C\|$ plyne z nerovnosti

$$\begin{aligned}
\|\tau s_N\|^2 &= (s_C + \tau s_N - s_C)^T (s_C + \tau s_N - s_C)\\
&= \|s_C\|^2 + 2(\tau s_N - s_C)^T s_C + \|\tau s_N - s_C\|^2 \geq \|s_C\|^2.
\end{aligned}$$

Rovnost nastane právě tehdy, když $\tau s_N = s_C$, neboli když $\tau = s_C^T s_C / s_C^T s_N$ a vektory s_C a s_N jsou rovnoběžné.

(b) Nechť $g^T B g > 0$ a $(s_N - s_C)^T s_C < 0$. Ze vztahu (590) plyne, že $g^T B g g^T B^{-1} g - (g^T g)^2 < 0$. Platí tedy

$$\begin{aligned}
(s_N - s_C)^T B (s_N - s_C) &= g^T B^{-1} g - 2\frac{(g^T g)^2}{g^T B g} + \frac{(g^T g)^2}{g^T B g}\\
&= \frac{1}{g^T B g}\left(g^T B g g^T B^{-1} g - (g^T g)^2\right) < 0,\\
(s_N - s_C)^T (g + B s_C) &= -(s_N - s_C)^T B (s_N - s_C) > 0,
\end{aligned}$$

takže pro derivaci kvadratické funkce $Q(s_C + \alpha(s_C - s_C))$ dostaneme

$$(s_C - s_N)^T (g + B(s_C + \alpha(s_C - s_N)) = (1+\alpha)(s_C - s_N)^T B (s_C - s_N) < 0.$$

□

Věta 137 tvoří teoretický podklad pro jednoduchou a dvojitou metodu psí nohy. V případě, že $g^T Bg > 0$, $\|s_C\| < \Delta$ a $(s_N - s_C)^T s_C \geq 0$ pokládáme

$$\begin{aligned} s &= s_N, & \|s_N\| \leq \Delta, \\ s &= s_C + \alpha(\tau s_N - s_C), & \|s_N\| > \Delta, \end{aligned}$$

kde $\max(\underline{\tau}, \Delta/\|s_N\|) \leq \tau \leq 1$, $\underline{\tau} = s_C^T s_C / s_C^T s_N$, a kde parametr $0 < \alpha < 1$ se vybírá tak, aby platilo $\|s\| = \Delta$ (poznámka 223). Jednoduchá metoda psí nohy používá hodnotu $\tau = 1$. Dvojitá metoda psí nohy používá hodnotu $\tau = \max(\underline{\tau}, \Delta/\|s_N\|)$. Poznamenejme, že nemůže nastat případ, kdy $\|s_N\| \leq \Delta < \|s_C\|$, neboť podle věty 137 platí $\|s_N\| \geq \|s_C\|$, pokud $g^T Bg > 0$ a $(s_N - s_C)^T s_C \geq 0$. V případě, že $g^T Bg > 0$, $\|s_C\| < \Delta$ a $(s_N - s_C)^T s_C < 0$, není matice B positivně semidefinitní a nemá význam pokládat $s = s_N$, pokud $\|s_N\| \leq \Delta$, neboť tento vektor není minimem kvadratické funkce $Q(s)$. V tomto případě pokládáme

$$s = s_C + \alpha(s_C - s_N),$$

kde parametr $0 < \alpha < 1$ se vybírá tak, aby platilo $\|s\| = \Delta$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 11. Data $\Delta > 0$.

Krok 1 Pokud $g^T Bg \leq 0$, položíme $s := -(\Delta/\|g\|)g$ a ukončíme výpočet.

Krok 2 Vypočteme Cauchyův vektor $s_C = -(g^T g / g^T Bg)g$. Pokud $\|s_C\| \geq \Delta$, položíme $s := -(\Delta/\|g\|)g$ a ukončíme výpočet.

Krok 3 Vypočteme Newtonův vektor $s_N = -B^{-1}g$. Jestliže $(s_N - s_C)^T s_C \geq 0$ a $\|s_N\| \leq \Delta$, položíme $s := s_N$ a ukončíme výpočet.

Krok 4 Jestliže $(s_N - s_C)^T s_C \geq 0$ a $\|s_N\| > \Delta$, určíme číslo τ tak, aby platilo $\max(\underline{\tau}, \Delta/\|s_N\|) \leq \tau \leq 1$, kde $\underline{\tau} = s_C^T s_C / s_C^T s_N$, vybereme $\alpha > 0$ tak aby platilo $\|s_C + \alpha(\tau s_N - s_C)\| = \Delta$ (poznámka 223), položíme $s := s_C + \alpha(\tau s_N - s_C)$ a ukončíme výpočet.

Krok 5 Jestliže $(s_N - s_C)^T s_C < 0$, zvolíme $\alpha > 0$ tak aby platilo $\|s_C + \alpha(s_C - s_N)\| = \Delta$, položíme $s := s_C + \alpha(s_C - s_N)$ a ukončíme výpočet.

Věta 138. *Směrový vektor získaný algoritmem 11 vyhovuje podmínkám (T1a)–(T1d) s* $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$, $\underline{\nu} = 1$.

Důkaz (a) Vzhledem k tomu, že buď $\|s\| = \Delta$ nebo $\|s\| < \Delta$ a přitom $s = s_N$, jsou splněny podmínky (T1a)–(T1b) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, a $\overline{\omega} = 0$. Jestliže $g^T Bg \leq 0$ nebo $\|s_C\| \geq \Delta$, platí $s = s(\alpha^*)$, kde vektor $s(\alpha^*)$ je řešením úlohy (548). Podle lemmatu 59 je tedy splněna podmínka (T1c) s $\underline{\nu} = 1$ a podle lemmatu 63 platí (T1d) s $\underline{\nu} = 1$. Jelikož $s_C = s(\alpha^*)$, pokud $\|s_C\| < \Delta$, dostaneme v tomto případě stejné nerovnosti pro vektor s_C.

(b) Pokud $g^T Bg > 0$, $\|s_C\| < \Delta$ a $(s_N - s_C)^T s_C \geq 0$, je $s = s_C + \alpha(\tau s_N - s_C)$, kde $0 \leq \alpha \leq 1$. Jelikož podle věty 137 platí $Q(s) \leq Q(s_C)$ a použitím (591) dostaneme

$$g^T s = g^T(s_C + \alpha(\tau s_N - s_C)) = g^T s_C - \alpha \frac{g^T Bg}{g^T g} s_C^T(\tau s_N - s_C) \leq g^T s_C,$$

splňuje vektor $s = s_C + \alpha(\tau s_N - s_C)$ podmínky (T1c)–(T1d) s $\underline{\nu} = 1$.
(c) Pokud $g^T Bg > 0$, $\|s_C\| < \Delta$ a $(s_N - s_C)^T s_C < 0$, je $s = s_C + \alpha(s_C - s_N)$, kde $\alpha \geq 0$. Jelikož podle věty 137 platí $Q(s) \leq Q(s_C)$ a s použitím nerovnosti $(s_N - s_C)^T s_C < 0$ dostaneme

$$g^T s = g^T(s_C + \alpha(s_C - s_N)) = g^T s_C + \alpha \frac{g^T Bg}{g^T g} s_C^T(s_N - s_C) \leq g^T s_C,$$

splňuje vektor $s = s_C + \alpha(s_C - s_N)$ podmínky (T1c)–(T1d) s $\underline{\nu} = 1$. □

Poznámka 226. V algoritmu 11 se předpokládá, že matice B je regulární (v opačném případě nelze použít Newtonův vektor $s_N = -B^{-1}g$). Tento nedostatek se obvykle obchází tím, že se matice B při určování Choleského rozkladu mírně modifikuje. Bližší podrobnosti jsou uvedeny v části 6.3.

6.3 Nepřesné metody s lokálně omezeným krokem

K určení lokálně omezeného kroku můžeme velmi efektivně použít předpodmíněnou metodu sdružených gradientů aplikovanou na minimalizaci kvadratické funkce

$$Q(s) = g^T s + \frac{1}{2} s^T B s. \tag{592}$$

Připomeňme, že předpodmíněná metoda sdružených gradientů používá rekurentní vztahy

$$s_1 = 0, \qquad g_1 = g, \qquad p_1 = -C^{-1}g$$

a

$$\begin{aligned} q_i &= Bp_i, & \alpha_i &= g_i^T C^{-1} g_i, / p_i^T q_i, \\ s_{i+1} &= s_i + \alpha_i p_i, & g_{i+1} &= g_i + \alpha_i q_i, \\ \beta_i = g_{i+1}^T C^{-1} g_{i+1} / g_i^T C^{-1} g_i, & & p_{i+1} &= -C^{-1} g_{i+1} + \beta_i p_i \end{aligned}$$

pro $1 \le i \le n$. Můžeme používat větu 69, větu 71 a důsledek 5.

Chceme-li určit lokálně omezený krok pomocí metody sdružených gradientů, zastavujeme iterační proces nejen tehdy, když $\|g_i\| \le \omega \|g\|$ (kde $0 < \omega \le \overline{\omega} < 1$), ale také tehdy, když $\|s_i\| < \Delta$ a buď $p_i^T B p_i \le 0$ nebo $\|s_{i+1}\| \ge \Delta$. Pokud $p_i^T B p_i \le 0$, můžeme použít větu 71, podle které $Q(s_i + \alpha p_i) < Q(s_i)$ a $g^T(s_i + \alpha p_i) < g^T s_i$ pro $\alpha > 0$. Pokud $\|s_{i+1}\| \ge \Delta$, můžeme použít větu 69, podle které $Q(s_i + \alpha p_i) < Q(s_i)$ a $g^T(s_i + \alpha p_i) < g^T s_i$ pro $0 < \alpha \le \alpha_i$. V obou případech určíme číslo $\alpha_i \ge 0$ tak, aby platilo $\|s_i + \alpha_i p_i\| = \Delta$ (poznámka 223) a položíme $s = s_i + \alpha_i p_i$.

Dosavadní úvahy tvoří základ jednoduchého algoritmu:

Algoritmus 12. Data $C \succ 0$, $0 < \omega \le \overline{\omega} < 1$, $\Delta > 0$, $m \ge n$ (obvykle $m = n + 3$).

Krok 1 Položíme $s := 0$, $r := -g$, $v := C^{-1}r$, $\sigma := r^T v$, $\overline{\sigma} := \sigma$, $p := r$ a $k := 1$.

Krok 2 Položíme $\rho := \sigma$, vypočteme vektor $q = Bp$ a číslo $\tau = p^T q$. Jestliže $\tau \le 0$, určíme číslo $\alpha \ge 0$ tak, aby platilo $\|s + \alpha p\| = \Delta$, položíme $s := s + \alpha p$ a ukončíme výpočet.

Krok 3 Položíme $\alpha := \rho / \tau$. Jestliže $\|s + \alpha p\| \ge \Delta$, určíme číslo $\alpha \ge 0$ tak, aby platilo $\|s + \alpha p\| = \Delta$, položíme $s := s + \alpha p$ a ukončíme výpočet.

Krok 4 Položíme $s := s + \alpha p$, $r := r - \alpha q$ $v := C^{-1}r$ a $\sigma := r^T v$. Jestliže $\sigma \le \omega^2 \overline{\sigma}$ nebo $k \ge m$, ukončíme výpočet.

Krok 5 Položíme $\beta := \sigma / \rho$, $p := r + \beta p$, $k := k + 1$ a přejdeme na krok 2.

Věta 139. *Směrový vektor získaný algoritmem 12 vyhovuje podmínkám (T1a)–(T1d) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} < 1$, $\underline{\nu} = 1/\kappa(C)$.*

Důkaz Jak již bylo zmíněno, z věty 69 a věty 71 plyne, že $Q(s) < Q(s_i)$ a $g^T s < g^T s_i$. Pokud $i > 1$, můžeme použít důsledek 5 podle kterého platí

$$Q(s) \le Q(s_2) \le -\frac{\|g\|^2}{2\kappa(C)\|B\|}, \qquad g^T s \le g^T s_2 \le -\frac{\|g\|^2}{\kappa(C)\|B\|}.$$

Pokud $i = 1$, můžeme použít lemma 59 a lemma 579, takže

$$Q(s) = Q(s(\alpha^*)) \le -\frac{1}{2}\|g\|\|s\|, \qquad g^T s = g^T s(\alpha^*) \le -\|g\|\|s\|.$$

□

Směrový vektor s_i získaný metodou sdružených gradientů můžeme kombinovat s vektorem s_N tak jako v metodách psí nohy (kde kombinujeme vektor $s_C = s_2$ s vektorem s_N). Ztratí se však výlučně iterační charakter nepřesné metody s lokálně omezeným krokem (podřebujeme získat vektor s_N přímým řešením soustavy lineárních rovnic). Nicméně použití několika kroků metody sdružených gradientů může urychlit konvergenci metody psí nohy. Následující věta udává teoretický podklad pro konstrukci vícekrokové metody psí nohy.

Věta 140. *Nechť jsou splněny předpoklady věty 69 pro $i \geq 1$, přičemž $\|s_i\| < \Delta$ a $Bs_i + g \neq 0$. Nechť $s_N \in R^n$ je vektor takový, že $Bs_N + g = 0$. Pak pro $0 \leq \alpha < 1$ platí*

$$\frac{dQ(s_i + \alpha(s_N - s_i))}{d\alpha} = (1-\alpha)(s_N - s_i)^T g_i.$$

Důkaz Jelikož

$$Q(s_i + \alpha(s_N - s_i)) = g^T(s_i + \alpha(s_N - s_i)) + \frac{1}{2}(s_i + \alpha(s_N - s_i))^T B(s_i + \alpha(s_N - s_i)),$$

platí

$$\begin{aligned}
\frac{dQ(s_i + \alpha(s_N - s_i))}{d\alpha} &= (s_N - s_i)^T g + (s_N - s_i)^T B(s_i + \alpha(s_N - s_i)) \\
&= (s_N - s_i)^T B(s_i - s_N + \alpha(s_N - s_i)) \\
&= (1-\alpha)(s_N - s_i)^T B(s_i - s_N) \\
&= (1-\alpha)(s_N - s_i)^T (Bs_i + g) \\
&= (1-\alpha)(s_N - s_i)^T g_i.
\end{aligned}$$

□

Z věty 140 vyplývá, že pokud $(s_N - s_i)^T g_i \leq 0$, je funkce $Q(s_i + \alpha(s_N - s_i))$ nerostoucí pro $0 \leq \alpha \leq 1$ (pokud $(s_N - s_i)^T g_i < 0$ je tato funkce klesající pro $0 \leq \alpha < 1$). Jestliže naopak $(s_N - s_i)^T g_i > 0$ je funkce $Q(s_i + \alpha(s_i - s_N))$ klesající pro $\alpha \geq 0$. Uvedené úvahy tvoří základ následujícího algoritmu:

Algoritmus 13. Data $0 < \Delta$, $m \ll n$.

Krok 1 Jako v algoritmu 12.

Krok 2 Jako v algoritmu 12.

Krok 3 Jako v algoritmu 12.

Krok 4 Položíme $s := s + \alpha p$, $r := r - \alpha q$ a $\sigma := \|r\|^2$. Jestliže $k < m$ položíme $\beta := \sigma/\rho$, $p := r + \beta p$, $k := k + 1$ a přejdeme na krok 2.

Krok 5 Řešíme soustavu rovnic $Bs^* + g = 0$. Pokud $(s^* - s)^T r \geq 0$ a $\|s^*\| \leq \Delta$, položíme $s := s^*$ a ukončíme výpočet. Pokud $(s^* - s)^T r \geq 0$ a $\|s^*\| > \Delta$, určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha(s^* - s)\| = \Delta$, položíme $s := s + \alpha(s^* - s)$ a ukončíme výpočet. Pokud $(s^* - s)^T r < 0$, určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha(s - s^*)\| = \Delta$, položíme $s := s + \alpha(s - s^*)$ a ukončíme výpočet.

Obvykle volíme $m = 5$. Pro $m = 1$ dostaneme jednoduchou metodu psí nohy popsanou v oddílu 6.2.

6.4 Použití symetrické Lanczosovy metody

Metodu sdružených gradientů popsanou v předchozím odstavci musíme přerušit, pokud v i-tém iteračním kroku platí buď $g_i^T B g_i \leq 0$ nebo $s_{i+1} \geq \Delta$. V tomto případě určíme směrový vektor d takový, že $\|d\| = \Delta$ a ukončíme výpočet. Abychom nalezli přesnější aproximaci optimálního lokálně omezeného kroku je třeba v iteračním procesu pokračovat. K tomuto účelu lze použít symetrický Lanczosův proces.

Definice 44. *Nechť $B \in R^{n\times n}$ je symetrická matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$q_0 = 0, \qquad \delta_1 q_1 = g$$

a

$$\gamma_i = q_i^T B q_i, \quad \delta_{i+1} q_{i+1} = B q_i - \gamma_i q_i - \delta_i q_{i-1}$$

pro $1 \leq i \leq n$, kde koeficienty $\delta_i \geq 0$, $1 \leq i \leq n$, se volí tak, aby vektory q_i, $1 \leq i \leq n$, měly jednotkovou normu, nazveme symetrickým Lanczosovým procesem (LS) určeným maticí B a vektorem g.

Poznámka 227. Nechť $\delta_i \neq 0$, $1 \leq i \leq k$, pro nějaký index $1 \leq k \leq n$. Pak podle definice 44 platí $g = Q_k(\delta_1 e_1)$ a

$$BQ_k = Q_k T_k + \delta_{k+1} q_{k+1} e_k^T \tag{593}$$

kde $Q_k = [q_1, q_2, \ldots, q_{k-1}, q_k]$, $e_1^T = [1, 0, \ldots, 0, 0]$, $e_k^T = [0, 0, \ldots, 0, 1]$ a

$$T_k = \begin{bmatrix} \gamma_1, & \delta_2, & \ldots, & 0, & 0 \\ \delta_2, & \gamma_2, & \ldots, & 0, & 0 \\ --, & --, & --, & --, & -- \\ 0, & 0, & \ldots, & \gamma_{k-1}, & \delta_k \\ 0, & 0, & \ldots, & \delta_k, & \gamma_k \end{bmatrix} \tag{594}$$

(matice $T_k \in R^{k\times k}$ je tridiagonální). Můžeme se o tom snadno přesvědčit roznásobením a použitím rekurentních vztahů metody LS.

Věta 141. *Uvažujme symetrický Lanczosův proces (LS) určený symetrickou maticí $B \in R^{n\times n}$ a vektorem $g \in R^n$. Nechť $\delta_i \neq 0$, $1 \leq i \leq k$, pro nějaký index $1 \leq k \leq n$. Pak vektory q_i, $1 \leq i \leq k$, tvoří ortonormální bázi v Krylovově podprostoru $\mathcal{K}_k = \text{span}\,(g, Bg, \ldots, B^{k-1}g)$.*

Důkaz (indukcí). Pro $k = 1$ je tvrzení zřejmé, neboť $q_1 = g/\|g\|$. Předpokládejme, že tvrzení platí pro nějaký index $1 \leq k < n$ a že $\delta_{k+1} \neq 0$. Podle indukčního předpokladu platí $Q_k^T Q_k = I$, takže $Q_k^T B Q_k = T_k + \delta_{k+1} Q_k^T q_{k+1} e_k^T$. Matice $Q_k^T B Q_k$ je symetrická stejně jako matice T_k, takže nutně $Q_{k-1}^T q_{k+1} = 0$ (v opačném případě by matice $Q_k^T q_{k+1} e_k^T$ nebyla symetrická). Dále podle definice 44 platí $\delta_{k+1} q_k^T q_{k+1} = q_k^T B q_k - \gamma_k = \gamma_k - \gamma_k = 0$. Vektor q_{k+1} je tedy ortogonální k vektorům q_i, $1 \leq i \leq k$, a má jednotkovou normu. Podle definice 44 leží vektory q_i, $1 \leq i \leq k+1$ v Krylovově podprostoru $\mathcal{K}_{k+1}$ a jelikož jsou vzájemně ortogonální a mají jednotkovou normu, tvoří tam ortonormální bázi. □

Poznámka 228. Jelikož $Q_k^T Q_k = I$ a $Q_k^T q_{k+1} = 0$ (důkaz věty 141), můžeme psát

$$Q_k^T B Q_k = T_k,$$

takže symetrický Lanczosův proces lze použít k tridiagonalizaci matice B.

Poznámka 229. Jsou-li matice T_i, $1 \leq i \leq k$, regulární, můžeme je použít k určení stacionárních bodů s_{i+1} kvadratické funkce $Q(s)$ (definované vztahem (237)) na Krylovových podprostorech $\mathcal{K}_i$, $1 \leq i \leq k$. Jelikož $s \in \mathcal{K}_i$ právě tedy, když $s = Q_i z$, $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde z_i je stacionárním bodem funkce

$$Q(Q_i z) = \frac{1}{2} z^T Q_i^T B Q_i z + g^T Q_i z = \frac{1}{2} z^T T_i z + \delta_1 e_1^T z,$$

neboli $T_i z_i + \delta_1 e_1 = 0$ (plyne to ze vztahů $g = Q_i(\delta_1 e_1)$ a $Q_i^T Q_i = I$). Pokud $\delta_{k+1} = 0$, je vektor $s_{k+1} \in \mathcal{K}_k$ řešením soustavy rovnic $Bs + g = 0$. Podle (593) totiž platí $BQ_k = Q_k T_k$ a jelikož matice T_k je regulární, lze položit $z_k = -T_k^{-1}(\delta_1 e_1)$, což dává $Bs_{k+1} = BQ_k z_k = -Q_k T_k T_k^{-1}(\delta_1 e_1) = -Q_k(\delta_1 e_1) = -g$.

Symetrický Lanczosův proces je velmi úzce spjat s metodou sdružených gradientů. Poznamenejme, že iterační proces metody sdružených gradientů (definice 35 s $C = I$) je definován tehdy, když $p_i B p_i \neq 0$, $1 \leq i \leq k$ (nemusí nutně platit $p_i B p_i > 0$, $1 \leq i \leq k$). V tomto případě jsou vektory p_i, $1 \leq i \leq k$, lineárně nezávislé, platí (239)–(241) (s $C = I$), vektory g_i, $1 \leq i \leq k$, tvoří ortogonální bázi v Krylovově podprostoru $\mathcal{K}_k = \mathcal{L}((g, Bg, \ldots, B^{k-1}g)$ (je to ukázáno v části (a) důkazu lemmatu 21) a vektory s_{i+1}, $1 \leq i \leq k$, jsou stacionárními body (ne nutně minimy) kvadratické funkce $Q(s)$ na $\mathcal{K}_i$, $1 \leq i \leq k$.

Věta 142. *Vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 229, jsou shodné s vektory s_{i+1}, $1 \leq i \leq k$, generovanými metodou sdružených gradientů (definice 35 s $C = I$), pokud $\alpha_i \neq 0$, $1 \leq i \leq k$. Navíc platí $\gamma_1 = 1/\alpha_1$, $\varepsilon_1 = 1$ a*

$$\gamma_{i+1} = \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}}, \qquad \delta_{i+1} = \frac{\sqrt{\beta_i}}{|\alpha_i|}, \qquad \varepsilon_{i+1} = -\varepsilon_i \mathrm{sgn}(\alpha_i)$$

a

$$q_i = \varepsilon_i \frac{g_i}{\|g_i\|}$$

pro $1 \leq i \leq k$.

Důkaz Z důkazu věty 40 plyne, že vektory s_{i+1}, $1 \leq i \leq k$, určené metodou sdružených gradientů, leží v Krylovových podprostorech $\mathcal{K}_i$, $1 \leq i \leq k$, a jsou tam stacionárními body kvadratické funkce $Q(s)$. To je však právě definice vektorů s_{i+1}, $1 \leq i \leq k$, v poznámce 229. Jelikož vektory g_i, $1 \leq i \leq k$, jsou vzájemně ortogonální a leží v Krylovových podprostorech $\mathcal{K}_i$, $1 \leq i \leq k$, musí být kolineární s vektory q_i, $1 \leq i \leq k$, neboli

$$G_k = Q_k N_k E_k,$$

kde $G_k = [g_1, \ldots, g_k]$, $N_k = \mathrm{diag}\,(\|g_1\|, \ldots, \|g_k\|)$, $E_k = \mathrm{diag}\,(\varepsilon_1, \ldots, \varepsilon_k)$, (čísla ε_i, $1 \leq i \leq k$, mohou nabývat pouze hodnot 1 a -1). Položme $P_k = [p_1, \ldots, p_k]$, kde p_i, $1 \leq i \leq k$, jsou vektory použité v definici 35 s $C = I$. Pak z rekurentních vztahů metody sdružených gradientů plyne

$$G_k = P_k B_k,$$

kde

$$B_k = \begin{bmatrix} -1, & \beta_1, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}$$

je horní bidiagonální matice. Z důkazu věty 40 plyne, že matice $P_k^T B P_k$ je diagonální. Použijeme-li vztahy $\alpha_i = \|g_i\|^2 / p_i^T B p_i$, $1 \leq i \leq k$, uvedené v definici 35 s $C = I$, dostaneme

$$P_k^T B P_k = N_k D_k N_k = N_k E_k D_k E_k N_k,$$

kde $D_k = \mathrm{diag}\,(1/\alpha_1, \ldots, 1/\alpha_k)$, (neboť $E_k^2 = I$ a diagonální matice E_k, D_k komutují), takže

$$\begin{aligned} T_k &= Q_k^T B Q_k = E_k^{-1} N_k^{-1} G_k^T B G_k N_k^{-1} E_k^{-1} = E_k N_k^{-1} B_k^T P_k^T B P_k B_k N_k^{-1} E_k = \\ &= E_k N_k^{-1} B_k^T N_k E_k D_k E_k N_k B_k N_k^{-1} E_k = L_k D_k L_k^T, \end{aligned}$$

kde

$$L_k = E_k N_k^{-1} B_k^T N_k E_k = \begin{bmatrix} -1, & 0, & \ldots, & 0 \\ \beta_1 \frac{\varepsilon_1 \|g_1\|}{\varepsilon_2 \|g_2\|}, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix} = \begin{bmatrix} -1, & 0, & \ldots, & 0 \\ \varepsilon_1 \varepsilon_2 \sqrt{\beta_1}, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}$$

je je dolní bidiagonální matice. Dosadíme-li tuto matici do vyjádření pro matici T_k, můžeme psát

$$T_k = \begin{bmatrix} \frac{1}{\alpha_1}, & \frac{-\varepsilon_1\varepsilon_2\sqrt{\beta_1}}{\alpha_1}, & \dots, & 0 \\ \frac{-\varepsilon_1\varepsilon_2\sqrt{\beta_1}}{\alpha_1}, & \frac{\beta_1}{\alpha_1} + \frac{1}{\alpha_2}, & \dots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \dots, & \frac{\beta_{k-1}}{\alpha_{k-1}} + \frac{1}{\alpha_k} \end{bmatrix},$$

což porovnáním se (593) dává $\gamma_1 = 1/\alpha_1$ a

$$\gamma_{i+1} = \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}}, \qquad \delta_{i+1} = -\frac{\varepsilon_i\varepsilon_{i+1}\sqrt{\beta_i}}{\alpha_i}$$

pro $1 \le i \le k$. Jelikož $\delta_{i+1} \ge 0$, musí platit $\varepsilon_i\varepsilon_{i+1} = -\mathrm{sgn}(\alpha_i)$ pro $1 \le i \le k$. Protože podle definice 44 platí $\delta_1 q_1 = g = g_1$ a $\delta_1 \ge 0$, dostaneme $\varepsilon_1 = 1$. □

Poznámka 230. Chceme-li se zbavit střídavých znamének, můžeme položit $\tilde{Q}_k = G_k N_k^{-1} = Q_k E_k$ a $\tilde{T}_k = \tilde{Q}_k^T B \tilde{Q}_k = E_k T_k E_k$. Pak platí

$$\tilde{Q}_k^T B \tilde{Q}_k = \tilde{T}_k = E_k L_k E_k D_k E_k L_k^T E_k = \tilde{L}_k D_k \tilde{L}_k^T, \tag{595}$$

kde

$$\tilde{L}_k = E_k L_k E_k = \begin{bmatrix} -1, & 0 & \dots, & 0 \\ \sqrt{\beta_1}, & -1, & \dots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \dots, & -1 \end{bmatrix}.$$

V případě, že $\alpha_i > 0$, $1 \le i \le k$, lze matici $\tilde{T}_k$ získat tak, že ve vzorcích uvedených v definici 44 pokládáme $-\delta_1 q_1 = g$ a $-\delta_{i+1} q_{i+1} = Bq_i - \gamma_i q_i - \delta_i q_{i-1}$, $1 \le i \le k$. V matici $\tilde{T}_k$ jsou pak mimodiagonální prvky záporné.

Poznámka 231. Koeficienty symetrického Lanczosova procesu jsou prvky tridiagonální matice T_k, zatímco koeficienty metody sdružených gradientů určují prvky jejího Choleského rozkladu. Vzorce uvedené ve větě 142 lze použít k rekurentnímu výpočtu koeficientů metody sdružených gradientů z prvků matice T_k. Označíme-li $\tau_i = 1/\alpha_i$, $1 \le i \le k$, platí $\tau_1 = \gamma_1$ a

$$\tau_{i+1} = \gamma_{i+1} - \frac{\delta_{i+1}}{\tau_i}, \qquad \beta_i = \left(\frac{\delta_{i+1}}{\tau_i}\right)^2.$$

Jelikož τ_i, $1 \le i \le k$, jsou prvky diagonální matice D_k v Choleského rozkladu $T_k = L_k D_k L_k^T$, je matice T_k pozitivně definitní právě tehdy, jsou-li všechna čísla τ_i, $1 \le i \le k$, kladná.

Poznámka 232. Symetrický Lanczosův proces můžeme použít k přibližnému určení optimálního lokálně omezeného kroku. Pokládáme $s_1 = 0$ a vektory s_{i+1}, $1 \le i \le k$, hledáme tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i, \|s\| \le \Delta} \left(\frac{1}{2} s^T B s + g^T s\right). \tag{596}$$

Jelikož $s \in \mathcal{K}_i$ právě tedy, když $s = Q_i z$, kde $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i, \|z\| \le \Delta} \left(\frac{1}{2} z^T T_i z + \delta_1 e_1^T z\right) \tag{597}$$

(plyne to ze vztahů $g = Q_i(\delta_1 e_1)$, $Q_i^T Q_i = I$ a z toho, že ortogonalita sloupců matice Q_i implikuje rovnost $\|s\| = \|Q_i z\| = \|z\|$).

Je-li vektor z_i řešením úlohy (597), zajímá nás, jak dobře aproximuje vektor s_{i+1} řešení úlohy (547). Podle věty 124 je vektor z_i řešením úlohy (597) právě tehdy, existuje-li číslo $\lambda_i \geq 0$ takové že matice $T_i+\lambda_i I$ je positivně semidefinitní, $(T_i + \lambda_i I)z_i + \delta_1 e_1 = 0$, $\|z_i\| \leq \Delta$ a $\lambda_i(\|z_i\| - \Delta) = 0$. Protože $\|s_{i+1}\| = \|z_i\|$, splňuje dvojice s_{i+1}, λ_i většinu podmínek uvedených ve z větě 124. Kriteriem aproximace tedy může být hodnota $\|(B + \lambda_i)s_{i+1} + g\|$ (norma rezidua).

Věta 143. *Nechť $s_{i+1} = Q_i z_i$, kde vektor z_i je řešením úlohy (597). Pak platí*

$$(B + \lambda_i)s_{i+1} + g = \delta_{i+1} e_i^T z_i q_{i+1},$$

takže $\|(B + \lambda_i)s_{i+1} + g\| = \delta_{i+1}|e_i^T z_i|$.

Důkaz Použijeme-li vztah (593) a podmínku $(T_i + \lambda_i I)z_i + \delta_1 e_1 = 0$, dostaneme

$$\begin{aligned}(B + \lambda_i I)s_{i+1} + g &= (B + \lambda_i I)Q_i z_i + \delta_1 Q_i e_1 \\ &= Q_i((T_i + \lambda_i I)z_i + \delta_1 e_1) + \delta_{i+1} q_{i+1} e_i^T z_i \\ &= \delta_{i+1} q_{i+1} e_i^T z_i.\end{aligned}$$

Zbytek tvrzení plyne z toho, že $\|q_{i+1}\| = 1$. □

Nyní si podrobněji všimneme vlastností úlohy (597).

Definice 45. *řekneme, že matice T_i (jejíž tvar je uveden v poznámce 227) je ireducibilní, jestliže $\delta_j \neq 0$ $\forall 1 < j \leq i$.*

Věta 144. *Je-li matice T_i ireducibilní, nenastane v úloze (597) singulární případ (matice $T_i + \lambda_i I$ je positivně definitní). Je-li matice T_n ireducibilní, nenastane singulární případ ani v úloze (547). Nenastane-li singulární případ v úloze (547) a platí-li $\delta_{i+1} = 0$, je vektor $s_{i+1} = Q_i z_i$ řešením úlohy (547).*

Důkaz (a) Je-li vektor z_i řešením úlohy (597), je matice $T_i + \lambda_i I$ positivné semidefinitní. Je-li tato matice singulární, musí existovat nenulový vektor v_i takový, že $(T_i + \lambda_i I)v_i = 0$. Pak ale

$$z_i^T(T_i + \lambda_i I)v_i = -\delta_1 e_1^T v_i = 0,$$

takže vektor v_i má nulovou první složku. Předpokládejme, že matice T_i je ireducibilní. Z rovnice $T_i v_i = \lambda_i v_i$ vidíme, že je-li první složka vektoru v_i nulová a $\delta_2 \neq 0$, je i druhá složka vektoru v_i nulová (matice T_i je tridiagonální). Takto lze pokračovat dále a jsou-li všechna čísla δ_j, $1 < j \leq i$, nenulová, musí platit $v_i = 0$, což je ve sporu s předpokladem, že $v_i \neq 0$. Matice $T_i + \lambda_i I$ tedy nemůže být singulární a jelikož je positivné semidefinitní, musí být positivné definitní. V úloze (597) tedy nenastane singulární případ.

(b) Pro $i = n$ je úloha (547) ekvivalentní úloze (596) a tedy i úloze (597). Je-li matice T_n ireducibilní, nenastane singulární případ v úloze (597) a tedy ani v úloze (547).

(c) Platí-li $\delta_{i+1} = 0$, můžeme podle (593) psát $BQ_i = Q_i T_i$. Jelikož matice T_i je symetrická, můžeme ji vyjádřit ve tvaru $T_i = V_i \Lambda_i V_i^T$, kde Λ_i je diagonální matice obsahující vlastní čísla matice T_i a V_i je ortogonální (a tedy regulární) čtvercová matice, jejímiž sloupci jsou odpovídající vlastní vektory. Platí tedy

$$BQ_i = Q_i V_i \Lambda_i V_i^T \quad \Rightarrow \quad BQ_i V_i = Q_i V_i \Lambda_i,$$

takže diagonální prvky matice Λ jsou vlastními čísly matice B a sloupce matice $Q_i V_i$ jsou odpovídajícími vlastními vektory. Ukážeme, že matice Λ_i musí obsahovat nejmenší vlastní číslo λ_1 matice B. Kdyby tomu tak nebylo, musel by být příslušný vlastní vektor v_1 kolmý ke všem sloupcům matice $Q_i V_i$ (vlastní vektory odpovídající různým vlastním číslům jsou ortogonální), neboli $V_i^T Q_i^T v_1 = 0$. Protože čtvercová matice V_i je regulární, muselo by platit $Q_i^T v_1 = 0$ a jelikož vektor g je podle konstrukce rovnoběžný s vektorem q_1, také $g^T v_1 = 0$. To však není možné, neboť v úloze (547) nenastane singulární případ takže podle poznámky 217 nemůže platit $g^T v_1 = 0$. Jelikož λ_1 je vlastním číslem matice T_i, musí platit $\lambda_i \geq -\lambda_1$, takže matice $B + \lambda_i I$ je positivné definitní. Spojíme li tento fakt s tvrzením věty 143, vidíme, že jsou splněny nutné a postačující podmínky pro to, aby vektor $s_{i+1} = Q_i z_i$ byl řešením úlohy (547). □

Poznámka 233. Symetrický Lanczosův proces můžeme předpodmínit tak že ho aplikujeme na kvadratickou funkci (238). Označíme-li $\tilde{q}_i = C^{-1/2}q_i$ a $v_i = C^{-1/2}\tilde{q}_i = C^{-1}q_i$ můžeme rekurentní vztahy předpodmíněného symetrického Lancoszova procesu zapsat pomocí rekurentních vztahů

$$q_0 = 0, \quad \delta_1 q_1 = g, \quad v_1 = C^{-1}q_1$$

a

$$\gamma_i = v_i^T B v_i, \quad \delta_{i+1} q_{i+1} = B v_i - \gamma_i q_i - \delta_i q_{i-1}, \quad v_{i+1} = C^{-1} q_{i+1}$$

pro $1 \leq i \leq n$, kde koeficienty $\delta_i \geq 0$, $1 \leq i \leq n$ se volí tak, aby platilo $q_i^T v_i = 1$, $1 \leq i \leq n$. Pak vektory v_i jsou C-ortogonální a vektory q_i C^{-1}-ortogonální. Pro libovolný index $1 \leq k \leq n$ platí

$$V_k^T C V_k = Q_k^T V_k = Q_k^T C^{-1} Q_k = I$$

a

$$B V_k = Q_k T_k + \delta_{k+1} q_{k+1} e_k^T,$$

kde $T_k = V_k^T B V_k$ je symetrická tridiagonální matice. Poznamenejme, že je-li vektor z_i řešením problému (597), kde matice T_i byla získána předpodmíněným symetrickým Lanczosovým procesem, je třeba v (597) nahradit podmínku $\|s\| \leq \Delta$ podmínkou $\|s_i\|_C = \sqrt{s_i^T C s_i} \leq \Delta$.

Nyní můžeme přistoupit k popisu algoritmu pro výpočet lokálně omezeného kroku pomocí symetrického Lanczosova procesu. Nejprve uvedeme základní myšlenky, které tvoří základ tohoto algoritmu.

(a) Jelikož metoda sdružených gradientů je výpočetně ekonomičtější než symetrický Lanczosův proces, používáme metodu CG vždy, kdy je to možné. V každém iteračním kroku metody CG počítáme a ukládáme čísla δ_i, γ_i a vektory q_i (věta 142).

(b) Na začátku iteračního procesu počítáme vektory s_{i+1} metodou CG. Pokud v nějakém iteračním kroku platí $p_i^T B p_i \leq 0$, nebo $\|s_i + \alpha_i p_i\| > \Delta$, začneme pokládat $s_{i+1} = Q_i z_i$, kde vektor z_i je řešením úlohy (597).

(c) Výpočet ukončíme, platí-li $\|g_{i+1}\| \leq \omega \|g\|$ v případě, že $s_{i+1} = s_i + \alpha_i p_i$, nebo $\delta_{i+1}|e_i^T z_i| \leq \omega\|g\|$ v případě, že $s_{i+1} = Q_i z_i$. Přitom ω je předepsaná přesnost.

Dosavadní úvahy jsou použity v algoritmu 14. V tomto algoritmu je $L = 1$, používáme-li rekurentní vztahy metody sdružených gradientů, nebo $L = 0$, používáme-li rekurentní vztahy symetrické Lanczosovy metody. Podobně je $M = 1$, počítáme-li vektor s_{k+1} metodou sdružených gradientů, nebo $M = 0$, používáme-li k určení vektoru s_{i+1} řešení úlohy (597).

Algoritmus 14. Data $0 < \omega < 1$, $\Delta > 0$, $\varepsilon > 0$, $m \leq n$ (obvykle $m = \min(n, 100)$).

Krok 1 Položíme $s_1 := 0$, $g_1 := g$, $p_1 := -g$, $q_1 := g/\|g\|$, $\beta_1 := \|g\|$, $\sigma_1 := g^T g$, $\varepsilon_1 = 1$, $L := 1$, $M := 1$ a $k := 1$.

Krok 2 Jestliže $L = 0$, přejdeme na krok 5. V opačném případě vypočteme vektor $u_k := B p_k$ a číslo $\tau_k := u_k^T p_k$. Jestliže $|\tau_k| \leq \varepsilon \sigma_k$, položíme $L := 0$ a přejdeme na krok 5.

Krok 3 Položíme $\alpha_k := \sigma_k/\tau_k$ a vypočteme číslo γ_k podle věty 142, tedy $\gamma_k := 1/\alpha_k$, pokud $k = 1$, nebo $\gamma_k := 1/\alpha_k + \beta_{k-1}/\alpha_{k-1}$, pokud $k > 1$. Je-li $\alpha_k \leq 0$ nebo $\|s_k + \alpha_k p_k\| > \Delta$, položíme $M := 0$

Krok 4 Položíme $g_{k+1} := g_k + \alpha_k u_k$, $\sigma_{k+1} := g_{k+1}^T g_{k+1}$, $\beta_k := \sigma_{k+1}/\sigma_k$, vypočteme číslo δ_{k+1} podle věty 142, tedy $\delta_{k+1} := \sqrt{\beta_k}/|\alpha_k|$ a přejdeme na krok 6.

Krok 5 Položíme $M := 0$, $\gamma_k := q_k^T B q_k$ a vypočteme číslo δ_{k+1} a vektor $v_{k+1} := \delta_{k+1} q_{k+1}$ podle definice 44, tedy $\delta_{k+1} := \|v_{k+1}\|$, kde $v_{k+1} := B q_k - \gamma_k q_k$, pokud $k = 1$, nebo $v_{k+1} := B q_k - \gamma_k q_k - \delta_k q_{k-1}$, pokud $k > 1$.

Krok 6 Jestliže $M = 1$ a $k \leq m$, položíme $s_{k+1} := s_k + \alpha_k p_k$ a pokud $\|g_{k+1}\| \leq \omega\|g\|$, ukončíme výpočet. Jestliže $M = 0$ nebo $k > m$, položíme $s_{i+1} := Q_i z_i$, kde vektor z_i je řešením úlohy (597) a pokud $\delta_{k+1}|e_k^T z_k| \leq \omega\|g\|$ nebo $k > m$, ukončíme výpočet.

Krok 7 Jestliže $L = 0$, položíme $q_{k+1} := v_{k+1}/\delta_{k+1}$. Jestliže $L = 1$, položíme $\varepsilon_{k+1} := -\varepsilon_k \mathrm{sgn}(\alpha_k)$, $q_{k+1} := \varepsilon_{k+1} g_{k+1}/\|g_{k+1}\|$ a $p_{k+1} := -g_{k+1} + \beta_k p_k$. Zvětšíme k o jednotku a přejdeme na krok 2.

V metodách používajících symetrický Lanczosův proces není účelné používat předpodmínění, neboť se tím mění původní omezení $\|s_i\| \leq \Delta$ na $\|s_i\|_C = \sqrt{s_i^T C s_i} \leq \Delta$ (výjimku tvoří případy, kdy je z nějakých důvodů třeba řešit úlohu s omezením $\|s_i\|_C \leq \Delta$). Předpodmiňovač C se obvykle odvozuje od matice B, takže může být špatně podmíněný a navíc se mění v každé iteraci.

6.5 Posunuté nepřesné metody s lokálně omezeným krokem

V tomto oddílu ukážeme jiný způsob použití symetrického Lanczosova procesu. Symetrický Lanczosův proces použijeme k určení aproximace λ Lagrangeova multiplikátoru λ^* vystupujícího ve větě 124 a směrový vektor $s = s(\lambda)$ budeme hledat řešením úlohy

$$s(\lambda) = \arg \min_{\|s\| \leq \Delta} Q_\lambda(s), \qquad Q_\lambda(s) = \frac{1}{2} s^T (B + \lambda I) s + g^T s. \tag{598}$$

To znamená, že budeme metodu sdružených gradientů aplikovat na soustavu rovnic s maticí $B + \lambda I$. Aby získaný směrový vektor splňoval podmínku (T1b), potřebujeme aby $\lambda = 0$, pokud úloha (547) má řešení takové, že $\|s^*\| < \Delta$. To je zaručeno, pokud je splněna nerovnost $\lambda \leq \lambda^*$, kterou nyní dokážeme. Budeme přitom používat označení

$$\mathcal{K}_k(\lambda) = \mathrm{span}\{g, (B + \lambda I)g, \ldots, (B + \lambda I)^{k-1} g\}$$

pro Krylovův podprostor dimenze k definovaný maticí $B + \lambda I$ a vektorem g, a $Z_k \in R^{n \times k}$ pro matici jejíž sloupce tvoří ortonormální bázi v $\mathcal{K}_k(\lambda)$.

Poznámka 234. Indukcí ukážeme, že pro libovolné číslo $\lambda \in R$ platí $\mathcal{K}_k(\lambda) = \mathcal{K}_k$, kde $\mathcal{K}_k = \mathcal{K}_k(0)$. Pro $k = 1$ je to zřejmé, neboť $\mathcal{K}_k(\lambda) = \mathrm{span}\{g\} = \mathcal{K}_k$. Předpokládejme, že uvažovaná rovnost platí pro nějaký index $1 \leq k < n$. Pak

$$(B + \lambda I)^k g = (B + \lambda I)(B + \lambda I)^{k-1} g = (B + \lambda I)v = Bv + \lambda v,$$

kde $v \in \mathcal{K}_k(\lambda) = \mathcal{K}_k$. Jelikož $\lambda v \in \mathcal{K}_k$ a $Bv \in \mathcal{K}_{k+1}$, platí $(B + \lambda I)^k g \in \mathcal{K}_{k+1}$, takže $\mathcal{K}_{k+1}(\lambda) \subset \mathcal{K}_{k+1}$. Aplikujeme-li stejný postup na matice $B + \lambda I$ a $B = (B + \lambda I) - \lambda I$, dostaneme opačnou inkluzi.

Lemma 66. *Nechť $Z_k^T B Z_k + \lambda_1 I$, $Z_k^T B Z_k + \lambda_2 I$ jsou symetrické pozitivně definitní matice a nechť*

$$s_k(\lambda_1) = \arg \min_{s \in \mathcal{K}_k} Q_{\lambda_1}(s), \qquad s_k(\lambda_2) = \arg \min_{s \in \mathcal{K}_k} Q_{\lambda_2}(s),$$

kde $Q_\lambda(s)$ je funkce definovaná v (598). Pak

$$\lambda_2 \leq \lambda_1 \quad \Longleftrightarrow \quad \|s_k(\lambda_2)\| \geq \|s_k(\lambda_1)\|.$$

Důkaz (a) Nechť B_1 a B_2 jsou dvě symetrické pozitivně definitní matice. Pak ze vztahů

$$B_1 - B_2 = B_2^{1/2}(B_2^{-1/2} B_1 B_2^{-1/2} - I) B_2^{1/2}, \quad B_2^{-1} - B_1^{-1} = B_1^{-\frac{1}{2}}(B_1^{\frac{1}{2}} B_2^{-1} B_1^{1/2} - I) B_1^{-1/2}$$

a z toho, že matice $B_2^{-1/2} B_1 B_2^{-1/2}$ a $B_1^{1/2} B_2^{-1} B_1^{1/2}$ mají stejná vlastní čísla, plyne

$$\begin{aligned} B_1 - B_2 \succeq 0 \quad &\Longleftrightarrow \quad B_2^{-1} - B_1^{-1} \succeq 0, \\ B_1 - B_2 \succ 0 \quad &\Longleftrightarrow \quad B_2^{-1} - B_1^{-1} \succ 0. \end{aligned}$$

(b) Je-li matice $Z_k^T B Z_k + \lambda I$ pozitivně definitní, můžeme minimum $s_k(\lambda)$ funkce $Q_\lambda(s)$ na $\mathcal{K}_k$ vyjádřit ve tvaru

$$s_k(\lambda) = -Z_k (Z_k^T (B + \lambda I) Z_k)^{-1} Z_k^T g. \tag{599}$$

Pokud $s \in \mathcal{K}_k$, platí $s = Z_k\tilde{s}$, kde $\tilde{s} \in R^k$, takže

$$Q_\lambda(s) = \frac{1}{2}s^T(B+\lambda I)s + g^T s = \frac{1}{2}\tilde{s}^T Z_k^T(B+\lambda I)Z_k\tilde{s} + g^T Z_k\tilde{s} \triangleq \tilde{Q}_\lambda(\tilde{s}).$$

Minimum $\tilde{s}_k(\lambda)$ funkce $\tilde{Q}_\lambda(\tilde{s})$ na R^k určíme podle vzorce $\tilde{s}_k(\lambda) = -(Z_k^T(B+\lambda I)Z_k)^{-1}Z_k^T g$, což po dosazení do $s_k(\lambda) = Z_k\tilde{s}_k(\lambda)$ dává (599).

(c) Použijeme-li (599), dostaneme

$$\|s_k(\lambda)\|^2 = g^T Z_k(Z_k^T(B+\lambda I)Z_k)^{-2}Z_k^T g = g^T Z_k(Z_k^T BZ_k + \lambda I)^{-2}Z_k^T g.$$

Platí tedy

$$\|s_k(\lambda_2)\|^2 - \|s_k(\lambda_1)\|^2 = g^T Z_k\left((Z_k^T BZ_k + \lambda_2 I)^{-2} - (Z_k^T BZ_k + \lambda_1 I)^{-2}\right)Z_k^T g.$$

Označíme-li $\tilde{B}_2 = (Z_k^T BZ_k + \lambda_2 I)$ a předpokláme-li, že $\lambda_2 \le \lambda_1$, můžeme psát

$$(Z_k^T BZ_k + \lambda_1 I)^2 - (Z_k^T BZ_k + \lambda_2 I)^2 = (\tilde{B}_2 + (\lambda_1-\lambda_2)I)^2 - \tilde{B}_2^2 = 2(\lambda_1-\lambda_2)\tilde{B}_2 + (\lambda_1-\lambda_2)^2 I \succeq 0,$$

což spolu s první ekvivalencí v (a) dává

$$(Z_k^T BZ_k + \lambda_2 I)^{-2} - (Z_k^T BZ_k + \lambda_1 I)^{-2} \succeq 0,$$

neboli $\|s_k(\lambda_2)\|^2 - \|s_k(\lambda_1)\|^2 \ge 0$. Použijeme-li druhou ekvivalenci v (a), dostaneme stejným postupem $\lambda_2 < \lambda_1 \Rightarrow \|s_k(\lambda_2)\|^2 > \|s_k(\lambda_1)\|^2$. Protože nezáleží na indexování, můžeme psát $\lambda_1 < \lambda_2 \Rightarrow \|s_k(\lambda_1)\|^2 > \|s_k(\lambda_2)\|^2$, což dává $\|s_k(\lambda_2)\| \ge \|s_k(\lambda_1)\| \Rightarrow \lambda_2 \le \lambda_1$. □

Věta 145. *Nechť pro libovolný index $1 \le k \le n$ je vektor s_k řešením úlohy*

$$s_k = \arg\min_{s\in\mathcal{K}_k, \|s\|\le\Delta} Q(s), \qquad Q(s) = \frac{1}{2}\,s^T Bs + g^T s \tag{600}$$

a číslo λ_k odpovídajícím Lagrangeovým multiplikátorem. Pak pro $1 \le i \le j \le n$ platí $\lambda_i \le \lambda_j$. Speciálně $\lambda_k \le \lambda^*$ pro libovolný index $1 \le k \le n$.

Důkaz Podle věty 124 je vektor s_k řešením úlohy (600) právě tehdy, když $\|s_k\| = \|Z_k\tilde{s}_k\| \le \Delta$, kde $Z_k^T(B+\lambda_k I)Z_k\tilde{s}_k = -Z_k^T g$, $Z_k^T(B+\lambda_k I)Z_k \succeq 0$, $\lambda_k \ge 0$ a $\lambda_k(\Delta - \|s_k\|) = 0$. Toto řešení je nepodmíněným minimem (stejné řešení dostaneme i po odstranění omezení $s_k \le \Delta$) právě tehdy, když $\lambda_k = 0$.

(a) Nechť $\lambda_j = 0$ (což znamená, že vektor s_j je nepodmíněným minimem funkce $Q(s)$ na $\mathcal{K}_j$) a $i \le j$. Pak podle věty 69 platí $\|s_i\| \le \|s_j\| \le \Delta$ a vektor s_i je nepodmíněným minimem funkce $Q(s)$ na $\mathcal{K}_i$, což dává $\lambda_i = 0$. Jestliže $\lambda_j > 0$ a $\lambda_i = 0$, není co dokazovat. Zbývá tedy případ, kdy $\lambda_j > 0$ a $\lambda_i > 0$, takže $\|s_j\| = \|s_i\| = \Delta$, což budeme v dalších krocích předpokládat.

(b) Je-li matice $Z_i^T(B+\lambda_i I)Z_i$ singulární, musí platit $\lambda_i \le \lambda_j$, neboť v opačném případě by existoval nenulový vektor $v \in \mathcal{K}_i$ takový, že $v^T(B+\lambda_j I)v = v^T(B+\lambda_i I)v + (\lambda_j - \lambda_i)\|v\| = (\lambda_j - \lambda_i)\|v\| < 0$. To však není možné, neboť $\mathcal{K}_i \subset \mathcal{K}_j$ a podle věty 124 platí $Z_j^T(B+\lambda_j I)Z_j \succeq 0$.

(c) Předpokládejme, že $Z_i^T(B+\lambda_i I)Z_i \succ 0$ a $Z_j^T(B+\lambda_j I)Z_j \succ 0$. Jelikož podle poznámky 234 platí $\mathcal{K}_i(\lambda_i) = \mathcal{K}_i$ a $\mathcal{K}_j(\lambda_j) = \mathcal{K}_j$, jsou vektory s_i a s_j řešením nepodmíněných úloh

$$s_i = \arg\min_{s\in\mathcal{K}_i} Q_{\lambda_i}(s), \qquad s_j = \arg\min_{s\in\mathcal{K}_j} Q_{\lambda_j}(s).$$

Předpokládejme, že $\lambda_i > \lambda_j$, takže $Z_j^T(B+\lambda_i I)Z_j \succ 0$. Nechť

$$s_j(\lambda_i) = \arg\min_{s\in\mathcal{K}_j} Q_{\lambda_i}(s).$$

Pak použitím věty 69 dostaneme $\|s_j(\lambda_i)\| \geq \|s_i\| = \Delta = \|s_j\|$, neboli $\|s_j(\lambda_j)\| \leq \|s_j(\lambda_i)\|$, takže podle lemmatu 66 platí $\lambda_i \leq \lambda_j$. To je však ve sporu s předpokladem, že $\lambda_i > \lambda_j$.

(d) Předpokládejme nakonec, že matice $Z_j^T(B + \lambda_j I)Z_j$ je singulární. V tomto případě podle věty 133 platí $\|s_j(\lambda_j + \varepsilon)\| \leq \Delta$ pro libovolné číslo $\varepsilon > 0$. Jelikož matice $Z_j^T(B + (\lambda_j + \varepsilon)I)Z_j$ je pozitivně definitní, je i matice $Z_i^T(B + (\lambda_j + \varepsilon)I)Z_i$ pozitivně definitní a podle věty 69 platí $\|s_i(\lambda_j + \varepsilon)\| \leq \|s_j(\lambda_j + \varepsilon)\| \leq \Delta$. Protože $\|s_i\| = \Delta$, můžeme psát $\|s_i(\lambda_j + \varepsilon)\| \leq \|s_i(\lambda_i)\|$ a s použitím lemmatu 66 dostaneme $\lambda_i \leq \lambda_j + \varepsilon$. Jelikož číslo ε je libovolné, platí $\lambda_i \leq \lambda_j$. □

Nyní se vrátíme k problému (598). Položíme-li $\lambda = \lambda_k$ pro nějaký index $k \leq n$, věta 145 zaručuje, že $0 \leq \lambda = \lambda_k \leq \lambda_n = \lambda^*$. Důsledkem této nerovnosti je, že $\lambda = 0$, pokud $\lambda^* = 0$. Je-li matice B pozitivně definitní a $\lambda > 0$, platí $\Delta \leq \|(B + \lambda I)^{-1}g\| < \|B^{-1}g\|$ podle věty 69, takže nepodmíněné minimum funkce $Q_\lambda(s)$ je blíže k hranici oblasti určené omezením $\|s\| \leq \Delta$ než Newtonův krok $d_N = B^{-1}g$ a můžeme očekávat, že $s(\lambda)$ je blíže k optimálnímu lokálně omezenému kroku než s_N. Navíc, jelikož $\lambda > 0$, je matice $B + \lambda I$ lépe podmíněná a můžeme očekávat, že posunutá nepřesná metoda s lokálně omezeným krokem bude konvergovat rychleji než standardní metoda (s $\lambda = 0$). Posunutá nepřesná metoda s lokálně omezeným krokem se skládá ze tří základních kroků.

Algoritmus 15. Data $C \succ 0$, $0 < \omega \leq \overline{\omega} < 1$, $\Delta > 0$, $m \ll n$.

Krok 1: Použijeme m kroků nepředpodmíněného symetricého Lanczosova procesu a získáme tak symetrickou tridiagonální matici $T = T_k = Z_k^T B Z_k$.

Krok 2: řešíme úlohu

$$z_k = \arg \min_{z \in R^k, \|s\| \leq \Delta} \left(\frac{1}{2} z^T T_k z + \delta_1 e_1^T z \right)$$

metodou pro výpočet optimálního lokálně omezeného kroku (oddíl 6.1). Získáme přitom Lagrangeův multiplikátor λ.

Krok 3: Aplikujeme nepřesnou metodu s lokálně omezeným krokem na úlohu (598) a získáme tak vektor s, který je aproximací vektoru $s(\lambda)$.

Obvykle volíme $m = 5$. Pokud $m = 1$ dostaneme nepřesnou metodu s lokálně omezeným krokem popsanou v oddílu 6.3.

6.6 Numerické porovnání jednotlivých algoritmů

V tomto oddílu ukážeme, jak jednotlivé algoritmy pro výpočet lokálně omezeného kroku ovlivňují účinnost diferenční verze Newtonovy metody popsané v oddílu 9.6 (realizované jako metody s lokálně omezeným krokem). K testování bylo použito 22 úloh s 1000 a 5000 proměnnými (TEST14 z oddílu 1.5). V následující tabulce jsou uvedeny výsledky testů odpovídající těmto metodám:

- A10 - metoda s optimálním lokálně omezeným krokem (algoritmus 10),
- A11 - metoda psí nohy (algoritmus 11),
- A12 - nepřesná metoda s lokálně omezeným krokem (algoritmus 12 s $C = I$),
- A12* - předpodmíněná nepřesná metoda s lokálně omezeným krokem (algoritmus 12 s $C \neq I$),
- A13 - vícekroková metoda psí nohy (algoritmus 13 s $m = 5$),
- A14 - metoda založená na použití symetrické Lanczosovy metody (algoritmus 14),
- A15* - předpodmíněná posunutá nepřesná metoda s lokálně omezeným krokem (algoritmus 15 s $C \neq I$).

Algoritmy A12* a A15* používají předpodmínění založené na neúplném Choleského rozkladu. V tabulce jsou uvedeny celkové počty iterací `NIT`, funkčních hodnot NFV, gradientů NFG, vnitřních terací `NCG` a celkový čas výpočtu.

N	Metoda	NIT	NFV	NFG	NCG	čas
1000	A10	1911	1952	8724	-	1.27
	A11	2286	2425	10779	-	1.26
	A12	3475	4021	17242	55930	2.49
	A12*	2611	2807	12836	913	1.40
	A13	2132	2232	9998	11059	1.34
	A14	3283	3688	16250	60467	2.51
	A15*	2007	2077	9239	11467	1.27
5000	A10	8177	8273	34781	-	21.31
	A11	9666	10146	42283	-	19.17
	A12	16932	19137	84417	341925	62.49
	A12*	11055	11760	53057	3740	28.62
	A13	8913	9244	38846	47234	21.52
	A14	14917	16664	72972	360790	62.30
	A15*	8320	8454	35629	46493	20.12

Tabulka 5a: TEST14 – 22 úloh

Úlohy obsažené ve sbírce TEST14 jsou jednodušší v tom smyslu, že čísla podmíněnosti příslušných Hessových matic nejsou příliš vysoká. V následující tabulce jsou tytéž metody testovány pomocí 74 hůře podmíněných úloh obsažených ve sbírce TEST25 (8 úloh ze sbírky TEST25 bylo vynecháno, protože je některá z testovaných metod nevyřešila). V posledním sloupci tabulky je uveden počet úloh (z celkového počtu 82), které daná metoda vyřešila, což vyjadřuje robustnost této metody.

N	Metoda	NIT	NFV	NFG	NCG	čas	počet
1000	A10	65591	6835	34298	-	6.05	81
	A11	87136	9557	55002	-	6.47	79
	A12	10456	11198	65232	1751061	56.11	75
	A12*	9746	10379	59695	658361	28.76	77
	A13	8717	9399	60611	43681	7.31	79
	A14	10593	11291	68784	2186546	74.25	75
	A15*	8728	9285	54650	777410	33.18	76

Tabulka 5b: TEST25 – 74 úloh

Výsledky testů ukazují, že pro velmi řídké úlohy je vhodné používat přímé metody (zejména algoritmus 10) a že iterační metody vyžadují kvalitní předpodmínění.

Na závěr uvedeme porovnání některých realizací metod pro středně velké husté úlohy. Použijene k tomu sbírku TEST01, která obsahuje analytická vyjádření prvků Hessovy matice, takže je možné testovat různé verze Newtonovy metody. V následující tabulce jsou uvedeny výsledky testů odpovídající těmto metodám:

`LSVM` - metoda BFGS s řízeným škálováním,
`TRVM-xx` - metoda BFGS s řízeným škálováním,
`LSMN` - Newtonova metoda s korekcí zajišťující spádovost,
`TRMN-xx` - Newtonova metoda s analytickým výpočtem Hessovy matice,

Metody, jejichž označení začíná písmeny TR jsou realizovány jako metody s lokálně omezeným krokem a písmena LS označují metody spádových směrů. Přitom číslice xx udávají číslo použitého algoritmu.

Pro srovnání jsou též uvedeny výsledky získané metodou sdružených gradientů `CG`, což je algoritmus 5, kde používáme volbu `HST+` (vzorce (172), 173), (174)) a proceduru pro výběr délky kroku převzatou z programu `CG-DESCENT`).

Metoda	NIT	NFV	NFG	čas
LSVM	2547	2971	2971	0.59
TRVM-11	3075	3409	3090	1.34
LSMN	1765	2697	2697	7.02
TRMN-10	571	608	586	3.26
TRMN-11	730	804	745	5.61
CG	7586	14843	8089	0.71

Tabulka 6: TEST01 – 15 úloh

Z výsledků uvedených v této tabulce lze vyvodit několik závěrů:

- Metody s proměnnou metrikou není vhodné realizovat jak metody s lokálně omezeným krokem.
- Newtonovu metodu není vhodné realizovat jako metodu spádových směrů.
- Použití přesné Hessovy matice zvyšuje rychlost konvergence (snižuje se počet iterací a použitých funkčních hodnot). Pro 200 proměnných se však již projevuje nutnost řešení soustavy rovnic (vyžaduje to $O(n^3)$ aritmetických operací).
- Pro úlohy s 200 proměnnými dominují metody s proměnnou metrikou. Newtonova metoda konverguje rychleji, ale spotřebuje více strojového času. Pokud neřešíme příliš obtížné úlohy roste efektivita metody sdružených gradientů s počtem proměnných (neboť je třeba pouze $O(n)$ aritmetických iterací na iteraci).

6.7 Iterační metody pro řešení lineárních soustav se symetrickou indefinitní maticí

V oddílu 2.7 jsme ukázali, jak lze získat spádové směry a směry se zápornou křivostí pomocí maticových rozkladů pro symetrické indefinitní matice. Směry se zápornou křivostí lze také určit pomocí iteračních metod pro symetrické indefinitní matice. V tomto oddílu popíšeme dvě iterační metody pro řešení soustavy rovnic $Bs + g = 0$ se symetrickou indefinitní maticí B. První z nich je modifikací metody sdružených gradientů popsané v oddílu 3.8 a druhá vychází se symetrické Lanczosovy metody uvedené v oddílu 6.4.

Metodu sdružených gradientů můžeme formálně použít i v případě, že matice B není pozitivně definitní, pokud platí $p_i^T q_i = p_i^T B p_i \neq 0$, $1 \leq i \leq n$ (definice 35 s $C = I$). V tomto případě není bod s_{i+1} minimem nekonvexní kvadratické funkce (237), ale jejím stacionárním bodem, na podprostoru definovaném vektory p_j, $1 \leq j \leq i$. Nastanou však potíže, pokud $p_i^T B p_i \approx 0$, a metoda zcela selže, pokud $p_i^T B p_i = 0$ (dělení nulou). Metoda sdružených gradientů je založena na tom, že se hledá stacionární bod s_{i+1} funkce (237) na přímce $s_i + \mathcal{L}(p_i)$, takže $s_{i+1} = s_i + \alpha_i p_i$, kde koeficient α_i se vybírá tak, aby platilo $p_i^T g_{i+1} = 0$, neboli $\alpha_i = -p_i^T g_i / p_i^T B p_i$. Základní planární metoda sdružených gradientů, kterou nyní popíšeme, určuje (v případě, že $p_i^T B p_i = 0$) nový vektor p_{i+1} takový, že $p_i^T B p_{i+1} \neq 0$ a $p_{i+1}^T g_i = 0$, a hledá stacionární bod s_{i+2} funkce (237) v rovině $s_i + \mathcal{L}(p_i, p_{i+1})$, takže $s_{i+2} = s_i + \alpha_i p_i + \alpha_{i+1} p_{i+1}$, kde koeficienty α_i, α_{i+1} se vybírají tak, aby platilo

$$p_i^T g_{i+2} = 0, \qquad p_{i+1}^T g_{i+2} = 0, \tag{601}$$

neboli

$$\begin{aligned} \alpha_{i+1}\, p_i^T B p_{i+1} &= -p_i^T g_i, \\ \alpha_i\, p_{i+1}^T B p_i + \alpha_{i+1}\, p_{i+1}^T B p_{i+1} &= 0 \end{aligned}$$

(neboť $g_{i+2} = g_i + \alpha_i B p_i + \alpha_{i+1} B p_{i+1}$, $p_i^T B p_i = 0$ a $p_{i+1}^T g_i = 0$), což dává

$$\alpha_{i+1} = -\frac{p_i^T g_i}{p_i^T B p_{i+1}}, \qquad \alpha_i = -\frac{p_{i+1}^T B p_{i+1}}{p_i^T B p_{i+1}}\, \alpha_{i+1}.$$

Základní planární metoda sdružených gradientů používá jednoduché kroky standardní metody sdružených gradientů, pokud $p_i^T B p_i \neq 0$, a dvojité planární kroky, pokud $p_i^T B p_i = 0$. Abychom zjednodušili příslušné úvahy, označíme I_1 množinu indexů jednoduchých kroků (takže z $i \in I_1$ plyne $p_i^T B p_i \neq 0$) a I_2 množinu prvních indexů planárních kroků (takže z $i \in I_2$ plyne $p_i^T B p_i = 0$ a $p_i^T B p_{i+1} \neq 0$). Poznamenejme, že $i \in I_2$ implikuje $i+1 \notin I_1$, $i+1 \notin I_2$, takže $I_1 \cup I_2 \neq \{1, \ldots, n\}$.

Definice 46. *Nechť $B \in R^{n\times n}$ je symetrická regulární matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \qquad g_1 = g, \qquad p_1 = -g$$

a buď

$$\begin{aligned} q_i &= B p_i, & \alpha_i &= -p_i^T g_i / p_i^T q_i, \\ s_{i+1} &= s_i + \alpha_i p_i, & g_{i+1} &= g_i + \alpha_i q_i, \\ \beta_i &= q_i^T g_{i+1} / p_i^T q_i, & p_{i+1} &= -g_{i+1} + \beta_i p_i, \end{aligned}$$

pokud $i \in I_1$ (jednoduchý krok), nebo

$$\begin{aligned} q_i &= B p_i, & \gamma_i &= \|p_i\| / \|q_i\|, \\ p_{i+1} &= \gamma_i q_i, & q_{i+1} &= B p_{i+1}, \\ \alpha_{i+1} &= -p_i^T g_i / p_{i+1}^T q_i & \alpha_i &= \alpha_{i+1}\, p_{i+1}^T q_{i+1} / p_{i+1}^T q_i, \\ s_{i+1} &= s_i + \alpha_i p_i, & g_{i+1} &= g_i + \alpha_i q_i, \\ s_{i+2} &= s_{i+1} + \alpha_{i+1} p_{i+1}, & g_{i+2} &= g_{i+1} + \alpha_{i+1} q_{i+1}, \\ \beta_i &= q_{i+1}^T g_{i+2} / p_{i+1}^T q_i, & p_{i+2} &= -g_{i+2} + \beta_i p_i, \end{aligned}$$

pokud $i \in I_2$ (planární krok), nazveme základní planární metodou sdružených gradientů pro řešení soustavy rovnic $Bs + g = 0$.

Poznámka 235. Planární krok obsahuje dvě násobení matice vektorem, ale protože tento krok odpovídá dvěma jednoduchým krokům, je základní planární metoda sdružených gradientů co do celkového počet operací srovnatelná se standardní metodou sdružených gradientů.

Poznámka 236. V planárním kroku se používá škálovací parametr $\gamma_i > 0$. Hodnotu $\gamma_i = \|p_i\|/\|q_i\|$ volíme proto, aby platilo $\|p_{i+1}\| = \|p_i\|$. Lze volit i jinou hodnotu, například $\gamma_i = 1$. Pak platí $\|p_{i+1}\| = \|Bp_i\|$.

Poznámka 237. Planární krok se používá pouze tehdy, když $p_i^T B p_i = 0$. Není tudíž ošetřen případ, kdy $p_i^T B p_i \approx 0$. Algoritmus základní planární metody sdružených gradientů lze upravit tak, že se planární krok použije i tehdy, když $0 < |p_i^T B p_i| < \underline{c}\|p_i\|^2$, kde $\underline{c} > 0$ je (malá) předepsaná tolerance. V tom případě však nezískáme přesné řešení soustavy rovnic $Bs + g = 0$ po konečném počtu kroků. Proto byly vyvinuty složitější planární metody sdružených gradientů (definice 47). Základní planární metodu sdružených gradientů zde uvádíme proto, abychom ukázali její vlastnosti. Podobné vlastnosti mají i ostatní planární metody, jejichž vyšetřování je však formálně složitější.

Věta 146. *Základní planární metoda sdružených gradientů (definice 46) najde řešení soustavy rovnic $Bs + g = 0$, kde B je symetrická regulární matice, po nejvýše n krocích (planární krok počítáme za dva jednoduché kroky).*

Důkaz Předpokládejme, že $g_i \neq 0$, $1 \leq i \leq n$ (není-li tato podmínka splněna, platí $g_{m+1} = 0$ pro nějaký index $m < n$, takže vektor s_{m+1} je řešením soustavy rovnic $Bs + g = 0$). Dokážeme indukcí, že pro libovolný index $1 \leq i \leq n$ je $p_i \neq 0$, $p_i^T g_i < 0$,

$$g_i \in \mathcal{L}(P_i), \quad Bp_{i-1} \in \mathcal{L}(P_i), \tag{602}$$

kde $P_i = [p_1, \ldots, p_i]$, přičemž pro $i \in I_1$ platí

$$p_j^T g_{i+1} = 0, \qquad 1 \leq j \leq i, \tag{603}$$

$$g_j^T g_{i+1} = 0, \qquad 1 \leq j \leq i, \tag{604}$$

$$p_j^T B p_i = 0, \qquad 1 \leq j < i \tag{605}$$

a pro $i \in I_2$ platí

$$p_j^T g_{i+2} = 0, \qquad 1 \leq j \leq i+1, \tag{606}$$

$$g_j^T g_{i+2} = 0, \qquad 1 \leq j \leq i+1, \tag{607}$$

$$p_j^T B p_i = 0, \qquad 1 \leq j < i+1 \tag{608}$$

(místo (608) stačí dokázat (605), neboť z $i \in I_2$ plyne $p_i^T B p_i = 0$). Pro $i = 1$ je $p_1 = -g_1 \neq 0$, $g_1^T p_1 = -g_1^T g_1 < 0$ a $g_1 = -p_1 \in \mathcal{L}(P_1)$. Nechť pro nějaký index $1 \leq i < n$ platí $p_j \neq 0$, $p_j^T g_j < 0$, $g_j \in \mathcal{L}(P_j)$, $Bp_{j-1} \in \mathcal{L}(P_j)$, pro $1 \leq j \leq i$, a $p_j^T g_i = 0$, $g_j^T g_i = 0$, $p_j^T B p_i = 0$, pro $1 \leq j < i$ (indukční předpoklad).

(a) Ukážeme nejprve, že iterační krok v definici 46 je dobře definován, čili že pro $i \in I_1$ platí $\alpha_i \neq 0$ a pro $i \in I_2$ platí $p_i^T B p_{i+1} \neq 0$, $p_{i+1}^T g_i = 0$ a $\alpha_{i+1} \neq 0$. Jelikož $p_i^T g_i < 0$, můžeme pro $i \in I_1$ psát $\alpha_i = -p_i^T g_i / p_i^T B p_i \neq 0$. Jelikož $p_{i+1} = \gamma_i B p_i$ a $\gamma_i \neq 0$, můžeme pro $i \in I_2$ psát $p_i^T B p_{i+1} = \gamma_i p_i^T B^2 p_i \neq 0$ (neboť $p_i \neq 0$ a matice B je symetrická a regulární). Podle definice 46 platí $g_i = -p_i + \beta_{i-1} p_{i-1}$, je-li předchozí krok jednoduchý, nebo $g_i = -p_i + \beta_{i-2} p_{i-2}$, je-li předchozí krok planární. V obou případech s použitím (605) dostaneme $p_{i+1}^T g_i = \gamma_i p_i^T B g_i = -\gamma_i p_i^T B p_i = 0$, odkud plynou vzorce pro α_i a α_{i+1} v definici 46, přičemž $\alpha_{i+1} = -p_i^T g_i / p_i^T B p_{i+1} \neq 0$.

(b) Důkaz incidencí (602). Pokud $i \in I_1$, platí $g_{i+1} = -p_{i+1} + \beta_i p_i$ a $g_{i+1} = g_i + \alpha_i B p_i$, takže $g_{i+1} \in \mathcal{L}(P_{i+1})$ a $Bp_i = (g_{i+1} - g_i)/\alpha_i \in \mathcal{L}(P_{i+1})$. Pokud $i \in I_2$, můžeme psát $Bp_i = (1/\gamma_i) p_{i+1}$ a $g_{i+1} = g_i + \alpha_i B p_i$, takže $Bp_i \in \mathcal{L}(P_{i+1})$ a $g_{i+1} \in \mathcal{L}(P_{i+1})$. Dále platí $g_{i+2} = -p_{i+2} + \beta_i p_i$ a $g_{i+2} = g_{i+1} + \alpha_{i+1} B p_{i+1}$, takže $g_{i+2} \in \mathcal{L}(P_{i+2})$ a $Bp_{i+1} = (g_{i+2} - g_{i+1})/\alpha_{i+1} \in \mathcal{L}(P_{i+2})$.

(c) Důkaz vztahů (603) a (606). Pokud $i \in I_1$, můžeme pro $1 \leq j < i$ psát $p_j^T g_{i+1} = p_j^T (g_i + \alpha_i B p_i) = 0$ (neboť podle indukčního předpokladu pro $1 \leq j < i$ platí $p_j^T g_i = 0$ a $p_j^T B p_i = 0$). Jelikož koeficient α_i vybíráme tak, aby platilo $p_i^T g_{i+1} = 0$, dostaneme (603). Nechť $i \in I_2$. Pak podle (608) pro $1 \leq j < i$ platí $p_j^T B p_{i+1} = \gamma_i (B p_j)^T B p_i = 0$ (neboť podle indukčního předpokladu $Bp_j \in \mathcal{L}(P_{j+1})$). Můžeme tedy psát

$$p_j^T g_{i+2} = p_j^T (g_i + \alpha_i B p_i + \alpha_{i+1} B p_{i+1}) = 0$$

pro $1 \leq j < i$ (neboť podle indukčního předpokladu pro $1 \leq j < i$ platí $p_j^T g_i = 0$ a $p_j^T B p_i = 0$). Jelikož koeficienty α_i a α_{i+1} vybíráme tak, aby platilo $p_i^T g_{i+2} = 0$ a $p_{i+1}^T g_{i+2} = 0$, dostaneme (606).

(d) Důkaz vztahů (604) a (607). Jelikož podle (b) pro $1 \leq j \leq i+1$ platí $g_j \in \mathcal{L}(P_j)$, plyne (604) z (603) a (607) z (606).

(e) Důkaz vztahů (605) a (608). Pro $i \in I_1$ podle definice 46 a podle (601), platí

$$g_{i+1}^T p_{i+1} = -g_{i+1}^T g_{i+1} + \beta_i g_{i+1}^T p_i = -g_{i+1}^T g_{i+1} < 0$$

(neboť $g_{i+1} \neq 0$), takže $p_{i+1} \neq 0$. Použijeme-li (604), (605) spolu s definicí 46, dostaneme

$$p_j^T B p_{i+1} = -p_j^T B g_{i+1} + \beta_i p_j^T B p_i = (1/\alpha_j)(g_j - g_{j+1})^T g_{i+1} = 0$$

pro $1 \le j < i$ a (601) spolu s definicí 46 dává

$$p_i^T B p_{i+1} = -p_i^T B g_{i+1} + \beta_i p_i^T B p_i = -p_i^T B g_{i+1} + \frac{p_i^T B g_{i+1}}{p_i^T B p_i} p_i^T B p_i = 0.$$

Nechť $i \in I_2$. Pak podle definice 46 a podle (601), platí

$$g_{i+2}^T p_{i+2} = -g_{i+2}^T g_{i+2} + \beta_i g_{i+2}^T p_i = -g_{i+2}^T g_{i+2} < 0$$

(neboť $g_{i+2} \neq 0$), takže $p_{i+2} \neq 0$. Použijeme-li (607), (608) spolu s definicí 46, dostaneme

$$p_j^T B p_{i+2} = -p_j^T B g_{i+2} + \beta_i p_j^T B p_i = (1/\alpha_j)(g_j - g_{j+1})^T g_{i+2} = 0$$

pro $1 \le j < i$. Dále z $p_i^T B p_i = 0$, $p_{i+1} = \gamma_i B p_i$ a (601) plyne

$$p_i^T B p_{i+2} = -p_i^T B g_{i+2} + \beta_i p_i^T B p_i = -\frac{1}{\gamma_i} p_{i+1}^T g_{i+2} = 0,$$

a (601) spolu s definicí 46 dává

$$p_{i+1}^T B p_{i+2} = -p_{i+1}^T B g_{i+2} + \beta_i p_{i+1}^T B p_i = -p_{i+1}^T B g_{i+2} + \frac{p_{i+1}^T B g_{i+2}}{p_{i+1}^T B p_i} p_{i+1}^T B p_i = 0.$$

(f) Podle (605) a (608) je matice $P_n^T B P_n$ blokově diagonální. Pro $i \in I_1$ je prvek $p_i^T B p_i$ nenulový a pro $i \in I_2$ platí

$$\det [p_i, p_{i+1}]^T B [p_i, p_{i+1}] = \det \begin{bmatrix} 0, & p_i^T B p_{i+1} \\ p_{i+1}^T B p_i, & p_{i+1}^T B p_{i+1} \end{bmatrix} = -(p_i^T B p_{i+1})^2 < 0,$$

takže matice $P_n^T B P_n$ je regulární a jelikož B je regulární, jsou vektory $p_1, \ldots, p_n$ lineárně nezávislé. Odtud a z (603), (606) plyne, že $g_{n+1} = 0$, takže vektor s_{n+1} je řešením soustavy rovnic $Bs + g = 0$. □

Důsledek 21. *Pro $i \in I_1 \cup I_2$, platí $p_i^T g_i = p_i^T g$.*

Důkaz Podle definice 46 platí $g_i = g_1 + \sum_{j=1}^{i-1} \alpha_j B p_j$, kde $g_1 = g$. Použijeme-li (605) a (608) můžeme psát

$$p_i^T g_i = p_i^T \left(g + \sum_{j=1}^{i-1} \alpha_j B p_j \right) = p_i^T g$$

□

Poznámka 238. Pokud $i \in I_2$, jsou vektory p_{i+1} a g_{i+1} výjimečné v tom, že platí nerovnosti $p_i^T B p_{i+1} \neq 0$, $p_i^T g_{i+1} = p_i^T (g_i + \alpha_i B p_i) = p_i^T g_i \neq 0$ a $g_i^T g_{i+1} = g_i^T (g_i + \alpha_i B p_i) = g_i^T g_i - \alpha_i (p_i - \beta_{i-1} p_{i-1})^T B p_i = g_i^T g_i \neq 0$. Použijeme-li planární krok, netvoří nenulové gradienty ortogonální systém. Abychom dostali ortogonální systém $\tilde{g}_1, \ldots, \tilde{g}_m$ (kde $m \in I_1$ nebo $m - 1 \in I_2$), položíme $\tilde{g}_i = g_i$, $i \in I_1$, a $\tilde{g}_i = g_i$, $\tilde{g}_{i+1} = -p_{i+1}$, $i \in I_2$. Pak podle (602) pro $1 \le j \le i$ platí $\tilde{g}_j \in \mathcal{L}(P_j)$, což spolu s (608) dává $\tilde{g}_j^T \tilde{g}_{i+1} = -\tilde{g}_j^T p_{i+1} = -\gamma_i \tilde{g}_j^T B p_i = 0$, a podle (603) a (606) pro $k > i + 1$, $k \in I_1 \cup I_2$, platí $\tilde{g}_{i+1}^T \tilde{g}_k = -p_{i+1}^T g_k = 0$.

Poznámka 239. Jsou-li splněny předpoklady věty 146 a $n \in I_1 \cup I_2$ (čili n je počátečním indexem posledního kroku), musí platit $n \in I_1$, neboli $p_n^T B p_n \neq 0$. V opačném případě by existoval vektor $p_{n+1} = \gamma_n B p_n$ takový, že $p_n^T B p_{n+1} \neq 0$, a mohli bychom provést planární krok. Postupem použitým v důkazu věty 146 bychom zjistili, že matice $P_{n+1}^T B P_{n+1}$ je blokově diagonální a regulární, což je spor, neboť vektory $p_1, \ldots, p_{n+1}$ (kterých je $n+1$) dimenze n jsou lineárně zásvislé. Případ, kdy $p_n^T B p_n = 0$ může nastat pouze tehdy, je-li matice B singulární, nebo tehdy, když $p_n = 0$ (a tedy $g_n = 0$).

V oddílu 6.4 (věta 142) je ukázáno, že metodu sdružených gradientů můžeme použít k tridiagonalizaci symetrické pozitivně definitní matice. Nyní ukážeme, že základní planární metodu sdružených gradientů lze použít k tridiagonalizaci libovolné symetrické regulární matice.

Věta 147. *Uvažujme základní planární metodu sdružených gradientů (definice 46) aplikovanou na soustavu rovnic $Bs + g = 0$, kde B je symetrická regulární matice. Označme $P_m = [p_1, \ldots, p_m]$, $\tilde{G}_m = [\tilde{g}_1, \ldots, \tilde{g}_m]$, $\tilde{N}_m = \operatorname{diag}(\|\tilde{g}_1\|, \ldots, \|\tilde{g}_m\|)$ a $\tilde{Q}_m = \tilde{G}_m \tilde{N}_m^{-1}$, kde $1 \le m \le n$ ($m \in I_1$ nebo $m - 1 \in I_2$) a kde $\tilde{g}_1, \ldots, \tilde{g}_m$ je ortogonální systém uvedený v poznámce 238. Pak platí*

$$\tilde{Q}_m^T B \tilde{Q}_m = \tilde{T}_m = \tilde{L}_m \tilde{D}_m \tilde{L}_m^T, \tag{609}$$

kde $\tilde{T}_m$ je symetrická tridiagonální matice, $\tilde{L}_m$ je dolní blokově bidiagonální matice a $\tilde{D}_m$ je blokově diagonální matice (diagonální bloky mají rozměr 1×1 nebo 2×2). Pokud $m = n$, platí $B = \tilde{Q}_n \tilde{T}_n \tilde{Q}_n^T$, kde matice $\tilde{Q}_n$ je ortogonální.

Důkaz Bez újmy na obecnosti budeme předpokládat, že $m = 8$, první dva kroky jsou jednoduché, další dva planární a zbylé dva jednoduché, což nám umožní přehledně znázornit strukturu matice $\tilde{T}_m$.

(a) Použijme-li definiční vztahy pro vektory p_i (definice 46) a $\tilde{g}_i$ (poznámka 238), kde $1 \le i < m$, dostaneme

$$\begin{aligned}
\tilde{g}_1 &= -p_1 \\
\tilde{g}_{i+1} &= -p_{i+1} + \beta_i p_i, \quad i \in I_1, \\
\tilde{g}_{i+1} &= -p_{i+1}, \quad i \in I_2, \\
\tilde{g}_{i+2} &= -p_{i+2} + \beta_i p_i, \quad i \in I_2,
\end{aligned}$$

takže $\tilde{G}_m = P_m \tilde{B}_m$, kde

$$\tilde{B}_m = \left[\begin{array}{cc|cc|cc|cc}
-1 & \beta_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & -1 & \beta_2 & 0 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & -1 & 0 & \beta_3 & 0 & 0 & 0 \\
0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & -1 & 0 & \beta_5 & 0 \\
0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & -1 & \beta_7 \\ \hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & -1
\end{array}\right]$$

je horní blokově bidiagonální matice .

(b) Použijeme-li definiční vztahy pro koeficienty α_i a α_{i+1} (definice 46), můžeme pro $i \in I_1$ psát

$$p_i^T B p_i = \frac{\|\tilde{g}_i\|^2}{\alpha_i}$$

a pro $i \in I_2$ platí $p_i^T B p_{i+1} = \|\tilde{g}_i\|^2 / \alpha_{i+1}$. Jelikož také $p_i^T B p_{i+1} = p_{i+1}^T p_{i+1} / \gamma_i = \|\tilde{g}_{i+1}\|^2 / \gamma_i$, dostaneme

$$p_i^T B p_{i+1} = \frac{\|\tilde{g}_i\| \|\tilde{g}_{i+1}\|}{\sqrt{\gamma_i \alpha_{i+1}}}$$

a

$$p_{i+1}^T B p_{i+1} = \frac{\alpha_i}{\alpha_{i+1}} p_i^T B p_{i+1} = \frac{\alpha_i \|\tilde{g}_{i+1}\|^2}{\gamma_i \alpha_{i+1}}.$$

Podle (605) a (608) tedy platí

$$P_m^T B P_m = \left[\begin{array}{c|c|cc|cc|c|c} p_1^T B p_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & p_2^T B p_2 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & p_3^T B p_4 & 0 & 0 & 0 & 0 \\ 0 & 0 & p_4^T B p_3 & p_4^T B p_4 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & p_5^T B p_6 & 0 & 0 \\ 0 & 0 & 0 & 0 & p_6^T B p_5 & p_6^T B p_6 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & p_7^T B p_7 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & p_8^T B p_8 \end{array}\right] = \tilde{N}_m \tilde{D}_m \tilde{N}_m,$$

kde

$$D_m = \left[\begin{array}{c|c|cc|cc|c|c} \frac{1}{\alpha_1} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & \frac{1}{\alpha_2} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & \frac{1}{\sqrt{\gamma_3 \alpha_4}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{\gamma_3 \alpha_4}} & \frac{\alpha_3}{\gamma_3 \alpha_4} & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{\gamma_5 \alpha_6}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{\gamma_5 \alpha_6}} & \frac{\alpha_5}{\gamma_5 \alpha_6} & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\alpha_7} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\alpha_8} \end{array}\right].$$

(c) Označme $\tilde{Q}_m = \tilde{G}_m \tilde{N}_m^{-1}$ a $\tilde{T}_m = \tilde{Q}_m^T B \tilde{Q}_m$. Pak podle (a) a (b) platí

$$\begin{aligned} \tilde{T}_m &= \tilde{Q}_m^T B \tilde{Q}_m = \tilde{N}_m^{-1} \tilde{G}_m^T B \tilde{G}_m \tilde{N}_m^{-1} = \tilde{N}_m^{-1} \tilde{B}_m^T P_m^T B P_m \tilde{B}_m \tilde{N}_m^{-1} \\ &= \tilde{N}_m^{-1} \tilde{B}_m^T \tilde{N}_m \tilde{D}_m \tilde{N}_m \tilde{B}_m \tilde{N}_m^{-1} = \tilde{R}_m^T \tilde{D}_m \tilde{R}_m, \end{aligned}$$

kde $\tilde{R}_m = \tilde{N}_m \tilde{B}_m \tilde{N}_m^{-1}$. Použijeme-li (603)–(608), můžeme pro $i \in I_1$ psát

$$\beta_i = \frac{p_i^T B g_{i+1}}{p_i^T B p_i} = \frac{(g_{i+1} - g_i)^T g_{i+1}}{(g_{i+1} - g_i)^T p_i} = \frac{\|g_{i+1}\|^2}{\|g_i\|^2}$$

a pro $i \in I_2$ platí

$$\beta_i = \frac{p_{i+1}^T B g_{i+2}}{p_{i+1}^T B p_i} = \frac{(g_{i+2} - g_{i+1})^T g_{i+2}}{(g_{i+2} - g_{i+1})^T p_i} = \frac{\|g_{i+2}\|^2}{\|g_i\|^2},$$

neboť $p_i^T g_{i+1} = p_i^T (g_i + \alpha_i B p_i) = p_i^T g_i = -g_i^T g_i$, což po dosazení dává

$$\tilde{R}_m = \tilde{N}_m \tilde{B}_m \tilde{N}_m^{-1} = \left[\begin{array}{c|c|c|c|c|c|c|c} -1 & \beta_1 \frac{\|g_1\|}{\|g_2\|} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -1 & \beta_2 \frac{\|g_2\|}{\|g_3\|} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & \beta_3 \frac{\|g_3\|}{\|g_5\|} & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & 0 & \beta_5 \frac{\|g_5\|}{\|g_7\|} & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & -1 & \beta_7 \frac{\|g_7\|}{\|g_8\|} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right]$$

$$= \left[\begin{array}{c|c|c|c|c|c|c|c} -1 & \sqrt{\beta_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -1 & \sqrt{\beta_2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -1 & 0 & \sqrt{\beta_3} & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & -1 & 0 & \sqrt{\beta_5} & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & -1 & \sqrt{\beta_7} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right].$$

Označme $\tilde{\beta}_i = \|\tilde{g}_{i+1}\|/\|\tilde{g}_i\|$, $1 \leq i \leq m$, takže $\beta_i = \tilde{\beta}_i$ pro $i \in I_1$ a $\beta_i = \tilde{\beta}_i \tilde{\beta}_{i+1}$ pro $i \in I_2$. Pro $i \in I_2$ pak platí

$$\frac{1}{\gamma_i \alpha_{i+1}} = \frac{p_i^T B p_{i+1}}{\gamma g_i^T g_i} = \frac{\|\tilde{g}_{i+1}\|}{\gamma_i^2 \|\tilde{g}_i\|}. \tag{610}$$

Vynásobením matic $\tilde{R}_m^T \tilde{D}_m \tilde{R}_m$ a použitím (610) dostaneme

$$\tilde{T}_m = \tilde{R}_m^T \tilde{D}_m \tilde{R}_m = \left[\begin{array}{c|c|c|c|c|c|c|c} \frac{1}{\alpha_1} & -\frac{\sqrt{\beta_1}}{\alpha_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline -\frac{\sqrt{\beta_1}}{\alpha_1} & \frac{\beta_1}{\alpha_1}+\frac{1}{\alpha_2} & -\frac{\sqrt{\beta_2}}{\alpha_2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\frac{\sqrt{\beta_2}}{\alpha_2} & \frac{\beta_2}{\alpha_2} & \frac{1}{\sqrt{\gamma_3\alpha_4}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{\gamma_3\alpha_4}} & \frac{\alpha_3}{\gamma_3\alpha_4} & -\frac{\sqrt{\beta_3}}{\sqrt{\gamma_3\alpha_4}} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -\frac{\sqrt{\beta_3}}{\sqrt{\gamma_3\alpha_4}} & 0 & \frac{1}{\sqrt{\gamma_5\alpha_6}} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{\gamma_5\alpha_6}} & \frac{\alpha_5}{\gamma_5\alpha_6} & -\frac{\sqrt{\beta_5}}{\sqrt{\gamma_5\alpha_6}} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{\beta_5}}{\sqrt{\gamma_5\alpha_6}} & \frac{1}{\alpha_7} & -\frac{\sqrt{\beta_7}}{\alpha_7} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{\beta_7}}{\alpha_7} & \frac{\beta_7}{\alpha_7}+\frac{1}{\alpha_8} \end{array}\right]$$

$$= \left[\begin{array}{c|c|c|c|c|c|c|c} \frac{1}{\alpha_1} & -\frac{\sqrt{\tilde{\beta}_1}}{\alpha_1} & 0 & 0 & 0 & 0 & 0 & 0 \\ \hline -\frac{\sqrt{\tilde{\beta}_1}}{\alpha_1} & \frac{\tilde{\beta}_1}{\alpha_1}+\frac{1}{\alpha_2} & -\frac{\sqrt{\tilde{\beta}_2}}{\alpha_2} & 0 & 0 & 0 & 0 & 0 \\ \hline 0 & -\frac{\sqrt{\tilde{\beta}_2}}{\alpha_2} & \frac{\tilde{\beta}_2}{\alpha_2} & \frac{\sqrt{\tilde{\beta}_3}}{\gamma_3} & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{\tilde{\beta}_3}}{\gamma_3} & \frac{\alpha_3\tilde{\beta}_3}{\gamma_3^2} & -\frac{\tilde{\beta}_3\sqrt{\tilde{\beta}_3}}{\gamma_3^2} & 0 & 0 & 0 \\ \hline 0 & 0 & 0 & -\frac{\tilde{\beta}_3\sqrt{\tilde{\beta}_3}}{\gamma_3^2} & 0 & \frac{\sqrt{\tilde{\beta}_5}}{\gamma_5} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{\sqrt{\tilde{\beta}_5}}{\gamma_5} & \frac{\alpha_5\tilde{\beta}_5}{\gamma_5^2} & -\frac{\tilde{\beta}_5\sqrt{\tilde{\beta}_5}}{\gamma_5^2} & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -\frac{\tilde{\beta}_5\sqrt{\tilde{\beta}_5}}{\gamma_5^2} & \frac{1}{\alpha_7} & -\frac{\sqrt{\tilde{\beta}_7}}{\alpha_7} \\ \hline 0 & 0 & 0 & 0 & 0 & 0 & -\frac{\sqrt{\tilde{\beta}_7}}{\alpha_7} & \frac{\tilde{\beta}_7}{\alpha_7}+\frac{1}{\alpha_8} \end{array}\right],$$

což je symetrická tridiagonální matice. Položíme-li $\tilde{L}_m = \tilde{R}_m^T$, dostaneme (609).

(d) Matice $\tilde{G}_m$ má ortogonální a matice $\tilde{Q}_m$ ortonormální sloupce, takže pro $m = n$ je matice $\tilde{Q}_n$ čtvercová a ortogonální a podle (609) platí $\tilde{Q}_n \tilde{T}_n \tilde{Q}_n^T = \tilde{Q}_n \tilde{Q}_n^T B \tilde{Q}_n \tilde{Q}_n^T = B$. □

Rozklad (609) je zobecněním rozkladu (595), kde matice $\tilde{D}_m$ je nyní blokově diagonální.

Jak již bylo poznamenáno, při použití základní planární metody sdružených gradientů mohou nastat potíže, pokud $p_i^T B p_i \approx 0$. Proto byly vyvinuty modifikované planární metody sdružených gradientů, které používají planární krok i tehdy, když $0 < |p_i^T B p_i| < \underline{c}\|p_i\|^2$ a které naleznou řešení soustavy rovnic $Bs + g = 0$ po nejvýše n krocích [80], [47]. Jestliže $i \in I_2$, pokud $0 < |p_i^T B p_i| < \underline{c}\|p_i\|^2$, nemůžeme položit $p_{i+1} = Bp_i$, neboť v tomto případě nelze splnit indukční předpoklady (602)–(608). Proto se pokládá $p_{i+1} = q_i$, kde vektor q_i se počítá rekurentně, přičemž obecně $q_i \neq Bp_i$ a $q_i^T g_i \neq 0$. Aby platilo $p_i^T g_{i+2} = 0$, $q_i^T g_{i+2} = 0$, je třeba koeficienty α_i, α_{i+1}, vystupující ve vztahu $s_{i+2} = s_i + \alpha_i p_i + \alpha_{i+1} q_i$, určit řešením soustavy rovnic

$$\begin{aligned} \alpha_i\, p_i^T B p_i + \alpha_{i+1}\, p_i^T B q_i &= -p_i^T g_i, \\ \alpha_i\, q_i^T B p_i + \alpha_{i+1}\, q_i^T B q_i &= -q_i^T g_i, \end{aligned}$$

což dává

$$\alpha_i = \frac{1}{\delta_i}(b_i q_i^T g_i - c_i p_i^T g_i), \qquad \alpha_{i+1} = \frac{1}{\delta_i}(b_i p_i^T g_i - a_i q_i^T g_i),$$

kde $a_i = p_i^T B p_i$, $b_i = p_i^T B q_i$, $c_i = q_i^T B q_i$ a $\delta_i = a_i c_i - b_i^2$. Na základě těchto úvah lze definovat následující modifikovanou planární metodu sdružených gradientů, kde $I_1 = \{i \in N : |p_i^T B p_i| \geq \underline{c} \|p_i\|^2\}$ a $\underline{c} > 0$.

Definice 47. *Nechť $B \in R^{n \times n}$ je symetrická regulární matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \qquad g_1 = g, \qquad p_1 = -g, \qquad u_1 = Bp_1, \qquad q_1 = u_1, \qquad a_1 = u_1^T p_1,$$

a buď

$$\begin{array}{lll} \alpha_i = -p_i^T g_i / a_i, & s_{i+1} = s_i + \alpha_i p_i, & g_{i+1} = g_i + \alpha_i u_i, \\ \beta_i = u_i^T g_{i+1} / a_i, & p_{i+1} = -g_{i+1} + \beta_i p_i, & u_{i+1} = B p_{i+1}, \\ \gamma_i = u_i^T u_{i+1} / a_i, & q_{i+1} = u_{i+1} + \gamma_i p_i, & a_{i+1} = u_{i+1}^T p_{i+1} \end{array}$$

pokud $i \in I_1$ (jednoduchý krok), nebo

$$\begin{array}{rl} v_i = B q_i, & b_i = v_i^T p_i, \\ c_i = v_i^T q_i, & \delta_i = a_i c_i - b_i^2, \\ \alpha_i = (b_i q_i^T g_i - c_i p_i^T g_i)/\delta_i, & \alpha_{i+1} = (b_i p_i^T g_i - a_i q_i^T g_i)/\delta_i, \\ s_{i+2} = s_i + \alpha_i p_i + \alpha_{i+1} q_i, & g_{i+2} = g_i + \alpha_i u_i + \alpha_{i+1} v_i, \\ \beta_i = v_i^T g_{i+2}/\delta_i, & p_{i+2} = -g_{i+2} + \beta_i (a_i q_i - b_i p_i), \\ u_{i+2} = B p_{i+2}, & a_{i+2} = u_{i+2}^T p_{i+2}, \\ \gamma_i = v_i^T u_{i+2}/\delta_i, & q_{i+2} = u_{i+2} - \gamma_i (a_i q_i - b_i p_i), \end{array}$$

pokud $i \in I_2$ (planární krok), nazveme modifikovanou planární metodou sdružených gradientů pro řešení soustavy rovnic $Bs + g = 0$.

Poznámka 240. Nejstarší modifikovaná planární metoda sdružených gradientů, která používá rekurentní vztahy uvedené v definici 47 a která najde řešení soustavy rovnic $Bs + g = 0$ po konečném počtu kroků, je popsána v práci [80]. Tato metoda však používá množinu $I_1 = \{i \in N : |\delta_i| \geq \underline{c}\, b_i^2\}$, takže je třeba v každém (i jednoduchém) iteračním kroku použít dvě násobení matice vektorem (to druhé pro výpočet vektoru v_i a čísla c_i), což prodlužuje čas výpočtu. Úprava, která používá množinu $I_1 = \{i \in N : |p_i^T B p_i| \geq \underline{c} \|p_i\|^2\}$, pochází z práce [47].

Modifikovaná planární metoda sdružených gradientů má podobné vlastnosti jako základní planární metoda sdružených gradientů.

Tvrzení 5. *Modifikovaná planární metoda sdružených gradientů (definice 46) najde řešení soustavy rovnic $Bs + g = 0$, kde B je symetrická regulární matice, po nejvýše n krocích (planární krok počítáme za dva jednoduché kroky). Navíc jsou splněny podmínky (602)–(608), kde $p_{j+1} = q_j$, pokud $j \in I_2$ (rovnosti (608) jsou ovšem splněny pouze pro $1 \leq j < i$, neboť pro $i \in I_2$ již obecně neplatí $p_i^T B p_i = 0$).*

Poznámka 241. Důkaz tvrzení 5 je podobný důkazu věty 146. Opět se indukcí dokazují vztahy (602)–(608). Postup důkazu je však technicky komplikovaný a zde ho uvádět nebudene (lze ho nalézt v pracech [80] a [47]).

Poznámka 242. ze vztahů (602)–(608) plynou rovnosti $p_i^T g_i = p_i^T g$ pro $i \in I_1$ a $p_i^T g_i = p_i^T g$, $q_i^T g_i = q_i^T g$ pro $i \in I_2$. K jejich odvození lze použít postup uvedený v důkazu důsledku 21.

Věta 148. *Uvažujme modifikovanou planární metodou sdružených gradientů (definice 47) a položme*

$$s = -\sum_{i \in I_1} \frac{p_i^T g_i}{|p_i^T B p_i|} p_i - \sum_{i \in I_2} \left(\frac{p_i^T g_i}{\|Bp_i\|^2} p_i + \frac{q_i^T g_i}{\|Bq_i\|^2} q_i \right). \tag{611}$$

Nechť $\underline{\lambda} = \min_{1\le i\le n}|\lambda_i|$ *a* $\overline{\lambda} = \max_{1\le i\le n}|\lambda_i|$, *kde* λ_i, $1 \le i \le n$, *jsou vlastní čísla matice* B *(takže* $0 < \underline{\lambda} \le \underline{\lambda}$*). Pak platí*

$$s^T g \le -\underline{s}\|g\|^2, \qquad \|s\| \le \overline{s}\|g\|,$$

kde $\underline{s} = 1/\max(\overline{\lambda}, \overline{\lambda}^2)$ *a* $\overline{s} = n/\min(\underline{c}, \underline{\lambda}^2)$.

Důkaz (a) Jelikož podle poznámky 242 platí $p_i^T g_i = p_i^T g$ a $q_i^T g_i = q_i^T g$, můžeme psát

$$s^T g = -\sum_{i\in I_1} \frac{(p_i^T g)^2}{|p_i^T B p_i|} - \sum_{i\in I_2} \left(\frac{(p_i^T g)^2}{\|Bp_i\|^2} + \frac{(q_i^T g)^2}{\|Bq_i\|^2} \right).$$

Pokud $1 \in I_1$, dostaneme

$$s^T g \le -\frac{(p_1^T g)^2}{|p_1^T B p_1|} = -\frac{(g^T g)^2}{|g^T B g|} \le -\frac{1}{\overline{\lambda}}\|g\|^2.$$

Pokud $1 \in I_2$, dostaneme

$$s^T g \le -\frac{(p_1^T g)^2}{\|Bp_1\|^2} = -\frac{(g^T g)^2}{\|Bg\|^2} \le -\frac{1}{\overline{\lambda}^2}\|g\|^2.$$

Platí tedy $s^T g \le -\underline{s}\|g\|^2$, kde $\underline{s} = 1/\max(\overline{\lambda}, \overline{\lambda}^2)$.

(b) Podobným způsobem dostaneme

$$\begin{aligned}
\sum_{i\in I_1} \left\| \frac{p_i^T g_i}{|p_i^T B p_i|} p_i \right\| &= \sum_{i\in I_1} \left\| \frac{p_i^T g}{|p_i^T B p_i|} p_i \right\| = \sum_{i\in I_1} \frac{\|p_i p_i^T\|}{|p_i^T B p_i|}\|g\| \le \sum_{i\in I_1} \frac{\|p_i\|^2}{\underline{c}\|p_i\|^2}\|g\| \le \frac{n}{\underline{c}}\|g\|, \\
\sum_{i\in I_2} \left\| \frac{p_i^T g_i}{\|Bp_i\|^2} p_i \right\| &= \sum_{i\in I_2} \left\| \frac{p_i^T g}{\|Bp_i\|^2} p_i \right\| = \sum_{i\in I_2} \frac{\|p_i p_i^T\|}{p_i^T B^2 p_i}\|g\| \le \sum_{i\in I_2} \frac{\|p_i\|^2}{\underline{\lambda}^2\|p_i\|^2}\|g\| \le \frac{n}{2\underline{\lambda}^2}\|g\|, \\
\sum_{i\in I_2} \left\| \frac{q_i^T g_i}{\|Bq_i\|^2} q_i \right\| &= \sum_{i\in I_2} \left\| \frac{q_i^T g}{\|Bq_i\|^2} q_i \right\| = \sum_{i\in I_2} \frac{\|q_i q_i^T\|}{q_i^T B^2 q_i}\|g\| \le \sum_{i\in I_2} \frac{\|q_i\|^2}{\underline{\lambda}^2\|q_i\|^2}\|g\| \le \frac{n}{2\underline{\lambda}^2}\|g\|.
\end{aligned}$$

Podle (611) tedy platí

$$\|s\| \le \sum_{i\in I_1} \left\| \frac{p_i^T g_i}{|p_i^T B p_i|} p_i \right\| + \sum_{i\in I_2} \left\| \frac{p_i^T g_i}{\|Bp_i\|^2} p_i \right\| + \sum_{i\in I_2} \left\| \frac{q_i^T g_i}{\|Bq_i\|^2} q_i \right\| \le \max\left(\frac{n}{\underline{c}}, \frac{n}{\underline{\lambda}^2} \right) \|g\| = \overline{s}\|g\|,$$

kde $\overline{s} = n/\min(\underline{c}, \underline{\lambda}^2)$.

Poznámka 243. Je-li matice B pozitivně definitní a platí-li $0 < \underline{c} \le \underline{\lambda}(B)$, generuje modifikovaná planární metoda sdružených gradientů stejný vektor jako klasická metoda sdružených gradientů. Jsou-li vsechny kroky jednoduché, je vektor s totožný s vektorem $s + z$ vystupujícím v lemmatu 30. Význam modifikované planární metody sdružených gradientů spočívá v tom, že iterační proces pokračuje i když $|p_i^T B p_i| < \underline{c}\|p_i\|^2$ a výsledný směrový vektor lépe odpovídá vlastnostem matice B.

Druhá skupina iteračích metod pro řešení lineárních soustav se symetrickou indefinitní maticí je založena na použití symetrického Lanczosova procesu (definice 44). V oddílu 6.4 je ukázáno, že je-li symetrická matice B pozitivně definitní, je možné najít řešení soustavy $Bs + g = 0$ postupným hledáním globálních minim kvadratické funkce $(1/2)s^T Bs + g^T s$ na Krylovových podprostorech $\mathcal{K}_i$, $1 \le i \le n$, určených maticí B a vektorem g, jejichž báze tvoří vektory q_i, $1 \le i \le n$, generované symetrickým Lanczosovým procesem. To znamená, že pro $1 \le i \le n$ generujeme vektory $s_{i+1} = \arg\min_{s\in\mathcal{K}_i}(1/2)s^T Bs + g^T s$, neboli (vzhledem k tomu, že $Q_i^T B Q_i = T_i$ a $Q_i^T g = \delta_1 e_1$) řešíme soustavy rovnic $T_i z_i + \delta_1 e_1 = 0$, $1 \le i \le n$, a pokládáme $s_{i+1} = Q_i z_i$. Soustavu rovnic $T_i z_i + \delta_1 e_1 = 0$ řešíme pomocí Choleského rozkladu $T_i = L_i D_i L_i^T$, což podle poznámky 231 vede na metodu sdružených gradientů. Je-li matice B indefinitní, nelze tento postup použít,

neboť kvadratická funkce $(1/2)s^T Bs + g^T s$ není zdola omezená. Vektor $s_{i+1} = Q_i z_i$ lze však určit řešením úlohy

$$\begin{aligned} s_{i+1} &= \arg\min_{s\in\mathcal{K}_i} \|Bs+g\| = \arg\min_{z\in R^i} \|BQ_i z + g\| = \arg\min_{z\in R^i} \|Q_i T_i z + \delta_{i+1} q_{i+1} + g\| \\ &= \arg\min_{z\in R^i} \left\| Q_{i+1} \left(\begin{bmatrix} T_i \\ \delta_{i+1} e_i^T \end{bmatrix} + \delta_1 e_1 \right) \right\| = \arg\min_{z\in R^i} \|H_i z + \delta_1 e_1\|, \end{aligned} \tag{612}$$

kde

$$H_i = \begin{bmatrix} T_i \\ \delta_{i+1} e_i^T \end{bmatrix} = \begin{bmatrix} \gamma_1, & \delta_2, & \ldots, & 0, & 0 \\ \delta_2, & \gamma_2, & \ldots, & 0, & 0 \\ --, & --, & --, & --, & -- \\ 0, & 0, & \ldots, & \gamma_{i-1}, & \delta_i \\ 0, & 0, & \ldots, & \delta_i, & \gamma_i \\ 0, & 0, & \ldots, & 0, & \delta_{i+1} \end{bmatrix}$$

je tridiagfonální horní Hessenbergova matice (používáme rovnici (593), vztah $g = \delta_1 q_1$ a skutečnost, že matice Q_{i+1} má ortonormální sloupce, takže $Q_{i+1}^T Q_{i+1} = I$). Vektor $s_{i+1} = Q_i z_i$ lze tedy určit řešením lineární úlohy nejmenších čtverců $H_i z_i + \delta_1 e_1 \approx 0$ (řešení této úlohy je popsáno v oddílu 8.4). Je-li matice B regulární, má matice H_i lineárně nezávislé sloupce a lze ji pomocí ortogonálních transformací popsaných v oddílu 8.4 převést na horní trojúhelníkovou matici

$$P_i H_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \qquad P_i(\delta_1 e_1) = \begin{bmatrix} h_i \\ \bar{\eta}_i \end{bmatrix}$$

(poznámka 272), kde

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \tau_3, & \ldots, & 0, & 0, & 0 \\ 0, & \rho_2, & \sigma_3 & \ldots, & 0, & 0, & 0 \\ --, & --, & --, & --, & --, & --, & -- \\ 0, & 0, & 0, & \ldots, & \rho_{i-2}, & \sigma_{i-1}, & \tau_i \\ 0, & 0, & 0, & \ldots, & 0, & \rho_{i-1}, & \sigma_i \\ 0, & 0, & 0, & \ldots, & 0, & 0, & \rho_i \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ -- \\ \eta_{i-2} \\ \eta_{i-1} \\ \eta_i \end{bmatrix}$$

Přitom P_i, $1 \le i \le k$, jsou ortogonální matice dimenze $i+1$ a R_i, $1 \le i \le k$, jsou tridiagonální horní trojúhelníkové matice dimenze i. Matice P_i, $1 \le i \le k$, se počítají rekurentně pomocí Givensových matic elementárních rotací $\tilde{P}_i \in R^{2\times 2}$, $1 \le i \le k$, studovaných v oddílu 8.4.

Poznámka 244. Předpokládejme, že

$$\begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} [H_i, t_{i+1}, t_{i+2}, \delta_1 e_1] = \begin{bmatrix} \rho_1, & \sigma_2, & \tau_3, & \ldots, & 0, & 0, & 0, & 0, & \eta_1 \\ 0, & \rho_2, & \sigma_3, & \ldots, & 0, & 0, & 0, & 0, & \eta_2 \\ --, & --, & --, & --, & --, & --, & --, & --, & -- \\ 0, & 0, & 0, & \ldots, & \rho_{i-1}, & \sigma_i, & \tau_{i+1}, & 0, & \eta_{i-1} \\ 0, & 0, & 0, & \ldots, & 0, & \tilde{\rho}_i, & \tilde{\sigma}_{i+1}, & 0, & \tilde{\eta}_i \\ 0, & 0, & 0, & \ldots, & 0, & \delta_{i+1}, & \gamma_{i+1}, & \delta_{i+2}, & 0 \end{bmatrix}, \tag{613}$$

kde t_{i+1}, t_{i+2} jsou vektory, které obsahují prvních $i+1$ prvků posledních dvou sloupců matice T_{i+2}. Abychom vynulovali prvek δ_{i+1}, sestrojíme Givensovu ortogonální matici

$$\tilde{P}_i = \frac{1}{\sqrt{\tilde{\rho}_i^2 + \delta_{i+1}^2}} \begin{bmatrix} \tilde{\rho}_i, & \delta_{i+1} \\ -\delta_{i+1}, & \tilde{\rho}_i \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i, & \delta_{i+1} \\ -\delta_{i+1}, & \tilde{\rho}_i \end{bmatrix}, \qquad \rho_i = \sqrt{\tilde{\rho}_i^2 + \delta_{i+1}^2}.$$

Pak podle věty 144 platí

$$\tilde{P}_i \begin{bmatrix} \tilde{\rho}_i \\ \delta_{i+1} \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i^2 + \delta_{i+1}^2 \\ 0 \end{bmatrix} = \begin{bmatrix} \rho_i \\ 0 \end{bmatrix}, \qquad \tilde{P}_i \begin{bmatrix} \tilde{\eta}_i \\ 0 \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i \tilde{\eta}_i \\ -\delta_{i+1} \tilde{\eta}_i \end{bmatrix} = \begin{bmatrix} \eta_i \\ \tilde{\eta}_{i+1} \end{bmatrix}. \tag{614}$$

$$\tilde{P}_i \begin{bmatrix} \tilde{\sigma}_{i+1}, & 0 \\ \gamma_{i+1}, & \delta_{i+2} \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i\tilde{\sigma}_{i+1} + \delta_{i+1}\gamma_{i+1}, & \delta_{i+1}\delta_{i+2} \\ \tilde{\rho}_i\gamma_{i+1} - \delta_{i+1}\tilde{\sigma}_{i+1}, & \tilde{\rho}_i\delta_{i+2} \end{bmatrix} = \begin{bmatrix} \sigma_{i+1}, & \tau_{i+2} \\ \tilde{\rho}_{i+1}, & \tilde{\sigma}_{i+2} \end{bmatrix}, \tag{615}$$

Ortogonální matice P_i, $1 \le i \le k$, budeme hledat ve tvaru $P_1 = \tilde{P}_1$ a

$$P_i = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix},$$

kde I je jednotková matice řádu $i-2$. Pak podle (613) a (614) pro $1 < i \le k$ platí

$$P_iH_i = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} H_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \quad P_i(\delta_1 e_1) = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} (\delta_1 e_1) = \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix}.$$

Poznamenejme, že v i-tém iterační kroku ještě neznáme koeficienty γ_{i+1} a δ_{i+2}, takže nelze použít transformaci (615). Místo toho na začátku i-tého iteračního kroku transformujeme koeficienty γ_i a δ_{i+1}, získané v předchozím iteračním kroku, podle vzorce

$$\tilde{P}_{i-1} \begin{bmatrix} \tilde{\sigma}_i, & 0 \\ \gamma_i, & \delta_{i+1} \end{bmatrix} = \frac{1}{\rho_{i-1}} \begin{bmatrix} \tilde{\rho}_{i-1}\tilde{\sigma}_i + \delta_i\gamma_i, & \delta_i\delta_{i+1} \\ \tilde{\rho}_{i-1}\gamma_i - \delta_i\tilde{\sigma}_i, & \tilde{\rho}_{i-1}\delta_{i+1} \end{bmatrix} = \begin{bmatrix} \sigma_i, & \tau_{i+1} \\ \tilde{\rho}_i, & \tilde{\sigma}_{i+1} \end{bmatrix}, \tag{616}$$

Použijeme-li vztahy (614) a (616), dostaneme následující tvrzení.

Lemma 67. *Prvky matic R_i a vektorů h_i, $1 \le i \le k$, lze počítat podle rekurentních vztahů $\lambda_0 = 1$, $\mu_0 = 0$, $\tilde{\sigma}_1 = 0$, $\tilde{\eta}_1 = \delta_1$ a*

$$\begin{aligned} \sigma_i &= \lambda_{i-1}\tilde{\sigma}_i + \mu_{i-1}\gamma_i, & \tilde{\rho}_i &= \lambda_{i-1}\gamma_i + \mu_{i-1}\tilde{\sigma}_i, \\ \tilde{\sigma}_{i+1} &= \lambda_{i-1}\delta_{i+1} & \tau_{i+1} &= \mu_{i-1}\delta_{i+1} \\ \rho_i &= \sqrt{\tilde{\rho}_i^2 + \delta_{i+1}^2}, & \lambda_i &= \frac{\tilde{\rho}_i}{\rho_i}, \qquad \mu_i = \frac{\delta_{i+1}}{\rho_i}, \\ \eta_i &= \lambda_i\tilde{\eta}_i, & \tilde{\eta}_{i+1} &= -\mu_i\tilde{\eta}_i \end{aligned}$$

pro $1 \le i \le k$.

Nyní odvodíme rekurentní vztahy pro vektory s_{i+1}, $1 \le i \le k$. Jelikož

$$P_i(B_iz + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix}$$

a $P_i^TP_i = I$, můžeme položit $s_{i+1} = Q_iz_i$, kde

$$z_i = \arg\min_{z \in R^i} \left\| \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix} \right\|.$$

Jelikož matice $R_i \in R^{i \times i}$ je regulární, musí podle věty 167 platit $R_iz_i + h_i = 0$. Vzhledem k jednoduché struktuře matic R_i, $1 \le i \le k$, můžeme vektory z_i, $1 \le i \le k$, a tudíž i vektory s_{i+1}, $1 \le i \le k$, určovat rekurentně.

Lemma 68. *Vektory $s_{i+1} = Q_iz_i$, $1 \le i \le k$, kde $R_iz_i + h_i = 0$, lze určit pomocí rekurentních vztahů $p_0 = 0$, $p_1 = q_1$, $s_1 = 0$ a*

$$\begin{aligned} s_{i+1} &= s_i - \frac{\eta_i}{\rho_i}p_i, \\ p_{i+1} &= q_{i+1} - \frac{\sigma_{i+1}}{\rho_i}p_i - \frac{\tau_{i+1}}{\rho_i}p_{i-1} \end{aligned}$$

pro $1 \le i \le k$.

Důkaz Platí $R_1 = [\rho_1]$, $R_1^{-1} = [1/\rho_1]$ a

$$R_i = \begin{bmatrix} R_{i-1}, & r_{i-1} \\ 0, & \rho_i \end{bmatrix}, \quad R_i^{-1} = \begin{bmatrix} R_{i-1}^{-1}, & -R_{i-1}^{-1} r_{i-1}/\rho_i \\ 0, & 1/\rho_i \end{bmatrix}$$

pro $1 < i \leq k$, kde $r_{i-1} = \sigma_i e_{i-1} + \tau_i e_{i-2}$ a e_{i-2}, e_{i-1} jsou poslední dva sloupce jednotkové marice řádu $i-1$, přičemž $e_0 = 0$ (vztah pro R_i^{-1} můžeme ověřit dosazením do rovnosti $R_i R_i^{-1} = I$). Označíme-li $\tilde{r}_{i-2}$ poslední sloupec matice R_{i-2}^{-1} doplněný o nulu a položíme-li $z_i = -R_i^{-1} h_i$, dostaneme z předchozích rovností rekurentní vztahy $r_1 = [1/\rho_1]$, $z_1 = -[\eta_1/\rho_1]$ a

$$r_i = \begin{bmatrix} -(\sigma_i r_{i-1} + \tau_i \tilde{r}_{i-2})/\rho_i \\ 1/\rho_i \end{bmatrix}, \quad z_i = \begin{bmatrix} z_{i-1} + (\eta_i/\rho_i)(\sigma_i r_{i-1} + \tau_i \tilde{r}_{i-2}) \\ -\eta_i/\rho_i \end{bmatrix}.$$

pro $1 < i \leq k$, takže přihlédneme-li k tomu, že $Q_{i-1}\tilde{r}_{i-2} = Q_{i-2} r_{i-2}$, dostaneme

$$\begin{aligned} p_i &\triangleq \rho_i Q_i r_i = q_i - \frac{\sigma_i}{\rho_{i-1}} \rho_{i-1} Q_{i-1} r_{i-1} - \frac{\tau_i}{\rho_{i-2}} \rho_{i-2} Q_{i-2} r_{i-2} = q_i - \frac{\sigma_i}{\rho_{i-1}} p_{i-1} - \frac{\tau_i}{\rho_{i-2}} p_{i-2}, \\ s_{i+1} &= Q_i z_i = Q_{i-1} z_{i-1} + \frac{\eta_i}{\rho_i} \sigma_i Q_{i-1} r_{i-1} + \frac{\eta_i}{\rho_i} \tau_i Q_{i-2} r_{i-2} - \frac{\eta_i}{\rho_i} q_i \\ &= s_i - \frac{\eta_i}{\rho_i} \left(q_i - \frac{\sigma_i}{\rho_{i-1}} \rho_{i-1} Q_{i-1} r_{i-1} - \frac{\tau_i}{\rho_{i-2}} \rho_{i-2} Q_{i-2} r_{i-2} \right) = s_i - \frac{\eta_i}{\rho_i} p_i. \end{aligned}$$

□

Rekurentní vztahy uvedené v předchozích dvou lemmatech tvoří základ metody MINRES.

Definice 48. *Nechť $B \in R^{n \times n}$ je symetrická matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$q_0 = 0, \quad p_0 = 0, \quad \lambda_0 = 1, \quad \mu_0 = 0, \quad \delta_1 q_1 = g, \quad p_1 = q_1, \quad \tilde{\sigma}_1 = 0, \quad \tilde{\eta}_1 = \delta_1$$

a

$$\gamma_i = q_i^T B q_i, \quad \delta_{i+1} q_{i+1} = B q_i - \gamma_i q_i - \delta_i q_{i-1}$$

$$\sigma_i = \lambda_{i-1} \tilde{\sigma}_i + \mu_{i-1} \gamma_i, \quad \tilde{\rho}_i = \lambda_{i-1} \gamma_i + \mu_{i-1} \tilde{\sigma}_i,$$

$$\tilde{\sigma}_{i+1} = \lambda_{i-1} \delta_{i+1}, \quad \tau_{i+1} = \mu_{i-1} \delta_{i+1}$$

$$\rho_i = \sqrt{\tilde{\rho}_i^2 + \delta_{i+1}^2}, \quad \lambda_i = \frac{\tilde{\rho}_i}{\rho_i}, \quad \mu_i = \frac{\delta_{i+1}}{\rho_i},$$

$$\eta_i = \lambda_i \tilde{\eta}_i, \quad \tilde{\eta}_{i+1} = -\mu_i \tilde{\eta}_i$$

$$s_{i+1} = s_i - \frac{\eta_i}{\rho_i} p_i,$$

$$p_{i+1} = q_{i+1} - \frac{\sigma_{i+1}}{\rho_i} p_i - \frac{\tau_{i+1}}{\rho_i} p_{i-1}$$

pro $1 \leq i \leq n$, kde koeficienty $\delta_i \geq 0$, $1 \leq i \leq n$, se volí tak, aby vektory q_i, $1 \leq i \leq n$, měly jednotkovou normu, nazveme metodou MINRES určenou maticí B a vektorem g.

7 Metody kubické regularizace

7.1 Základní vlastnosti metod kubické regularizace

Myšlenka metod kubické regularizace je velmi podobná myšlence metod kvadratické regularizace. K lokálnímu kvadratickému modelu, který se používá u metod s lokálně omezeným krokem, se přidá regularizační člen, který je nyní třetího řádu. Dostaneme tak kubický model

$$C_i(s) = g_i^T s + \frac{1}{2} s^T B_i s + \frac{1}{3}\sigma_i \|s\|^3. \tag{617}$$

Podíl skutečného a předpověděného poklesu funkce F je pak definován vztahem

$$\rho_i(s) = \frac{F(x_i + s) - F(x_i)}{C_i(s)}. \tag{618}$$

Definice 49. *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou kubické regularizace, jestliže směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že*

$$C_i(s_i) \le \underline{\nu} C_i(s_i(\alpha^*)), \qquad s_i(\alpha_i^*) = \arg \min_{s_i(\alpha) = -\alpha g_i} C_i(s_i(\alpha)), \tag{$C1$}$$

kde $0 < \underline{\nu} \le 1$, délky kroku $\alpha_i \ge 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \le 0 \quad \Rightarrow \quad \alpha_i = 0, \tag{$C2a$}$$

$$\rho_i(s_i) > 0 \quad \Rightarrow \quad \alpha_i = 1, \tag{$C2b$}$$

a čísla $\sigma_i \ge 0$, $i \in N$, se volí tak, že

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \underline{\gamma}\sigma_i \le \sigma_{i+1} \le \overline{\gamma}\sigma_i, \tag{$C3a$}$$

$$\rho_i(s_i) \ge \underline{\rho} \quad \Rightarrow \quad 0 < \sigma_{i+1} \le \sigma_i, \tag{$C3b$}$$

kde $0 < \underline{\rho} < 1$ a $1 < \underline{\gamma} < \overline{\gamma}$. Řekneme, že metoda kubické regularizace je striktní metodou kubické regularizace, jsou-li podmínky (C2a) a (C2b) nahraženy podmínkami

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \alpha_i = 0, \tag{$C2c$}$$

$$\rho_i(s_i) \ge \underline{\rho} \quad \Rightarrow \quad \alpha_i = 1. \tag{$C2d$}$$

Poznámka 245. Podmínka (C3b) se obvykle realizuje tak, že

$$\underline{\rho} \le \rho_i(s_i) \le \overline{\rho} \quad \Rightarrow \quad \sigma_{i+1} = \sigma_i, \tag{$C3c$}$$

$$\rho_i(s_i) > \overline{\rho} \quad \Rightarrow \quad \sigma_{i+1} = \min(\sigma_i, \max(\underline{\beta}\sigma_i, \|g_i\|)), \tag{$C3d$}$$

kde $0 < \underline{\rho} < \overline{\rho} < 1$ a $0 < \underline{\beta} < 1$. Nerovnost na levé straně (C3b) lze zapsat ve tvaru

$$F(x_i) - F(x_{i+1}) \ge -\underline{\rho} C_i(s_i). \tag{619}$$

Poznámka 246. Při vyšetřování metod kubické regularizace budeme používat označení

$$\begin{aligned} N_1 &= \{i \in N : \rho_i(s_i) < \underline{\rho}\}, \\ N_2 &= \{i \in N : \rho_i(s_i) \geq \underline{\rho}\}, \\ N_3 &= \{i \in N : \rho_i(s_i) > \overline{\rho}\}. \end{aligned}$$

Jelikož $0 \leq \underline{\rho} < \overline{\rho}$, platí $N_3 \subset N_2$.

Lemma 69. *Nechť $\|g_i\| > 0$, $\|B_i\| > 0$ a $\sigma_i > 0$. Pak platí*

$$-C_i(s_i(\alpha^*)) \geq \frac{\|g_i\|^2}{2(\|B_i\| + \sqrt{\sigma_i \|g_i\|})} \geq \frac{\|g_i\|^2}{4} \min\left(\frac{1}{\|B_i\|}, \frac{1}{\sqrt{\sigma_i \|g_i\|}}\right). \tag{620}$$

Důkaz Pro zjednodušení zápisu budeme index i vynechávat. Nechť $s(\alpha) = -\alpha g$, kde $\alpha > 0$. Pak můžeme psát

$$-C(s(\alpha)) = -g^T s(\alpha) - \frac{1}{2} s(\alpha)^T B s(\alpha) - \frac{1}{3}\sigma \|s(\alpha)\|^3 \geq \alpha \|g\|^2 \left(1 - \frac{1}{2}\alpha \|B\| - \frac{1}{3}\alpha^2 \sigma \|g\|\right).$$

Označme $\tilde{\alpha}$ hodnotu, která maximalizuje výraz na pravé straně této nerovnosti. Tuto hodnotu získáme vynulováním první derivace, což dává

$$-\|g\|^2 \left(-1 + \tilde{\alpha}\|B\| + \tilde{\alpha}^2 \sigma \|g\|\right) = 0.$$

Kladným kořenem této kvadratické rovnice je číslo

$$\tilde{\alpha} = \frac{1}{2\sigma\|g\|}\left(-\|B\| + \sqrt{\|B\|^2 + 4\sigma\|g\|}\right) = \frac{2}{\|B\| + \sqrt{\|B\|^2 + 4\sigma\|g\|}}.$$

Jelikož pro libovolná kladná čísla a, b platí $\sqrt{a^2 + b^2} \leq a + b$, můžeme psát

$$\tilde{\alpha} = \frac{2}{\|B\| + \sqrt{\|B\|^2 + 4\sigma\|g\|}} \geq \frac{1}{\|B\| + \sqrt{\sigma\|g\|}} \triangleq \bar{\alpha},$$

takže

$$\begin{aligned} -C(s(\alpha^*)) \geq -C(s(\bar{\alpha})) &= \bar{\alpha}\|g\|^2 \left(1 - \frac{1}{2}\frac{\|B\|}{\|B\| + \sqrt{\sigma\|g\|}} - \frac{1}{3}\frac{\sigma\|g\|}{(\|B\| + \sqrt{\sigma\|g\|})^2}\right) \\ &= \bar{\alpha}\|g\|^2 \, \frac{6(\|B\| + \sqrt{\sigma\|g\|})^2 - 3\|B\|(\|B\| + \sqrt{\sigma\|g\|}) - 2\sigma\|g\|}{6(\|B\| + \sqrt{\sigma\|g\|})^2} \\ &\geq \bar{\alpha}\|g\|^2 \, \frac{3\|B\|^2 + 6\|B\|\sqrt{\sigma\|g\|} + 3\sigma\|g\|}{6(\|B\| + \sqrt{\sigma\|g\|})^2} = \frac{1}{2}\bar{\alpha}\|g\|^2 \\ &= \frac{\|g\|^2}{2(\|B\| + \sqrt{\sigma\|g\|})}. \end{aligned}$$

□

Lemma 70. *Nechť $\|g_i\| > 0$, $\|B_i\| > 0$ a $\sigma_i > 0$. Pak pokud $C_i(s_i) \leq 0$, platí*

$$\|s_i\| \leq \frac{3}{\sigma_i} \max(\|B_i\|, \sqrt{\sigma_i \|g_i\|}). \tag{621}$$

Důkaz Platí

$$\begin{aligned} C(s) &= s^T g + \frac{1}{2}s^T Bs + \frac{1}{3}\sigma\|s\|^3 \geq -\|g\|\|s\| - \frac{1}{2}\|B\|\|s\|^2 + \frac{1}{3}\sigma\|s\|^3 \\ &= \left(\frac{1}{9}\sigma\|s\|^3 - \|g\|\|s\|\right) + \left(\frac{2}{9}\sigma\|s\|^3 - \frac{1}{2}\|B\|\|s\|^2\right). \end{aligned}$$

První závorka je kladná, pokud $\|s\| > 3\sqrt{\|g\|/\sigma}$, a druhá, pokud $\|s\| > 3\|B\|/\sigma > 9\|B\|/(4\sigma)$. Výraz $C(s)$ je tedy kladný, pokud

$$\|s\| > \frac{3}{\sigma}\max(\|B\|, \sqrt{\sigma\|g\|}),$$

což dokazuje tvrzení lemmatu. □

Lemma 71. *Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklad F4 a s_i je vektor určený podle (C1). Pak pokud $\|B_i\| \leq \overline{B}$ a*

$$\sqrt{\sigma_i\|g_i\|} \geq \frac{18}{(1-\underline{\rho})\underline{\nu}}(\overline{G} + \overline{B}) \tag{622}$$

platí $\rho_i(s_i) \geq \underline{\rho}$, takže $i \in N_2$.

Důkaz Z (622) plyne, že $\sqrt{\sigma\|g\|} \geq \overline{B}$, což spolu s (620), (621) a (C1) dává

$$-C(s) \geq \underline{\nu}\frac{\|g\|^2}{4\sqrt{\sigma\|g\|}} = \underline{\nu}\frac{\|g\|}{4}\sqrt{\frac{\|g\|}{\sigma}}, \qquad \|s\|^2 \leq 9\frac{\|g\|}{\sigma} \tag{623}$$

Podmínku $\rho(s) \geq \underline{\rho}$ můžeme zapsat ve tvaru $F_+ - F - C(s) \leq (\underline{\rho} - 1)C(s)$. Ale

$$F_+ - F - C(s) \leq s^T g + \frac{1}{2}\overline{G}\|s\|^2 - s^T g + \frac{1}{2}\overline{B}\|s\|^2 - \frac{\sigma}{3}\|s\|^3 \leq \frac{1}{2}(\overline{G} + \overline{B})\|s\|^2,$$

což spolu s (622), (623) a (C1) dává

$$F_+ - F - C(s) \leq \frac{1}{2}\left(\frac{(1-\underline{\rho})\underline{\nu}}{18}\sqrt{\sigma\|g\|}\right)\left(9\frac{\|g\|}{\sigma}\right) = (1-\underline{\rho})\underline{\nu}\frac{\|g\|}{4}\sqrt{\frac{\|g\|}{\sigma}} \leq (\underline{\rho} - 1)C(s)$$

□

Věta 149. *(globální konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou kubické regularizace taková, že $\|B_i\| \leq \overline{B}$, $i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F4. Pak platí*

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Důkaz Předpokládejme, že existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$. Pak podle (C3a)–(C3b) a podle lemmatu 71 platí

$$\sigma_i \leq \max\left(\sigma_1, \frac{\overline{\gamma}}{\underline{\varepsilon}}\left(\frac{18}{(1-\underline{\rho})\underline{\nu}}(\overline{G} + \overline{B})\right)^2\right) \triangleq \overline{\sigma} \qquad \forall i \in N. \tag{624}$$

Z této nerovnosti plyne, že množina N_2 je nekonečná (pokud $i \in N_1$ $\forall i \geq k$, pak z (C3a) plyne $\sigma_i \to \infty$). Použijeme-li (620), (624) a (C1), můžeme pro $i \in N_2$ psát

$$F_i - F_{i+1} \geq -\underline{\rho}C(s_i) \geq -\underline{\rho}\,\underline{\nu}C(s_i(\alpha^*)) \geq \frac{\underline{\rho}\,\underline{\nu}\|g_i\|}{4}\min\left(\frac{\|g_i\|}{\|B_i\|}, \sqrt{\frac{\|g_i\|}{\sigma_i}}\right) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{4}\min\left(\frac{\underline{\varepsilon}}{\overline{B}}, \sqrt{\frac{\underline{\varepsilon}}{\overline{\sigma}}}\right)$$

takže

$$F_1 - \underline{F} \geq F_1 - \lim_{i\to\infty} F_i = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i\in N_2}(F_i - F_{i+1}) \geq \sum_{i\in N_2} \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{4} \min\left(\frac{\underline{\varepsilon}}{\overline{B}}, \sqrt{\frac{\underline{\varepsilon}}{\overline{\sigma}}}\right)$$

což je spor, neboť množina N_2 je nekonečná a tudíž i výraz na pravé straně je nekonečný. □

Věta 150. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná striktní metodou kubické regularizace takovou, že $\|B_i\| \leq \overline{B}$ $\forall i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1 a F3. Pak platí*

$$\lim_{i\to\infty} \|g_i\| = 0.$$

Důkaz Důkaz této věty se liší od důkazu věty 119 pouze technickými detaily. Podle věty 149 platí

$$\liminf_{i\to\infty} \|g_i\| = 0. \tag{625}$$

Předpokládejme, že

$$\limsup_{i\to\infty} \|g_i\| > \underline{\varepsilon} > 0.$$

Za tohoto předpokladu je množina N_2 nekonečná a obsahuje nekonečnou podmnožinu $\overline{N}_2 \subset N_2$ takovou, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in \overline{N}_2$ (kdyby N_2 byla konečná, existoval by podle (C2c) index $k \in N$ takový, že $x_i = x_k$ $\forall i \geq k$ a podle (625) též $\|g_i\| = 0$ $\forall i \geq k$). Označme

$$\overline{N}_2 = \{k_1, k_2, k_3, \dots\}.$$

Jelikož posloupnost $F(x_{k_j})$, $j \in N$, je podle (C2) nerostoucí a podle předpokladu F1 zdola omezená, má tato posloupnost limitu. Existuje tedy index $m \in N$ takový, že

$$F(x_{k_j}) - F(x_{k_{j+1}}) < \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}^2}{48} \min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{G}}\right), \quad \forall j \geq m. \tag{626}$$

Nechť $j \geq m$ a l_j je největší index takový, že $k_j \leq l_j < k_{j+1}$, přičemž pro $k_j \leq l \leq l_j$ platí $\|g_l\| \geq \underline{\varepsilon}/2$ (takový index existuje, neboť pro $l = k_j \in \overline{N}_2$ platí $\|g_l\| \geq \underline{\varepsilon} > \underline{\varepsilon}/2$). Pak lze podle (619) a (620) psát

$$F(x_l) - F(x_{l+1}) \geq -\underline{\rho} C_l(s_l) \geq \underline{\rho}\,\underline{\nu} \frac{\|g_l\|}{4} \min\left(\frac{\|g_l\|}{\|B_l\|}, \sqrt{\frac{\|g_l\|}{\sigma_l}}\right) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{8} \min\left(\frac{\underline{\varepsilon}}{2\overline{B}}, \sqrt{\frac{\|g_l\|}{\sigma_l}}\right), \quad \forall k_j \leq l \leq l_j,$$

což spolu s (626) dává

$$\begin{aligned} \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}^2}{48} \min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{G}}\right) &> F(x_{k_j}) - F(x_{k_{j+1}}) \geq F(x_{k_j}) - F(x_{l_j+1}) \\ &= \sum_{l=k_j}^{l_j}(F(x_l) - F(x_{l+1})) \geq \frac{\underline{\rho}\,\underline{\nu}\,\underline{\varepsilon}}{8} \sum_{l=k_j}^{l_j} \min\left(\frac{\underline{\varepsilon}}{2\overline{B}}, \sqrt{\frac{\|g_l\|}{\sigma_l}}\right). \end{aligned} \tag{627}$$

Porovnáme-li obě strany této nerovnosti, vidíme, že případ, kdy $\sqrt{\|g_l\|/\sigma_l} \geq \underline{\varepsilon}/(2\overline{B})$ nemůže pro $k_j \leq l \leq l_j$ nastat (v opačném případě by pravá strana nebyla menší než levá). Platí tedy $\sqrt{\|g_l\|/\sigma_l} < \underline{\varepsilon}/(2\overline{B})$ pro $k_j \leq l \leq l_j$, neboli

$$\sqrt{\sigma_l \|g_l\|} > \frac{2\overline{B}\|g_l\|}{\underline{\varepsilon}} \geq \overline{B}$$

(neboť $\|g_l\| > \underline{\varepsilon}/2$ pro $k_j \le l \le l_j$), což spolu s (621) dává $\|s_l\| \le 3\sqrt{\|g_l\|/\sigma_l}$, takže podle (627) platí

$$\sum_{l=k_j}^{l_j} \|s_l\| \le 3 \sum_{l=k_j}^{l_j} \sqrt{\frac{\|g_l\|}{\sigma_l}} < \frac{\underline{\varepsilon}}{2} \min\left(\frac{1}{\overline{B}}, \frac{1}{\overline{G}}\right) \le \frac{\underline{\varepsilon}}{2\overline{G}}.$$

Použijeme-li tuto nerovnost spolu s nerovností (15), dostaneme

$$\|g(x_{k_j}) - g(x_{l_j+1})\| \le \overline{G}\|x_{k_j} - x_{l_j+1}\| \le \overline{G} \sum_{l=k_j}^{l_j} \|s_l\| < \frac{\underline{\varepsilon}}{2}.$$

Jelikož posloupnost N_2 je nekonečná a platí (625), musí existovat index $j \ge m$ takový, že $l_j + 1 < k_{j+1}$ (a tedy $\|g(x_{l_j+1})\| < \underline{\varepsilon}/2$). Pak podle toho co jsme dokázali platí

$$\|g(x_{k_j})\| \le \|g(x_{l_j+1})\| + \|g(x_{k_j}) - g(x_{l_j+1})\| < \frac{\underline{\varepsilon}}{2} + \frac{\underline{\varepsilon}}{2} = \underline{\varepsilon},$$

což je ve sporu s předpokladem, že $\|g_{k_j}\| \ge \underline{\varepsilon}\ \forall k_j \in \overline{N}_2$. □

Abychom dokázali superlineární konvergenci metod kubické regularizace, je třeba zavést předpoklady podobné předpokladům (544). První předpoklad se týká přesnosti určení lokálního minima funkce $C_i(s)$. Tuto přesnost lze vyjádřit nerovností

$$\|\nabla_s C_i(s_i)\| \le \overline{\omega}_i \|g_i\|, \tag{628}$$

kde $\nabla_s C_i(s)$ je gradient funkce $C_i(s)$ vzhledem k vektoru s a $0 \le \overline{\omega}_i \le \overline{\omega} < 1$. Druhý předpoklad se týká chování matic B_i, $i \in N$, a má stejný tvar jako (544).

Věta 151. *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou kubické regularizace takovou, že $\|B_i\| \le \overline{B}\ \forall i \in N$. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje předpokladům F4 a F5. Nechť platí*

$$\lim_{i\to\infty} \overline{\omega}_i = 0, \tag{629}$$

$$\lim_{i\to\infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0. \tag{630}$$

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz Nechť $0 < \underline{G} < \underline{\lambda}(G^*) \le \overline{\lambda}(G^*) < \overline{G}$. Stejným způsobem jako v části (a) důkazu věty 122 lze ukázat, že existuje index $k_1 \in N$ takový, že

$$\|g_i\| \ge \frac{1}{2}\underline{G}\|s_i\|, \quad \text{a} \quad s_i^T B_i s_i \ge \underline{G}\|s_i\|^2 \qquad \forall i \ge k_1. \tag{631}$$

(b) Ukážeme, že existují index $k_2 \ge k_1$ a číslo $\overline{\sigma}$ tak, že $i \in N_2$ a $\sigma_i \le \overline{\sigma}\ \forall i \ge k_2$. Pokud $\sqrt{\sigma_i\|g_i\|} \ge 18(\overline{G} + \overline{B})/(\underline{\nu}(1 - \underline{\rho}))$, platí podle lemmatu 622 $i \in N_2$. V opačném případě použitím (C1), (620) a (631) dostaneme

$$-C_i(s_i) \ge \underline{\nu}\frac{\|g_i\|^2}{4} \min\left(\frac{1}{\overline{B}}, \frac{1}{\sqrt{\sigma_i\|g_i\|}}\right) \ge \underline{\nu}\frac{\underline{G}^2}{16} \min\left(\frac{1}{\overline{B}}, \frac{\underline{\nu}(1-\underline{\rho})}{18(\overline{G}+\overline{B})}\right) \|s_i\|^2 \triangleq \frac{1}{2}\underline{C}\|s_i\|^2,$$

z čehož stejně jako v části (b) důkazu věty 122 plyne, že $\rho_i(s_i) \to 1$ a jelikož $\underline{\rho} < 1$, existuje index $k_2 \ge k_1$ takový, že $\rho_i(s_i) > \underline{\rho}\ \forall i \ge k_2$. Jelikož σ_i se pro $i \in N_2$ nemůže zvětšovat, můžeme položit $\overline{\sigma} = \sigma_{k_2}$.

(c) Označme

$$\tilde{G}_i = \int_0^1 G(x_i + \lambda s_i)\mathrm{d}\lambda.$$

Pak pro $i \in N_2$ platí $g_{i+1} - g_i = g(x_i + s_i) - g_i = \tilde{G}_i s_i$ a ze spojitosti druhých derivací plyne, že $\|\tilde{G}_i - G_i\| \to 0$, pokud $i \to \infty$. Jelikož $\nabla_s C_i(s_i) = g_i + B_i s_i + \sigma_i \|s_i\| s_i$ (důsledek 22), můžeme podle (628) psát

$$\begin{aligned} \|g_{i+1}\| &\leq \|g_{i+1} - \nabla_s C_i(s_i)\| + \|\nabla_s C_i(s_i)\| = \|g_{i+1} - g_i - B_i s_i - \sigma \|s_i\| s_i\| + \|\nabla_s C_i(s_i)\| \\ &\leq \|(\tilde{G}_i - G_i)s_i\| + \|(G_i - B_i)s_i\| + \overline{\sigma}\|s_i\|^2 + \overline{\omega}_i \|g_i\|. \end{aligned}$$

Dále z $g_{i+1} - g_i = \tilde{G}_i s_i$ plyne, že $\|g_i\| = \|g_{i+1} - \tilde{G}_i s_i\| \leq \|g_{i+1}\| + \overline{G}\|s_i\|$. Dosadíme-li tento vztah do předchozí nerovnosti a použijeme-li (631), dostaneme

$$\begin{aligned} (1 - \overline{\omega})\|g_{i+1}\| &\leq \left(\|(\tilde{G}_i - G_i)\| + \frac{\|(G_i - B_i)s_i\|}{\|s_i\|} + \overline{G}\,\overline{\omega}_i \right) \|s_i\| + \overline{\sigma}\|s_i\|^2 \\ &= (o(1) + \overline{\sigma}\|s_i\|)\|s_i\| \leq \frac{2}{\underline{G}} \left(o(1) + \frac{2}{\underline{G}} \overline{\sigma}\|g_i\| \right) \|g_i\|. \end{aligned}$$

(neboť $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$). Jelikož $x_i \to x^*$ a první derivace jsou spojité, platí $\|g_i\| = o(1)$, takže

$$\frac{\|g_{i+1}\|}{\|g_i\|} = \frac{2}{\underline{G}(1 - \overline{\omega})} o(1) \to 0.$$

□

7.2 Optimální metody kubické regularizace

Definice 50. *Optimální metody kubické regularizace používají směrový vektor*

$$s_i^* = \arg \min_{s \in R^n} C_i(s), \tag{632}$$

Abychom mohli zformulovat podmínky pro extrém funkce $C_i(s)$, potřebujeme znát její gradient a Hessovu matici.

Lemma 72. *Pro libovolný vektor $x \neq 0$ platí*

$$\frac{1}{3}\nabla \|x\|^3 = x\|x\|, \qquad \frac{1}{3}\nabla^2 \|x\|^3 = \nabla(x\|x\|) = \|x\| \left(I + \frac{xx^T}{\|x\|^2} \right),$$

kde I je jednotková matice řádu n.

Důkaz Podle lemmatu 64 platí

$$\frac{1}{3}\frac{\partial}{\partial x_k}\|x\|^3 = \|x\|^2 \frac{\partial}{\partial x_k}\|x\| = x_k \|x\|,$$

$$\frac{1}{3}\frac{\partial^2}{\partial x_k \partial x_l}\|x\|^3 = \frac{\partial}{\partial x_l}(x_k\|x\|) = \delta_{kl}\|x\| + x_k \frac{\partial}{\partial x_l}\|x\| = \delta_{kl}\|x\| + \frac{x_k x_l}{\|x\|} = \|x\| \left(\delta_{kl} + \frac{x_k x_l}{\|x\|^2} \right),$$

kde $\delta_{kl} = 1$, pokud $k = l$, a $\delta_{kl} = 0$, pokud $k \neq l$. □

Důsledek 22. *Gradient a Hessovu matici kubické funkce (617) určíme podle vzorců*

$$\nabla_s C_i(s) = g_i + B_i s + \sigma_i \|s\| s, \qquad \nabla_s^2 C_i(s) = B_i + \sigma_i \|s\| \left(I + \frac{ss^T}{\|s\|^2} \right).$$

Je-li směrový vektor s_i vybrán tak že $(B_i + \lambda_i I)s_i + g_i = 0$, platí $\nabla_s C_i(s_i) = (\sigma_i \|s_i\| - \lambda_i)s_i$.

Důkaz Zřejmě $C_i(s) = Q_i(x) + (\sigma_i/3)\|s\|^3$, kde $Q_i(s)$ je kvadratická funkce (516), pro kterou platí $\nabla_s Q_i(s) = g_i + B_i s$ a $\nabla_s^2 Q_i(s) = B_i$. Použijeme-li tento fakt a lemma 72, dostaneme dokazované tvrzení □

Věta 152. *Vektor $s_i^* \in R^n$ je řešením úlohy (632) právě tehdy, je-li řešením soustavy rovnic*

$$(B_i + \lambda_i^* I)s_i^* + g_i = 0, \tag{633}$$

kde $\lambda_i^* = \sigma_i\|s_i^*\|$ a matice $B_i + \lambda_i^* I$ je pozitivně semidefinitní.

Důkaz (a) Nejprve dokážeme nutnost. Podle důsledku 22 a věty 3 lze nutné podmínky pro to, aby vektor s_i^* byl lokálním minimem funkce $C_i(s)$, zapsat ve tvaru

$$\nabla_s C_i(s_i^*) = g_i + B_i s_i^* + \sigma_i\|s_i^*\|s_i^* = g_i + (B_i + \lambda_i^* I)s_i^* = 0, \qquad \lambda_i^* = \sigma_i\|s_i^*\|,$$

což je právě podmínka (633), a

$$\nabla_s^2 C_i(s_i^*) = B_i + \sigma_i\|s_i^*\|\left(I + \frac{s_i^*(s_i^*)^T}{\|s_i^*\|^2}\right) = B_i + \lambda_i^* I + \lambda_i^*\frac{s_i^*(s_i^*)^T}{\|s_i^*\|^2} \succeq 0.$$

Zbývá dokázat, že matice $B_i + \lambda_i^* I$ je pozitivně semidefinitní, čili že $v^T(B_i + \lambda_i^* I)v \geq 0$ pro libovolný vektor $v \in R^n$. Pokud $v^T s_i^* = 0$, je to zřejmé, neboť v tomto případě platí $v^T(B_i + \lambda_i^* I)v = v^T\nabla_s^2 C_i(s_i^*)v$ a matice $\nabla_s^2 C_i(s_i^*)$ je pozitivně semidefinitní. Nechť tedy $v^T s_i^* \neq 0$. Pak přímka $s_i^* + \alpha v$, $\alpha \in R$, protne kouli o poloměru $\|s_i^*\|$ v dalším bodě $s \neq s_i^*$. Podobně jako v důkazu věty 124 tedy existuje vektor $s \in R^n$ takový, že $\|s\| = \|s_i^*\|$ přičemž $s - s_i^* = \alpha v$, $\alpha \neq 0$. Protože $(B_i + \lambda_i^* I)s_i^* + g_i = 0$, můžeme psát

$$\begin{aligned} C_i(s) - C_i(s_i^*) &= g_i^T(s - s_i^*) + \frac{1}{2}s^T B_i s - \frac{1}{2}(s_i^*)^T B_i s_i^* + \frac{\sigma_i}{3}\|s\|^3 - \frac{\sigma_i}{3}\|s_i^*\|^3 \\ &= (s_i^*)^T(B_i + \lambda_i^* I)(s_i^* - s) + \frac{1}{2}s^T B_i s - \frac{1}{2}(s_i^*)^T B_i s_i^* + \frac{\sigma_i}{3}(\|s\|^3 - \|s_i^*\|^3) \\ &= \frac{1}{2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) + \frac{1}{2}\lambda_i^*(\|s_i^*\|^2 - \|s\|^2) - \frac{\sigma_i}{3}(\|s_i^*\|^3 - \|s\|^3) \end{aligned} \tag{634}$$

a protože $\|s\| = \|s_i^*\|$, dostaneme

$$v^T(B_i + \lambda_i^* I)v = \frac{1}{\alpha^2}(s_i^* - s)^T(B_i + \lambda_i^* I)(s_i^* - s) = \frac{1}{\alpha^2}(C_i(s) - C_i(s_i^*)) \geq 0.$$

(b) Nyní dokážeme postačitelnost. Je li matice $B_i + \lambda_i^* I$ pozitivně semidefinitní, můžeme podle (634) psát

$$\begin{aligned} C_i(s) - C_i(s_i^*) &\geq \frac{1}{2}\lambda_i^*(\|s_i^*\|^2 - \|s\|^2) - \frac{\sigma_i}{3}(\|s_i^*\|^3 - \|s\|^3) \\ &= \sigma_i\left(\frac{1}{2}\|s_i^*\|(\|s_i^*\|^2 - \|s\|^2) - \frac{1}{3}(\|s_i^*\|^3 - \|s\|^3)\right) \geq 0, \end{aligned}$$

neboť $\lambda_i^* = \sigma_i\|s_i^*\|$ a kubická funkce $\varphi(t) = a(a^2 - t^2)/2 - (a^3 - t^3)/3$ je pro $a \geq 0$ a $t \geq 0$ nezáporná (derivováním zjistíme, že $\varphi(t)$ nabývá svého lokálního minima $\varphi(t) = 0$ v bodě $t = a$ a lokálního maxima $\varphi(t) = a^3/6$ v bodě $t = 0$). □

Některé dobré vlastnosti optimálních metod kubické regularizace zůstanou zachovány i když řešíme úlohu (632) pouze přibližně. Proto zavádíme pojem kvazioptimálních metod kubické regularizace.

Definice 51. *Kvazioptimální metody kubické regularizace použijí místo podmínky (C1) podmínku*

$$C_i(s_i) \leq \underline{\nu} C_i(s_i^*) \tag{635}$$

s $0 < \underline{\nu} \leq 1$, kde vektor s_i^* je řešením úlohy (632). Poznamenejme, že z (635) plybe (C1), neboť podle definice 50 platí $C_i(s_i^*) \leq C_i(s_i(\alpha^*))$.

Při výpočtu optimálního směrového vektoru $s_i^* \in R^n$, vyhovujícího podmínkám uvedeným ve větě 152, je třeba opakovaně řešit soustavu rovnic (633) (je to ukázáno v oddílu 7.3). Proto je účelné, aproximovat vektor s_i^* vektorem s_i, který realizuje minimum funkce $C_i(s)$ na nějakém menším podprostoru $\mathcal{L}(Z_i)$, kde $Z_i \in R^{n\times m}$, $m < n$, je matice jejíž sloupce tvoří ortonormální bázi v $\mathcal{L}(Z_i)$. Tento vektor je řešením úlohy

$$s_i = \arg \min_{s\in\mathcal{L}(Z_i)} C_i(s), \tag{636}$$

Věta 153. *Vektor $s_i \in R^n$ je řešením úlohy (636) právě tehdy, platí-li $s_i = Z_i\tilde{s}_i$, kde vektor $\tilde{s}_i$ je řešením úlohy*

$$\tilde{s}_i = \arg \min_{\tilde{s}\in R^m} \tilde{C}_i(s), \qquad \tilde{C}_i(\tilde{s}) = g_i^T Z_i\tilde{s} + \frac{1}{2}\tilde{s}^T Z_i^T B_i Z_i\tilde{s} + \frac{1}{3}\sigma_i\|\tilde{s}\|^3, \tag{637}$$

neboli platí-li

$$(Z_i^T B_i Z_i + \lambda_i I)\tilde{s}_i + Z_i^T g_i = 0, \tag{638}$$

kde $\lambda_i = \sigma_i\|\tilde{s}_i\| = \sigma_i\|s_i\|$ a matice $Z_i^T B_i Z_i + \lambda_i I$ je pozitivně semidefinitní.

Důkaz Jelikož matice Z_i má ortonormální sloupce, platí $\|\tilde{s}\| = \|Z_i s\| = \|s\|$, takže dosadíme-li $s = Z_i\tilde{s}$ do (636) dostaneme (637). Na redukovanou úlohu (637) lze aplikovat větu 152, čímž dostaneme podmínky uvedené ve větě 153. □

Důsledek 23. *Je-li vektor s_i řešením úlohy (636), platí*

$$g_i^T s_i + s_i^T B_i s_i + \sigma_i\|s_i\|^3 = 0 \quad \Rightarrow \quad -C_i(s_i) = \frac{1}{2}s_i^T B_i s_i + \frac{3}{2}\sigma_i\|s_i\|^3, \tag{639}$$

$$s_i^T B_i s_i + \sigma_i\|s_i\|^3 \geq 0 \quad \Rightarrow \quad -C_i(s_i) \geq \frac{1}{6}\sigma_i\|s_i\|^3 \tag{640}$$

Důkaz Vynásobíme li rovnici (638) zleva transponovaným vektorem $\tilde{s}_i$ a položíme-li $s_i = Z_i\tilde{s}_i$, dostaneme rovnost na levé straně (639). Dosadíme-li tuto rovnosr do (617), dostaneme pravou rovnost v (639). Z pozitivní semidefinitnosti matice $Z_i^T B_i Z_i + \lambda_i I$ plyne, že $\tilde{s}_i^T(Z_i^T B_i Z_i + \lambda_i I)\tilde{s}_i \geq 0$, což spolu s $s_i = Z_i\tilde{s}_i$, $\|s_i\| = \|\tilde{s}_i\|$ a $\lambda_i = \sigma_i\|s_i\|$ dává nerovnost na levé straně (640). Dosadíme-li tuto nerovnosr do (617), dostaneme pravou nerovnost v (640). □

Poznámka 247. Je-li vektor s_i řešením úlohy (636) a $g_i \in \mathcal{L}(Z_i)$, platí (C1) s $\underline{\nu} = 1$.

Na závěr ukážeme, že norma vektoru

$$s_i(\sigma_i) = \arg \min_{s\in R^n} C_i(s)$$

je klesající funkcí parametru σ_i.

Věta 154. *Nechť $\sigma > 0$ a $s(\sigma) \neq 0$ je vektor, který je globálním minimem funkce $C(s)$. Pak $\|s(\sigma)\|$ je klesající funkcí parametru σ.*

Důkaz Podle lemmatu 64 platí

$$\|s(\sigma)\|' = \frac{s^T(\sigma)s'(\sigma)}{\|s(\sigma)\|}.$$

Je-li vektor $s(\sigma)$ globálním minimem funkce $C(s)$, jsou podle věty 633 matice $B + \sigma\|s(\sigma)\|I$ a $\nabla^2 C(s)$ pozitivně definitní a platí $(B + \sigma\|s(\sigma)\|I)s(\sigma) + g = 0$. Derivováním této rovnosti a použitím lemmatu 64 dostaneme

$$(B + \sigma\|s(\sigma)\|I)s'(\sigma) + \|s(\sigma)\|s(\sigma) + \frac{s(\sigma)s(\sigma)^T}{\|s(\sigma)\|}s'(\sigma) = 0,$$

neboli

$$(s'(\sigma))^T\nabla^2 C(s(\sigma))s'(\sigma) + \|s(\sigma)\|s^T(\sigma)s'(\sigma) = 0,$$

což dává

$$\|s(\sigma)\|' = \frac{s^T(\sigma)s'(\sigma)}{\|s(\sigma)\|} = -\frac{s'(\sigma)\nabla^2 C(s(\sigma))s'(\sigma)}{\|s(\sigma)\|^2} < 0$$

. □

7.3 Výpočet optimálního směrového vektoru

Podle věty 152 je optimální směrový vektor $s_i^* = s_i(\lambda_i^*)$ řešením soustavy rovnic $(B_i + \lambda_i^* I)s_i^*(\lambda_i) + g_i = 0$ se symetrickou pozitivně definitní maticí $B_i + \lambda_i^* I$, kde $\lambda_i^* = \sigma_i \|s_i^*(\lambda_i)\|$. Tuto nelineární úlohu řešíme přibližně tak, že hledáme číslo $\lambda_i > 0$ takové, že matice $B_i + \lambda_i I$ je pozitivně semidefinitní a

$$\|\nabla_s C_i(s_i)\| = |\sigma_i \|s_i\| - \lambda_i| \, \|s_i\| \leq \overline{\omega}_i \|g_i\|$$

(vzorec (628) a důsledek 22), kde $(B_i + \lambda_i I)s_i + g_i = 0$ a $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$. Tuto nerovnost lze zapsat ve tvaru

$$\frac{\lambda_i}{\sigma_i}\left(1 - \frac{\overline{\omega}_i \|g_i\|}{\lambda_i \|s_i\|}\right) \leq \|s_i\| \leq \frac{\lambda_i}{\sigma_i}\left(1 - \frac{\overline{\omega}_i \|g_i\|}{\lambda_i \|s_i\|}\right),$$

neboli $\underline{\delta}_i \lambda_i/\sigma_i \leq \|s_i\| \leq \overline{\delta}_i \lambda_i/\sigma_i$, což je analogie podmínky $\underline{\delta}\Delta_i \leq \|s_i\| \leq \overline{\delta}\Delta_i$ použité v oddílu 6.1. Poznamenejme, že pro praktické výpočty lze použít podmínku $\underline{\delta}\lambda_i/\sigma_i \leq \|s_i\| \leq \overline{\delta}\lambda_i/\sigma_i$, kde například $\underline{\delta} = 0.9$ a $\overline{\delta} = 1.1$, ale nelze dokázat nerovnost (635), neboť nelze jako v důkazu věty 134 zanedbat člen $(z_i^*)^T(B_i + \lambda_i I)z_i^*$ obsahující neznámý vektor z_i^*. V dalším výkladu se omezíme na jeden konkrétní iterační krok, takže index i budeme vynechávat.

Číslo $\lambda > 0$ vyhovující předpokladům věty 152 lze získat řešením nelineární rovnice ekvivalentní rovnici $\|s(\lambda)\| = \lambda/\sigma$. Přímé použití rovnice $\|s(\lambda)\| = \lambda/\sigma$ není vhodné, neboť funkce $\|s(\lambda)\|$ má póly v bodech, které odpovídají vlastním číslům matice B. Vhodnější (z hlediska omezenosti) je pro tento účel rovnice $\psi(\lambda) = 0$, kde $\psi(\lambda) = \sigma/\lambda - 1/\|s(\lambda)\|$. Tato rovnice se řeší pomocí Newtonovy metody.

Lemma 73. *Nechť $\psi(\lambda) = \sigma/\lambda - 1/\|s(\lambda)\|$, kde $(B + \lambda I)s(\lambda) + g = 0$, matice $B + \lambda I$ je pozitivně definitní a $g \neq 0$. Pak platí*

$$\psi'(\lambda) = -\frac{s(\lambda)^T (B + \lambda I)^{-1} s(\lambda)}{\|s(\lambda)\|^3} - \frac{\sigma}{\lambda^2} < 0$$

a $\psi''(\lambda) \geq 0$ (takže funkce $\psi(\lambda)$ je za daných předpokladů konvexní).

Důkaz Zřejmě $\psi(\lambda) = \phi(\lambda) + \sigma/\lambda$ kde $\phi(\lambda)$ je funkce použitá v oddílu 6.1 s konstantním členem $1/\Delta = 0$ (čili $\phi(\lambda) = -1/\|s(\lambda)\|$). Pak $\psi'(\lambda) = \phi'(\lambda) - \sigma/\lambda^2$ a $\phi''(\lambda) = \phi''(\lambda) + 2\sigma/\lambda^3$, takže dokazované tvrzení plyne z lemmatu 65. □

Poznámka 248. Aby matice $B + \lambda^* I$ byla pozitivně semidefinitní, musí platit $\lambda^* \geq -\lambda_1$, kde λ_1 je nejmenší vlastní číslo matice B. Abychom zjednodušili některé úvahy, budeme tak jako v oddílu 6.1 předpokládat, že nejmenší vlastní číslo λ_1 je jednoduché. Budeme rozlišovat dva případy: regulární případ, kdy $\lambda^* > -\lambda_1$, a singulární případ, kdy $\lambda^* = -\lambda_1$. V regulárním případě mohou nastat dvě možnosti. Pokud $\max(0, -\lambda_1) < \lambda < \lambda^*$ (takže $\sigma\|s(\lambda)\| > \lambda$ a $\psi(\lambda) > 0$) je krok Newtonovy metody $\lambda_+ = \lambda + \Delta\lambda_N$, kde

$$\begin{aligned}
\Delta\lambda_N &= \frac{\lambda^2 \|s(\lambda)\|^3}{\lambda^2 s(\lambda)^T (B + \lambda I)^{-1} s(\lambda) + \sigma \|s(\lambda)\|^3} \left(\frac{\sigma}{\lambda} - \frac{1}{\|s(\lambda)\|}\right) \\
&= \frac{\lambda\left(\|s(\lambda)\| - \dfrac{\lambda}{\sigma}\right)}{\|s(\lambda)\| + \dfrac{\lambda^2}{\sigma} \dfrac{s(\lambda)^T (B + \lambda I)^{-1} s(\lambda)}{\|s(\lambda)\|^2}},
\end{aligned}$$

dobře definován a platí $\lambda < \lambda_+ < \lambda^*$ (plyne to z konvexity funkce $\psi(\lambda)$). Pokud $\lambda^* < \lambda$ (takže $\sigma\|s(\lambda)\| < \lambda$ a $\psi(\lambda) < 0$), platí $\lambda_+ < \lambda^*$ a je třeba zajistit aby byla splněna podmínka $\max(0, -\lambda_1) < \lambda_+$. To lze provést použitím mezí $\underline{\lambda} < \lambda^* < \overline{\lambda}$ aktualizovaných v každém kroku algoritmu (poznámka 253). Newtonovu metodu ukončíme, pokud $\|\nabla_s C(s(\lambda))\| = \|(B + \sigma\|s(\lambda)\|I)s(\lambda) + g\| \leq \overline{\omega}\|g\|$. Řešíme-li rovnici $(B + \lambda I)s(\lambda) + g = 0$ přesně, můžeme iterační proces ukončit, pokud platí $|\lambda - \sigma\|s(\lambda)\|| \|s(\lambda)\| \leq \overline{\omega}\|g\|$.

Poznámka 249. Singulární případ může nastat jedině tehdy, když $\gamma_1 = v_1^T g = 0$, neboť pro $\lambda^* = -\lambda_1$ platí

$$v_1^T g = -v_1^T (B + \lambda^* I) s(\lambda^*) = -v_1^T (B - \lambda_1 I) s(\lambda_1) = 0$$

(používáme vztah $Bv_1 = \lambda_1 v_1$). V singulárním případě nastávají stejné potíže jako při výpočtu optimálního lokálně omezeného ktoku (poznámka 220).

Poznámka 250. Jestliže v regulárním případě platí $-\lambda_1 < \lambda < \lambda^*$, je možné nalézt vhodnější krok $\Delta\lambda = \lambda_+ - \lambda$, než poskytuje Newtonova metoda. Linearizujeme-li funkci $\psi(\lambda)$, dostaneme krok Newtonovy metody $\Delta\lambda_N$ řešením rovnice $\psi(\lambda + \Delta\lambda) \approx \psi(\lambda) + \psi'(\lambda)\Delta\lambda = 0$, neboli

$$\phi(\lambda) + \phi'(\lambda)\Delta\lambda_N + \frac{\sigma}{\lambda} - \frac{\sigma}{\lambda^2}\Delta\lambda_N = 0 \tag{641}$$

(kde $\phi(\lambda) = \psi(\lambda) - \sigma/\lambda = -1/\|s(\lambda)\|$). Linearizujeme-li v rovnici $\psi(\lambda+\Delta\lambda) = \phi(\lambda+\Delta\lambda) + \sigma/(\lambda+\Delta\lambda) = 0$ funkci $\phi(\lambda)$, můžeme psát

$$\phi(\lambda) + \phi'(\lambda)\Delta\lambda_C + \frac{\sigma}{\lambda + \Delta\lambda_C} = 0, \tag{642}$$

což vede na kvadratickou rovnici

$$(\Delta\lambda_C)^2 + \left(\frac{\phi(\lambda)}{\phi'(\lambda)} + \lambda\right)\Delta\lambda_C + \frac{\phi(\lambda)}{\phi'(\lambda)}\left(\lambda - \sigma\|s(\lambda)\|\right) = 0$$

(používáme identitu $\phi(\lambda)\|s(\lambda)\| = -1$). Tato rovnice má řešení

$$\begin{aligned}
\Delta\lambda_C &= -\frac{1}{2}\left(\left(\frac{\phi(\lambda)}{\phi'(\lambda)} + \lambda\right) - \sqrt{\left(\frac{\phi(\lambda)}{\phi'(\lambda)} + \lambda\right)^2 - 4\frac{\phi(\lambda)}{\phi'(\lambda)}\left(\lambda - \sigma\|s(\lambda)\|\right)}\right) \\
&= -\frac{2\frac{\phi(\lambda)}{\phi'(\lambda)}\left(\lambda - \sigma\|s(\lambda)\|\right)}{\frac{\phi(\lambda)}{\phi'(\lambda)} + \lambda + \sqrt{\left(\frac{\phi(\lambda)}{\phi'(\lambda)} - \lambda\right)^2 + 4\frac{\phi(\lambda)}{\phi'(\lambda)}\sigma\|s(\lambda)\|}} \\
&= \frac{2\lambda\left(\|s(\lambda)\| - \frac{\lambda}{\sigma}\right)}{\frac{\lambda}{\sigma} + \tau(\lambda) + \sqrt{(\frac{\lambda}{\sigma} - \tau(\lambda))^2 + 4\tau(\lambda)\|s(\lambda)\|}},
\end{aligned} \tag{643}$$

kde

$$\tau(\lambda) = \frac{\lambda^2}{\sigma}\frac{\phi'(\lambda)}{\phi(\lambda)} = \frac{\lambda^2}{\sigma}\frac{s(\lambda)^T(B + \lambda I)^{-1}s(\lambda)}{\|s(\lambda)\|^2}.$$

Lemma 74. *Nechť jsou splněny předpoklady lemmatu 73, přičemž $-\lambda_1 < \lambda < \lambda^*$ (takže $\sigma\|s(\lambda)\| > \lambda$ a $\psi(\lambda) > 0$). Pak platí $-\lambda_1 < \lambda + \Delta\lambda_N < \lambda + \Delta\lambda_C < \lambda^*$.*

Důkaz Jelikož podíl σ/λ je ryze konvexní funkcí parametru λ, platí

$$\frac{\sigma}{\lambda + \Delta\lambda_C} > \frac{\sigma}{\lambda} - \frac{\sigma}{\lambda^2}\Delta\lambda_C$$

Použijeme-li tuto nerovnost v (642), můžeme psát

$$\phi(\lambda) + \phi'(\lambda)\Delta\lambda_C + \frac{\sigma}{\lambda} - \frac{\sigma}{\lambda^2}\Delta\lambda_C < 0$$

a přihlédneme-li k tomu že $\psi(\lambda) = \phi(\lambda) + \sigma/\lambda$ a že podle lemmatu 73 platí $\psi'(\lambda) = \phi'(\lambda) - \sigma/\lambda^2 < 0$, dostaneme

$$\psi(\lambda) + \psi'(\lambda)\Delta\lambda_C < 0 = \psi(\lambda) + \psi'(\lambda)\Delta\lambda_N \quad \Rightarrow \quad \Delta\lambda_C > \Delta\lambda_N$$

□

Poznámka 251. V singulárním případě nelze použít Newtonovu metodu (z důvodů uvedených v oddílu 6.1). V tomto případě lze vektor $s(\lambda^*)$ vyjádřit ve tvaru $s(\lambda^*) = s + \alpha v_1$, kde s je libovolné řešení rovnice $(B - \lambda_1 I)s = -g$ (které existuje, ale není jediné) a α se vybírá tak aby platilo $\|s(\lambda^*)\| = \|s + \alpha v_1\| = \lambda/\sigma$. Potom

$$(B + \lambda^* I)s(\lambda^*) = (B - \lambda_1 I)s + \alpha(B - \lambda_1 I)v_1 = g.$$

Tento způsob je podkladem pro alternativní krok v případě, že $\psi(\lambda) < 0$, kdy krok Newtonovy metody může selhat. V tomto případě najdeme řešení s rovnice $(B + \lambda I)s + g = 0$ spolu s nějakou aproximací $\tilde{v}_1$ vektoru v_1 a testujeme, zda vektor $s + z = s + \alpha\tilde{v}_1$ takový, že $\|s + z\| = \lambda/\sigma$, vyhovuje podmínce (635). Ukážeme, jak lze získat kvazioptimální krok splňující podmínku (635) s $\underline{\nu} = 1/3$.

Věta 155. *Nechť $\lambda > 0$, $B + \lambda I \succeq 0$ a $\|s + z\| = \lambda/\sigma$, kde $(B + \lambda I)s + g = 0$. Jestliže $\lambda > \lambda^*$ a*

$$z^T(B + \lambda I)z \leq \frac{2}{3}s^T(B + \lambda I)s, \tag{644}$$

vyhovuje vektor $s + z$ podmínce (635) s $\underline{\nu} = 1/3$.

Důkaz Tak jako v důkazu věty 134 platí

$$\begin{aligned} C(s+z) &= Q(s+z) + \frac{\sigma}{3}\|s+z\|^3 \\ &= -\frac{1}{2}\left(s^T(B+\lambda I)s + \lambda(s+z)^T(s+z)\right) + \frac{1}{2}z^T(B+\lambda I)z + \frac{\sigma}{3}\|s+z\|^3 \\ &= -\frac{1}{2}\left(s^T(B+\lambda I)s + \lambda\left(\frac{\lambda}{\sigma}\right)^2\right) + \frac{1}{2}z^T(B+\lambda I)z + \frac{\sigma}{3}\left(\frac{\lambda}{\sigma}\right)^3 \\ &= -\frac{1}{2}s^T(B+\lambda I)s - \frac{\sigma}{6}\left(\frac{\lambda}{\sigma}\right)^3 + \frac{1}{2}z^T(B+\lambda I)z. \end{aligned} \tag{645}$$

Nechť vektor $s^* = s + z^*$ je řešním úlohy (632). Pak $(s + z^*)^T(s + z^*) = (\lambda^*/\sigma)^2$ a $(z^*)^T(B + \lambda I)z^* \geq 0$, takže po dosazení do předchozí rovnosti a po vynecháním kladných členů dostaneme

$$C(s+z^*) \geq -\frac{1}{2}\left(s^T(B+\lambda I)s + \lambda\left(\frac{\lambda^*}{\sigma}\right)^2\right) \geq -\frac{1}{2}s^T(B+\lambda I)s - \frac{\sigma}{2}\left(\frac{\lambda}{\sigma}\right)^3. \tag{646}$$

(neboť $\lambda > \lambda^*$). Je-li splněna nerovnost (644), můžeme psát

$$\begin{aligned} C(s+z) &\leq -\frac{1}{2}s^T(B+\lambda I)s - \frac{\sigma}{6}\left(\frac{\lambda}{\sigma}\right)^3 + \frac{2}{6}s^T(B+\lambda I)s \\ &= -\frac{1}{6}s^T(B+\lambda I)s - \frac{\sigma}{6}\left(\frac{\lambda}{\sigma}\right)^3 \leq \frac{1}{3}C(s+z^*). \end{aligned}$$

□

Poznámka 252. Nechť $s \in R^n$, $v \in R^n$ a $\|s\| < \Delta$. Číslo $\alpha \geq 0$, pro které platí $\|s + \alpha v\| = \lambda/\sigma$, určujeme podle vzorců

$$\alpha = \frac{\sqrt{(v^Ts)^2 + ((\lambda/\sigma)^2 - \|s\|^2)\|v\|^2} - v^Ts}{\|v\|^2} = \frac{(\lambda/\sigma)^2 - \|s\|^2}{\sqrt{(v^Ts)^2 + ((\lambda/\sigma)^2 - \|s\|^2)\|v\|^2} + v^Ts}.$$

První vzorec volíme pokud $v^Ts \leq 0$ a druhý v opačném případě. Oba vzorce se zjednoduší, pokud $\|v\| = 1$. Tyto vzorce lze snadno získat řešením kvadratické rovnice vzniklé roznásobením vztahu $\|s+\alpha v\|^2 = (\lambda/\sigma)^2$.

Poznámka 253. Abychom zabránili selhání Newtonovy metody, je účelné používat a aktualizovat dolní odhad $\underline{\mu}$ pro číslo $-\lambda_1$ a meze $0 \leq \underline{\lambda} < \lambda^* < \overline{\lambda}$. V prvním iteračním kroku Newtonovy metody můžeme jako $\underline{\mu}$ zvolit maximální diagonální prvek matice $-B$. Počáteční meze $\underline{\lambda} < \lambda^* < \overline{\lambda}$ lze určit z vlastností čísla λ^*. Jestliže $(B+\lambda^* I)s(\lambda^*)+g = 0$ (takže $s(\lambda^*)^T(B+\lambda^* I)^2 s(\lambda^*) = \|g\|^2$) a $\|s(\lambda^*)\| = \lambda^*/\sigma$, můžeme psát

$$(\underline{\lambda}(B)+\lambda^*)^2 \left(\frac{\lambda^*}{\sigma}\right)^2 \leq \|g\|^2 \leq (\overline{\lambda}(B)+\lambda^*)^2 \left(\frac{\lambda^*}{\sigma}\right)^2,$$

(používáme extrémní vlastnosti vlastních čísel matice B), neboli

$$(\lambda^* + \underline{\lambda}(B))\lambda^* \leq \sigma\|g\| \leq (\lambda^* + \overline{\lambda}(B))\lambda^*.$$

Řešením těchto dvou kvadratických nerovností dostaneme

$$\frac{1}{2}(\sqrt{\overline{\lambda}^2(B) + 4\sigma\|g\|} - \overline{\lambda}(B) \leq \lambda^* \leq \frac{1}{2}(\sqrt{\underline{\lambda}^2(B) + 4\sigma\|g\|} - \underline{\lambda}(B)$$

(funkce $\sqrt{x^2+a} - x$ je pro $a > 0$ klesaící). Jestliže $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$, platí

$$\underline{\lambda} = \frac{1}{2}(\sqrt{\overline{B}^2 + 4\sigma\|g\|} - \overline{B} \leq \lambda^* \leq \frac{1}{2}(\sqrt{\underline{B}^2 + 4\sigma\|g\|} - \underline{B} = \overline{\lambda}$$

Můžeme položit $\underline{B} = -\|B\|$ a $\overline{B} = \|B\|$, nebo použít jiné odhady pro vlastní čísla matice B, například Gerschgorinovy kruhy. Dolní mez $\underline{\lambda}$ je třeba ještě upravit tak, aby platilo $\underline{\lambda} \geq 0$.

Máme-li k disposici dolní odhad $\underline{\lambda}$ takový, že $\underline{\lambda} < \lambda^* < \lambda$ a platí-li $(\lambda - \underline{\lambda})/\lambda \to 0$, můžeme tvrzení věty 155 podstatně zesílit.

Věta 156. *Nechť jsou splněny předpoklady věty 155 a nechť $\underline{\lambda} < \lambda^* < \lambda$ a $(\lambda - \underline{\lambda})/\lambda \leq (1-\underline{\nu})/(2\underline{\nu})$, kde $0 < \underline{\nu} < 1$. Pak, platí-li*

$$z^T(B+\lambda I)z \leq (1-\underline{\nu})s^T(B+\lambda I)s, \tag{647}$$

vyhovuje vektor $s+z$ podmínce (635).

Důkaz Nechť vektor $s+z^*$ je řešením úlohy (632). Pak $(s+z^*)^T(s+z^*) = (\lambda^*/\sigma)^2$ a $(z^*)^T(B+\lambda I)z^* \geq 0$, což po dosazení do (645) dává

$$\begin{aligned} C(s+z^*) &\geq -\frac{1}{2}\left(s^T(B+\lambda I)s + \lambda\left(\frac{\lambda^*}{\sigma}\right)^2\right) + \frac{\lambda^*}{3}\left(\frac{\lambda^*}{\sigma}\right)^2 \\ &\geq -\frac{1}{2}s^T(B+\lambda I)s - \frac{\lambda}{2}\left(\frac{\lambda^*}{\sigma}\right)^2 + \frac{\underline{\lambda}}{3}\left(\frac{\lambda^*}{\sigma}\right)^2 \\ &\geq -\frac{1}{2}s^T(B+\lambda I)s + \frac{1}{6}(2\underline{\lambda} - 3\lambda)\left(\frac{\lambda}{\sigma}\right)^2 \end{aligned} \tag{648}$$

(neboť $\underline{\lambda} < \lambda^* < \lambda$). Porovnáme-li nerovnost (648) se vztahem (645) vidíme, že $C(s+z) \leq \underline{\nu}C(s+z^*)$ tehdy, když

$$\begin{aligned} \frac{1}{2}z^T(B+\lambda I)z &\leq \frac{1-\underline{\nu}}{2}s^T(B+\lambda I)s + \frac{\lambda}{6}\left(\frac{\lambda}{\sigma}\right)^2 - \frac{\underline{\nu}}{6}(2\underline{\lambda} - 3\lambda)\left(\frac{\lambda}{\sigma}\right)^2 \\ &= \frac{1-\underline{\nu}}{2}s^T(B+\lambda I)s + \frac{1}{6}(\lambda - \underline{\nu}(2\underline{\lambda} - 3\lambda))\left(\frac{\lambda}{\sigma}\right)^2. \end{aligned}$$

Tato nerovnost platí, je-li splněna podmínka (647) a je-li poslední člen na pravé straně nezáporný, neboli je-li splněna nerovnost $\lambda - \underline{\nu}(2\underline{\lambda} - 3\lambda) \geq 0$, což po úpravě dává $(\lambda - \underline{\lambda})/\lambda \leq (1-\underline{\nu})/(2\underline{\nu})$. □

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 16. Data $0 < \underline{\beta} < 1$ (obvykle $\underline{\beta} = 0.1$), $0 < \overline{\omega} < 1$, $0 < \underline{\nu} < 1$ (přesnosti výpočtu) a $\sigma > 0$.

Krok 1 Nechť $\underline{\mu}$ je maximální diagonální prvek matice $-B$ a $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$. Určíme meze $\underline{\lambda}$ a $\overline{\lambda}$ podle poznámky 253. Položíme $\lambda := \max(0,\ \underline{\mu}, \underline{\lambda})$ a $k := 0$.

Krok 2 Položíme $\underline{\lambda} := \max(0,\ \underline{\mu}, \underline{\lambda})$. Jestliže $k > 0$ a $\lambda \leq \underline{\mu}$, položíme $\lambda := \max(\sqrt{\underline{\lambda}\,\overline{\lambda}},\ \underline{\lambda} + \underline{\beta}(\overline{\lambda} - \underline{\lambda}))$.

Krok 3 Určíme Gillův-Murrayův rozklad $B + \lambda I + E = R^T R$ a položíme $k := k + 1$. Je-li $E = 0$ (takže $B + \lambda I \succ 0$), přejdeme na krok 4. V opačném případě určíme vektor $v \in R^n$ takový, že $\|v\| = 1$ a $v^T(B + \lambda I)v < 0$ (věta 29), položíme $\underline{\mu} := \max(\underline{\mu},\ \lambda - v^T(B + \lambda I)v)$, $\underline{\lambda} := \max(\underline{\mu}, \underline{\lambda})$ a přejdeme na krok 2.

Krok 4 Určíme vektor $s \in R^n$ řešením rovnice $R^T R s + g = 0$. Jestliže $|\lambda - \sigma\|s\|\|s\| \leq \overline{\omega}\|g\|$, ukončíme výpočet. Jestliže $\|s\| > \lambda/\sigma$, položíme $\underline{\lambda} := \lambda$ a přejdeme na krok 6. Jestliže $\|s\| < \lambda/\sigma$, položíme $\overline{\lambda} := \lambda$ a přejdeme na krok 5.

Krok 5 Jestliže $(\lambda - \underline{\lambda})/\lambda > (1 - \underline{\nu})/(2\underline{\nu})$, přejdeme na krok 6. V opačném případě Určíme vektor $v \in R^n$ tak, aby tento vektor byl dobrou aproximací vlastního vektoru matice B příslušného vlastnímu číslu $\underline{\lambda}(B)$ a aby platilo $\|v\| = 1$ a $v^T s \geq 0$ (tento vektor lze určit z rozkladu $R^T R$ způsobem, který používají programy knihovny LAPACK). Určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha v\| = \lambda/\sigma$ (poznámka 252). Jestliže $\alpha^2\|Rv\|^2 \leq (1 - \underline{\nu})\|Rs\|^2)$, položíme $s := s + \alpha v$ a ukončíme výpočet. V opačném případě položíme $\underline{\mu} := \max(\underline{\mu},\ \lambda - \|Rv\|^2)$ a přejdeme na krok 6.

Krok 6 Určíme vektor $v \in R^n$ řešením rovnice $R^T v = s$ a položíme $\tau := (\lambda^2/\sigma)(\|v\|/\|s\|)^2$. Položíme $\lambda := \lambda + \Delta\lambda$, kde buď

$$\Delta\lambda = \Delta\lambda_N = \frac{\lambda\left(\|s\| - \dfrac{\lambda}{\sigma}\right)}{\|s\| + \tau}$$

nebo

$$\Delta\lambda = \Delta\lambda_C = \frac{2\lambda\left(\|s\| - \dfrac{\lambda}{\sigma}\right)}{\dfrac{\lambda}{\sigma} + \tau + \sqrt{\left(\dfrac{\lambda}{\sigma} - \tau\right)^2 + 4\tau\|s\|}}$$

a přejdeme na krok 2

8 Metody pro minimalizaci součtu čtverců

V tomto oddílu budeme předpokládat, že minimalizovaná funkce má tvar

$$F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x), \tag{649}$$

kde $f : \mathcal{D}_F \to R^m$ je zobrazení definované na množině $\mathcal{D}_F \subset R^n$ (zobrazení f má stejný definiční obor jako funkce F). Budeme používat označení

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}, \qquad J(x) = \begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1}, & \cdots, & \dfrac{\partial f_1(x)}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \dfrac{\partial f_m(x)}{\partial x_1}, & \cdots, & \dfrac{\partial f_m(x)}{\partial x_n} \end{bmatrix}.$$

pro zobrazení f a jeho Jacobiovu matici. Je li zobrazení $f : \mathcal{D}_F \to R^m$ spojitě diferencovatelné na nějaké otevřené množině $\mathcal{D} \subset \mathcal{D}_F$, pak pro $x \in \mathcal{D}$ platí

$$g(x) = J^T(x) f(x) = \sum_{k=1}^{m} f_k(x) g_k(x), \tag{650}$$

kde $g(x)$ je gradient funkce F v bodě x. Je-li zobrazení f dvakrát spojitě diferencovatelné na $\mathcal{D}$, pak pro $x \in \mathcal{D}$ platí

$$G(x) = J^T(x) J(x) + C(x) = \sum_{k=1}^{m} g_k(x) g_k^T(x) + \sum_{k=1}^{m} f_k(x) G_k(x), \tag{651}$$

kde $G(x)$ je Hessova matice funkce F v bodě x.

Při vyšetřování konvergence metod pro minimalizaci součtu čtverců budeme předpokládat, že funkce F splňuje předpoklad F1 (který je splněn automaticky, neboť $F(x) \geq 0 \ \forall x \in \mathcal{D}_F$) a předpoklad F2. Aby funkce F splňovala předpoklady F3 a F4, je třeba aby tyto předpoklady splňovalo zobrazení f, neboli aby je splňovaly všechny funkce f_k, $1 \leq k \leq m$ (poznámka 5).

Definice 52. *Řekneme, že zobrazení f splňuje předpoklad F3 nebo F4, splňují-li tento předpoklad všechny funkce f_k, $1 \leq k \leq m$, tedy existují-li čísla $\overline{G}_k$, $1 \leq k \leq m$, taková, že pro $1 \leq k \leq m$ platí*

$$\|g_k(x_2) - g_k(x_1)\| \leq \overline{G}_k \|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}, \tag{652}$$

nebo

$$|d^T G_k(x) d| \leq \overline{G}_k \|d\|^2 \quad \forall x \in \mathcal{D} \quad \forall d \in R^n, \tag{653}$$

Podobně zobecníme předpoklady F3* a F4*.

Definice 53. *Řekneme, že zobrazení f splňuje předpoklad F3* nebo F4*, splňují-li tento předpoklad všechny funkce f_k, $1 \leq k \leq m$, tedy existují-li čísla $\overline{G}_k$, $1 \leq k \leq m$, a $\varepsilon > 0$ taková, že pro $1 \leq k \leq m$ platí*

$$\|g_k(x) - g_k(x^*)\| \leq \overline{G}_k \|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon), \tag{654}$$

nebo

$$|d^T G_k(x) d| \leq \overline{G}_k \|d\|^2 \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{655}$$

V tomto oddílu budeme na zobrazení f klást pouze předpoklady F3 a F4 (případně F3* a F4*). Další předpoklady F5 a F6 jsou uvedeny v oddílu 10.4.

Jelikož ve vzorcích (650)–(651) vystupuje Jacobiova matice zobrazení $f : \mathcal{D}_F \to R^m$, je účelné zavést další předpoklady týkající se zobrazení f a Jacobiovy matice J.

Předpoklad J1. *Zobrazení $f : \mathcal{D}_F \to R^m$ je omezené na $\mathcal{D}$, takže existuje konstanta $\overline{f}$ taková, že*

$$\|f(x)\| \leq \overline{f} \quad \forall x \in \mathcal{D}. \tag{656}$$

Předpoklad J3. *Zobrazení $f : \mathcal{D}_F \to R^m$ je lipschitzovské na $\mathcal{D}$, takže existuje konstanta $\overline{J} > 0$ taková, že*

$$\|f(x_2) - f(x_1)\| \leq \overline{J}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{657}$$

Předpoklad J4. *Zobrazení $f \in \mathcal{C}^1 : \mathcal{D} \to R^n$ má omezené Jacobiovy matice na $\mathcal{D}$, takže existuje konstanta $\overline{J} > 0$ taková, že*

$$\|J(x)d\| \leq \overline{J}\|d\| \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{658}$$

Podmínka (658) je ekvivalentní podmínce $\|J(x)\| \leq \overline{J}$ $\forall x \in \mathcal{D}$.

Poznámka 254. Použijeme-li tvrzení 6 a předpoklad J4, můžeme psát

$$\|f(x_2) - f(x_1)\| = \left\| \int_0^1 J(x_1 + \lambda(x_2 - x_1))(x_2 - x_1)\mathrm{d}\lambda \right\| \leq \overline{J}\|x_2 - x_1\|, \tag{659}$$

takže J4 implikuje J3.

Poznámka 255. Podle lemmatu 31 mají matice $J^T(x)J(x)$ a $J(x)J^T(x)$ stejná vlastní čísla , takže platí $\|J^T(x)\| = \|J(x)\|$. Jsou-li splněny předpoklady J3 a J4, můžeme psát

$$\|J^T(x_2)(f(x_2) - f(x_1))\| \leq \overline{J}\|f(x_2) - f(x_1)\| \leq \overline{J}^2\|x_2 - x_1\|. \tag{660}$$

Předpoklad J5. *Zobrazení $f \in \mathcal{C}^1 : \mathcal{D} \to R^n$ je stejnoměrně regulární na $\mathcal{D}$, takže existuje konstanta $\overline{J} > 0$ taková, že*

$$\|J(x)d\| \geq \underline{J}\|d\| \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{661}$$

Podmínka (661) je ekvivalentní podmínce $\|J^{-1}(x)\|^{-1} \geq \underline{J}$ $\forall x \in \mathcal{D}$.

Předpoklad J6. *Zobrazení $f \in \mathcal{C}^1 : \mathcal{D} \to R^n$ má lipschitzovské Jacobiovy matice na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$\|J(x_2) - J(x_1)\| \leq \overline{G}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{662}$$

Poznámka 256. Splňuje-li zobrazení f předpoklad F3, můžeme psát

$$\begin{aligned} \|J(x_2) - J(x_1)\| &\leq \|J(x_2) - J(x_1)\|_F = \sqrt{\sum_{k=1}^{m} \|g_k(x_2) - g_k(x_1)\|^2} \\ &\leq m \max_{1 \leq k \leq m} \|g_k(x_2) - g_k(x_1)\| \leq m \max_{1 \leq k \leq m} \overline{G}_k \|x_2 - x_1\|, \end{aligned}$$

takže F3 implikuje J6 s $\overline{G} = m \max_{1\le k\le m} \overline{G}_k$. Splňuje-li zobrazení f předpoklad J6, můžeme psát

$$\begin{aligned}\overline{G}\,\|x_2 - x_1\| &\geq \|J(x_2) - J(x_1)\| \geq \frac{1}{\sqrt{n}}\|J(x_2) - J(x_1)\|_F \\ &= \frac{1}{\sqrt{n}}\sqrt{\sum_{k=1}^{m}\|g_k(x_2) - g_k(x_1)\|^2} \geq \frac{1}{\sqrt{n}}\max_{1\le k\le m}\|g_k(x_2) - g_k(x_1)\|,\end{aligned}$$

takže J6 implikuje F3 s $\overline{G}_k = \sqrt{n}\,\overline{G}$, $1 \le k \le m$. Předpoklady F3 a J6 jsou tedy vzájemně zaměnitelné.

Poznámka 257. Z úvah použitých v poznámce 255 plyne, že předpoklad J6 implikuje nerovnost

$$\|J^T(x_2) - J^T(x_1)\| \le \overline{G}\|x_2 - x_1\|.$$

Abychom ukázali význam uvedených předpokladů uvedeme analogii tvrzení 3 o střední hodnotě.

Tvrzení 6. *Nechť $f \in \mathcal{C}^1 : \mathcal{D} \to R^m$, $x \in \mathcal{D}$ a $[x, x+d] \subset \mathcal{D}$. Pak platí*

$$f(x+d) = f(x) + \int_0^1 J(x+\lambda d)d\,\mathrm{d}\lambda.$$

Použijeme-li předpoklad J3 nebo tvrzení 6 a předpoklad J4, dostaneme

$$\|f(x+d) - f(x)\| \le \overline{J}\|d\|, \tag{663}$$

$$d^T(f(x+fd) - g(x)) \le \overline{J}\|d\|^2. \tag{664}$$

Použijeme-li tvrzení 6 a předpoklad J5, dostaneme

$$\|f(x+d) - f(x)\| \ge \underline{J}\|d\|, \tag{665}$$

$$d^T(f(x+d) - f(x)) \ge \underline{J}\|d\|^2. \tag{666}$$

Důkaz posledních dvou nerovností:

$$d^T(f(x+d) - f(x)) = \int_0^1 d^T J(x+\lambda d)d\,\mathrm{d}\lambda \ge \int_0^1 \underline{J}\|d\|^2\mathrm{d}\lambda = \underline{J}\|d\|^2,$$

$$\underline{J}\|d\|^2 \le d^T(f(x+d) - f(x)) \le \|d\|\|f(x+d) - f(x)\|.$$

Ukážeme, že předpoklady J1, J4 a F3 nebo J1, J4 a F4, kladené na zobrazení f, zaručují, že jsou splněny předpoklady F3 nebo F4, kladené na funkci $F = (1/2)f^T f$.

Věta 157. *Nechť zobrazení f splňuje předpoklady J1, J4 a F3. Pak je-li množina $\mathcal{D}$ konvexní, splňuje funkce F předpoklad F3. Nechť zobrazení f splňuje předpoklady J1, J4 a F4. Pak funkce F splňuje předpoklad F4.*

Důkaz (a) Nechť zobrazení f splňuje předpoklady J1, J4 a F3 a nechť $x_1 \in \mathcal{D}$, $x_2 \in \mathcal{D}$. Pak podle poznámky 257 je splněn předpoklad J6, takže použitím (660) dostaneme

$$\begin{aligned}\|g(x_2) - g(x_1)\| &= \|J^T(x_2)f(x_2) - J^T(x_1)f(x_1)\| \\ &\le \|J^T(x_2)(f(x_2) - f(x_1))\| + \|(J^T(x_2) - J^T(x_1))f(x_1)\| \\ &\le \overline{J}\|(f(x_2) - f(x_1)\| + \overline{G}\|x_2 - x_1\|\|f(x_1)\| \\ &\le \left(\overline{J}^2 + \overline{G}\,\overline{f}\right)\|x_2 - x_1\|. \end{aligned}\tag{667}$$

(b) Nechť zobrazení f splňuje předpoklady J1, J4 a F4 a nechť $x \in \mathcal{D}$. Pak podle (651) platí

$$\|G(x)\| \leq \|J(x)^T J(x)\| + \sum_{k=1}^{m} |f_k(x)| \|G_k(x)\| \leq \overline{J}^2 + m\overline{f}\,\overline{G}.$$

□

Poznámka 258. Při vyšetřování metod pro minimalizaci součtu čtverců můžeme přímo předpokládat, že funkce $F = (1/2)f^T f$ splňuje předpoklady F3 nebo F4. Musíme však mít na paměti, že to v obecném případě znamená, že zobrazení f splňuje předpoklady J1, J4 a F3 nebo J1, J4 a F4. V některých případech však tato nutnost odpadá. Je-li množina $\mathcal{D}$ omezená a je-li zobrazení f spojitě diferencovatelné na $\overline{\mathcal{D}}$ (uzávěr), splňuje funkce F předpoklad F3. Je-li navíc f dvakrát spojitě diferencovatelné na $\overline{\mathcal{D}}$, splňuje funkce F předpoklad F4.

Poznámka 259. Studujeme-li chování iteračního procesu v okolí limitního bodu $x^* \in R^n$, stačí předpokládat, že zobrazení f je spojitě diferencovatelné v nějakém okolí bodu x^* a pokládat $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$ v předpokladech F3, F6 a J3–J6 nebo používat předpoklady F3*, F6* a J3*–J6*.

Předpoklad J3*. *Zobrazení $f : \mathcal{D} \to R^m$ je klidné v okolí bodu $x^* \in \mathcal{D}$, takže existují čísla $\overline{J} > 0$ a $\varepsilon > 0$ taková, že platí*

$$\|f(x) - f(x^*)\| \leq \overline{J}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{668}$$

Předpoklad J4*. *Zobrazení $f : \mathcal{D} \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}$. Pak pro libovolnou konstantu $\overline{J} > \|J(x^*)\|$ existuje číslo $\varepsilon > 0$ takové, že*

$$\|J(x)d\| \leq \overline{J}\|d\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{669}$$

Předpoklad J5*. *Zobrazení $f : \mathcal{D} \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}$ a Jacobiova matice $J(x^*)$ je regulární. Pak pro libovolnou konstantu $0 < \underline{J} < 1/\|J^{-1}(x^*)\|$ existuje číslo $\varepsilon > 0$ takové, že*

$$\|J(x)d\| \geq \underline{J}\|d\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{670}$$

Předpoklad J6*. *Zobrazení $f : \mathcal{D} \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}$ a jeho Jacobiova matice je klidná v okolí bodu x^*, takže existují čísla $\overline{G} > 0$ a $\varepsilon > 0$ taková, že platí*

$$\|J(x) - J(x^*)\| \leq \overline{G}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{671}$$

Věta 158. *Splňuje-li zobrazení f předpoklady F4* a J4*, splňuje funkce F předpoklad F4*.*

Důkaz Ze spojité diferencovatelnosti zobrazení f v okolí bodu x^* vyplývá, že existuje číslo $\delta > 0$ takové, že $\|f(x)\| \leq 2\|f(x^*)\|$ a $\|J(x)\| \leq 2\|J(x^*)\|$ $\forall x \in \mathcal{B}(x^*, \delta)$. Položme $\overline{f} = 2\|f(x^*)\|$, $\overline{J} = 2\|J(x^*)\|$ a zvolme $0 < \varepsilon \leq \delta$ tak aby byla splněna podmínka (669). Pak v $\mathcal{B}(x^*, \varepsilon)$ platí

$$\|G(x)\| \leq \|J^T(x)J(x)\| + \sum_{k=1}^{m} |f_k(x)| \|G_k(x)\| \leq \overline{J}^2 + m\overline{f}\,\overline{G}.$$

□

Věta 159. *Splňuje-li zobrazení f předpoklady F4*, J5* a platí-li $f(x^*) = 0$, splňuje funkce F předpoklad F5*.*

Důkaz Jelikož $f(x^*) = 0$, existuje číslo $\delta > 0$ takové, že $\|f(x)\| \leq \underline{J}^2/(2m\overline{G})\ \forall x \in \mathcal{B}(x^*, \delta)$. Zvolme $0 < \varepsilon \leq \delta$ tak, aby byly splněny podmínky (669) a (670). Pak v $\mathcal{B}(x^*, \varepsilon)$ platí

$$\begin{aligned} d^T G(x) d &= d^T \left(J^T(x) J(x) + \sum_{k=1}^{n} f_k(x) G_k(x) \right) d \geq \|J(x)d\|^2 - \sum_{k=1}^{n} |f_k(x)| \|G_k(x)\| \|d\|^2 \\ &\geq \left(\underline{J}^2 - \frac{\underline{J}^2}{2m\overline{G}} m\overline{G} \right) \|d\|^2 = \frac{1}{2} \underline{J}^2 \|d\|^2. \end{aligned}$$

□

8.1 Gaussova–Newtonova metoda

Gaussova–Newtonova metoda vznikne z Newtonovy metody tím, že ve výrazu pro Hessovu matici $G(x_i)$ zanedbáme člen $C(x_i)$, takže

$$B_i = J_i^T J_i = \sum_{k=1}^{m} g_k(x_i) g_k^T(x_i),$$

kde B_i je matice, která se používá k určení směrového vektoru (řešením soustavy rovnic $B_i s_i = -g_i$ nebo minimalizací kvadratické funkce $Q_i(s)$ definované vzorcem (516)).

Poznámka 260. Existují dva důvody pro použití takto definované matice B_i:

(1) Úlohy s nulovým reziduem. Nechť $F(x^*) = 0$. Pak z $x_i \to x^*$ plyne $F(x_i) \to F(x^*) = 0$ a tedy $f_k(x_i) \to 0$, $1 \leq k \leq m$. Splňuje-li zobrazení f předpoklad F4, platí

$$\|C(x_i)\| = \|\sum_{k=1}^{m} f_k(x_i) G_k(x_i)\| \leq \overline{G} \sum_{k=1}^{m} |f_k(x_i)| \to 0$$

a tedy $\|G(x_i) - B_i\| = \|C(x_i)\| \to 0$, což je nutná podmínka pro Q-superlineární konvergenci.

(2) Linearizace. Platí

$$\begin{aligned} F(x_i + s) &= \frac{1}{2} f^T(x_i + s) f(x_i + s) \approx \frac{1}{2} (f(x_i) + J(x_i)s)^T (f(x_i) + J(x_i)s) = \\ &= \frac{1}{2} f^T(x_i) f(x_i) + f^T(x_i) J(x_i) s + \frac{1}{2} s^T J^T(x_i) J(x_i) s, \end{aligned}$$

takže

$$F(x_i + s) - F(x_i) \approx g^T(x_i) s + \frac{1}{2} s^T B_i s,$$

což je lokální kvadratická aproximace s maticí $B_i = J_i^T J_i$.

Věta 160. *Splňuje-li zobrazení f předpoklady J1, J4, F3, je Gaussova–Newtonova metoda, realizovaná jako metoda s lokálně omezeným krokem, globálně konvergentní. Splňuje-li zobrazení f předpoklady F4*, J4*, J5* a platí-li $x_i \to x^*$, $F(x^*) = 0$ a $\omega_i(s_i) \to 0$ (vzorec (517)), je rychlost konvergence Q-superlineární.*

Důkaz Splňuje-li zobrazení f předpoklady J1, J4, F3, splňuje funkce F předpoklady F1 a F3 (věta 157) a platí

$$\|B_i\| = \|J(x_i)J(x_i)^T\| \leq \overline{J}^2,$$

takže Gaussova–Newtonova metoda je podle věty 118 globálně konvergentní. Splňuje-li zobrazení f předpoklady F4*, J4*, J5* a platí-li $F(x^*) = 0$ (neboli $f(x^*) = 0$), splňuje funkce F podle věty 158 předpoklad F4* a podle věty 159 předpoklad F5*. Jak již bylo ukázáno (poznámka 260) z $F(x_i) \to F(x^*) = 0$ plyne $B_i \to G(x_i) \to G(x^*)$, neboli

$$\frac{\|(G^* - B_i)s_i\|}{\|s_i\|} \leq \|G^* - B_i\| \to 0,$$

což spolu s $\omega_i(s_i) \to 0$ a s F4*, F5* implikuje Q-superlineární konvergenci (věta 122, ve které jsme pro zjednodušení použili předpoklady F4, F5 místo F4*, F5*). □

Poznámka 261. Směrový vektor odpovídající Gaussově–Newtonově metodě můžeme určit několika různými způsoby:

(1) Řešením normální soustavy rovnic. Dosadíme-li $B_i = J_i^T J_i$ a $g_i = J_i^T f_i$ do vztahu $B_i s_i + g_i = 0$, dostaneme soustavu lineárních rovnic $J_i^T J_i s_i + J_i^T f_i = 0$, která se nazývá normální soustavou rovnic.

(2) Řešením linearizované úlohy nejmenších čtverců (přeurčené soustavy lineárních rovnic). Tato úloha má tvar $J_i s_i + f_i \approx 0$. Způsob jejího řešení je popsán v oddílu 8.4. Používá se stabilní QR-rozklad matice J_i. Při realizaci s lokálně omezeným krokem můžeme soustavu $(J_i^T J_i + \lambda I)s + J_i^T f_i = 0$ nahradit linearizovanou úlohou

$$\begin{bmatrix} J_i \\ \sqrt{\lambda} I \end{bmatrix} s + \begin{bmatrix} f_i \\ 0 \end{bmatrix} \approx 0.$$

(3) Řešením rozšířené soustavy rovnic. Označme $r_i = -(J_i s_i + f_i)$. Směrový vektor hledáme tak, aby platilo $J_i^T r_i = 0$. To dohromady dává

$$\begin{bmatrix} I, & J_i \\ J_i^T, & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0,$$

což je soustava $m + n$ rovnic se symetrickou indefinitní maticí. Tento způsob je vhodný pro řídké úlohy, neboť řídkost matice J_i implikuje řídkost matice rozšířené soustavy rovnic, zatímco matice normální soustavy rovnic může být hustá (například, má-li matice J_i hustý řádek). Použití rozšířené soustavy rovnic je vhodné i pro vážené úlohy. Jestliže

$$F(x) = \frac{1}{2} f^T(x) W f(x),$$

kde W je váhová matice, pak normální soustava má tvar

$$J_i^T W J_i s_i + J_i^T W f_i = 0,$$

a označíme-li $r_i = -W(J_i s_i + f_i)$, dostaneme

$$\begin{bmatrix} W^{-1}, & J_i \\ J_i^T, & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0.$$

8.2 Použití kvazinewtonovských aktualizací

Gaussova–Newtonova metoda je velmi efektivní pro úlohy s nulovými rezidui, může však selhávat v případě úloh s velkými rezidui. Proto se nabízí tato strategie:

(a) Jestliže $F_i \to F^* = 0$, použijeme Gaussovu–Newtonovu metodu.
(b) Jestliže $F_i \to F^* > 0$, použijeme nějakou superlineárně konvergentní metodu (buď Newtonovu metodu nebo metodu s proměnnou metrikou).

Následující věta udává způsob jak rozhodnout, která metoda bude použita.

Věta 161. *Nechť $F_i \to F^* = 0$ Q-superlineárně. Pak*

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 1.$$

Nechť $F_i \to F^* > 0$. Pak

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 0.$$

Důkaz Jestliže $F_i \to F^* = 0$ Q-superlineárně, pak platí

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = 1 - \lim_{i\to\infty} \frac{F_{i+1} - F^*}{F_i - F^*} = 1 - 0 = 1.$$

Jestliže $F_i \to F^* > 0$, pak

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = \frac{1}{F^*} \lim_{i\to\infty} (F_i - F_{i+1}) = 0.$$

□

Poznámka 262. Velmi efektivní hybridní metodu dostaneme, zkombinujeme-li Gaussovu–Newtonovu metodu s metodou BFGS: Nechť $B_1 = J_1^T J_1$. Jestliže $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme

$$B_{i+1} = J_{i+1}^T J_{i+1}.$$

Jestliže $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, položíme

$$B_{i+1} = B_i + \frac{y_i y_i^T}{y_i^T d_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i},$$

kde $d_i = x_{i+1} - x_i$ a $y_i = g_{i+1} - g_i = J_{i+1}^T f_{i+1} - J_i^T f_i$. Obvykle $\underline{\vartheta} = 0.01$ pro metody spádových směrů a $\underline{\vartheta} = 0.0001$ pro metody s lokálně omezeným krokem.

Nyní se budeme zabývat dalšími kombinacemi Gaussovy–Newtonovy metody s metodami s proměnnou metrikou, které se často nazývají strukturovanými metodami s proměnnou metrikou. Budeme předpokládat, že $B_i = J_i^T J_i + C_i$, kde C_i je nějaká aproximace matice $C(x_i)$ a budeme hledat matici C_{i+1} tak, aby matice $B_{i+1} = J_{i+1}^T J_{i+1} + C_{i+1}$ splňovala kvazinewtonovskou podmínku $B_{i+1} d_i = y_i$, kde opět $d_i = x_{i+1} - x_i$ a $y_i = g_{i+1} - g_i = J_{i+1}^T f_{i+1} - J_i^T f_i$. Existují dva způsoby, jak toho docílit. První způsob je založen na použití tranformované kvazinewtonovské podmínky

$$C_+ d = z \triangleq y - J_+^T J_+ d = J_+^T f_+ - J^T f - J_+^T J_+ d,$$

která bezprostředně plyne z podmínky $B_+ d = y$. Dostaneme tak aktualizaci

$$C_+ = C + \frac{z z^T}{d^T z} - \frac{Cd(Cd)^T}{d^T C d} + \frac{\beta}{d^T C d} \left(\frac{d^T C d}{d^T z} z - Cd \right) \left(\frac{d^T C d}{d^T z} z - Cd \right)^T. \tag{672}$$

Nevýhoda popsaného způsobu spočívá v tom, že číslo $d^T z$ nemusí být kladné, což komplikuje použití metody BFGS (s $\beta = 0$). V této souvislosti se nejvíce používá metoda hodnosti 1, kdy

$$C_+ = C + \frac{(z - Cd)(z - Cd)^T}{d^T(z - Cd)}. \tag{673}$$

(matice C_+ nemusí být positivně definitní, neboť aproximuje člen druhého řádu, který se přičítá k matici $J_+^T J_+$).

Druhý způsob je založen na aktualizaci matice $\bar{B} = J_+^T J_+ + C$ tak, aby matice $B_+ = J_+^T J_+ + C_+$ splňovala kvazinewtonovskou podmínku $B_+ d = y$. V tomto případě můžeme použít aktualizaci (306), kde matice B je nahražena maticí $\bar{B}$. Protože $y - \bar{B}d = z - Cd$, je výhodné použít vzorec (360). Označíme-li $v = Wd$, můžeme psát

$$\begin{aligned} C_+ &= C + \frac{(y - \bar{B}d)v^T + v(y - \bar{B}d)^T}{d^T v} - \frac{(y - \bar{B}d)^T d}{d^T v}\frac{vv^T}{d^T v} \\ &= C + \frac{(z - Cd)v^T + v(z - Cd)^T}{d^T v} - \frac{(z - Cd)^T d}{d^T v}\frac{vv^T}{d^T v}, \end{aligned} \tag{674}$$

kde $v = d$ pro aktualizaci PSB, $v = y$ pro aktualizaci DFP a $v = y + (y^T d/d^T \bar{B}d)^{1/2}\bar{B}d$ pro aktualizaci BFGS (metoda hodnosti 1 používá opět aktualizaci (673)).

Poznámka 263. Vektory y a z mohou být definovány různým způsobem, musí ale platit $z = y - J_+^T J_+ d$. Standardní volba

$$z = J_+^T f_+ - J^T f - J_+^T J_+ d \tag{675}$$

odpovídá kvazinewtonovské podmínce $(J_+^T J_+ + C_+)d = J_+^T f_+ - J^T f$. Velmi efektivní volba je založena na explicitním tvaru členu druhého řádu. Předpokládejme, že aproximace B_k^+ Hessových matic G_k splňují kvazinewtonovské podmínky $B_k^+ d = g_k^+ - g_k$, $1 \le k \le m$. Pak můžeme psát

$$z = C^+ d = \sum_{k=1}^{m} f_k^+ B_k^+ d = \sum_{k=1}^{m} f_k^+ (g_k^+ - g_k) = (J_+ - J)^T f_+. \tag{676}$$

Metody s proměnnou metrikou pro součet čtverců lze realizovat v součinovém tvaru (věta 86). Nyní se budeme zabývat strukturovanými metodami s proměnnou metrikou, které využívají znalost Jacobiovy matice. Abychom mohli tyto metody vyjádřit v součinovém tvaru, položíme $A = J + L$, $A_+ = J_+ + L_+$ a matici L budeme aktualizovat tak, aby platilo

$$B_+ d = A_+^T A_+ d = (J_+ + L_+)^T (J_+ + L_+)d = y.$$

Jelikož v případě součtu čtverců lze v součinovém tvaru efektivně realizovat pouze metodu BFGS (poznámka 136), omezíme se na podtřídu metod s proměnnou metrikou, která obsahuje metodu BFGS a pro níž je odvození součinového tvaru mnohem jednodušší než v obecném případě. K odvození součinového tvaru použijeme variační princip. Abychom ho mohli použít, zapíšeme kvazinewtonovskou podmínku ve tvaru

$$(J_+ + L_+)^T z = y, \qquad (J_+ + L_+)d = z, \qquad z^T z = d^T y, \tag{677}$$

kde z je volitelný vektor (parametr). Poznamenejme, že poslední rovnost, která je důsledkem prvních dvou rovností, je jediným omezením kladeným na volbu vektoru z.

Věta 162. *Nechť T je symetrická pozitivně definití matice. Pak Frobeniova norma $\|T^{-1/2}(L_+ - L)\|_F$ je minimální na množině všech matic vyhovujících rovnosti $(J_+ + L_+)^T z = y$ právě tehdy, platí-li*

$$L_+ = L - \frac{Tz(y - \bar{A}^T z)^T}{z^T T z},$$

kde $\bar{A} = J_+ + L$. Kvazinewtonovská podmínka (677) je v tomto případě splněna právě tehdy, jestliže $Tz = z - \bar{A}d$ a $z^T z = y^T d$.

Důkaz (a) Nutnost první části tvrzení dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L(L_+, u) &= \frac{1}{2}\left\|T^{-1/2}(L_+ - L)\right\|_F^2 + u^T\left((J_+ + L_+)^T z - y\right) \\ &= \sum_{i=1}^{m}\left[\frac{1}{2}\left(l_i^+ - l_i\right)^T T^{-1}\left(l_i^+ - l_i\right) + u_i z^T l_i^+\right] + u^T(J_+^T z - y), \end{aligned}$$

kde $L_+ = [l_1^+, \dots l_m^+]$ a $L = [l_1, \dots l_m]$. Postačitelnost je pak bezprostředním důsledkem konvexity Frobeniovy normy. Derivováním Langrangeovy funkce dostaneme

$$\frac{\partial L(L_+, u)}{\partial l_i^+} = T^{-1}\left(l_i^+ - l_i\right) + u_i z.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $T^{-1}(l_i^+ - l_i) + u_i z = 0$, $1 \le i \le m$, neboli

$$A_+ - \bar{A} = L_+ - L = -Tzu^T.$$

Z rovnosti $A_+^T z = y$ dostaneme $(A_+ - \bar{A})^T z = -z^T Tzu = y - \bar{A}^T z$, takže

$$u = -\frac{y - \bar{A}^T z}{z^T T z},$$

což po dosazení do předchozí rovnosti dává

$$A_+ - \bar{A} = L_+ - L = \frac{Tz(y - \bar{A}^T z)^T}{z^T T z}.$$

(b) Předpokládejme, že je splněna kvazinewtonovská podmínka (677), takže $(A_+ - \bar{A})d = z - \bar{A}d$. Pak platí

$$\frac{Tz(y - \bar{A}^T z)^T d}{z^T T z} = z - \bar{A}d.$$

Z tohoto vyjádření je zřejmé, že vektor Tz je rovnoběžný s vektorem $z - \bar{A}d$. Jelikož matici T můžeme vynásobit libovolným číslem aniž se změní zlomek na levé straně, můžeme položit $Tz = z - \bar{A}d$ (to lze provést pouze tehdy, když $T \neq I$, případ $T = I$ je vyšetřen v poznámce 264). Nechť naopak $Tz = z - \bar{A}d$ a $z^T z = d^T y$. Pak platí

$$A_+ - \bar{A} = L_+ - L = \frac{(z - \bar{A}d)(y - \bar{A}^T z)^T}{z^T(z - \bar{A}d)}.$$

a

$$A_+ d = \bar{A}d + (z - \bar{A}d)\frac{(y - \bar{A}^T z)^T d}{z^T(z - \bar{A}d)} = \bar{A}d + (z - \bar{A}d) = z,$$

takže je splněna i druhá podmínka z (677). □

Poznámka 264. Metodu BFGS dostaneme, zvolíme-li $T = I$. Pak vektor z je rovnoběžný s vektorem $\bar{A}d$, tedy $z = \lambda\bar{A}d$ a $Tz = (\lambda - 1)\bar{A}d$. Z poslední podmínky v (677) plyne, že $z^T z = \lambda^2 d^T \bar{A}^T \bar{A}d = d^T y$, což po dosazení do vztahu uvedeného ve větě 162 dává

$$L_+ = L + \frac{\bar{A}d}{d^T \bar{A}^T \bar{A}d}\left(\sqrt{\frac{d^T \bar{A}^T \bar{A}d}{d^T y}}\, y - \bar{A}^T \bar{A}d\right)^T. \tag{678}$$

Pokud $J_+ = 0$, takže $\bar{A} = A$, přejde tento výraz v (350).

Jistá nevýhoda aktualizace (678) spočívá v tom, že řešení lineárního problému nejmenších čtverců $(J+L)d+f \approx 0$ není řešením normální soustavy rovnic $(J+L)^T(J+L)d = -g = -J^T f$, která se používá pro výpočet směrového vektoru. Nelze tedy použít efektivní metody založené na QR rozkladu ani metodu LSQR (definice 77). Tuto nevýhodu lze odstranit, volíme-li matici L tak, aby platilo $(J+L)^T f = J^T f$, neboli $L^T f = 0$. Je tedy výhodné přidat omezení $L_+^T f_+ = 0$ k variační úloze definující metodu BFGS. Pokud $L_+^T f_+ = 0$, je minimalizace Frobeniovy normy $\|L_+ - L\|_F$ ekvivalentní minimalizaci Frobeniovy normy $\|P(L_+ - L)\|_F$, kde $P = I - f_+ f_+^T / f_+^T f_+$ je matice ortogonální projekce (připomeňme si, že $P^2 = P$). Plyne to z toho, že $PL_+ = L_+$, takže

$$\begin{aligned}(L_+ - L)^T P(L_+ - L) &= L_+^T PL_+ - L^T PL_+ - L_+^T PL + L^T PL \\ &= L_+^T L_+ - L^T L_+ - L_+^T L + L^T PL \\ &= (L_+ - L)^T (L_+ - L) + L^T (P - I)L,\end{aligned}$$

kde poslední člen je konstantní.

Věta 163. *Frobeniova norma $\|P(L_+ - L)\|_F$ je minimální na množině všech matic vyhovujících kvazi-newtonovské podmínce (677) a omezení $L_+^T f_+ = 0$ právě tehdy, platí-li*

$$L_+ = PL + \frac{\tilde{A}d}{d^T \tilde{A}^T \tilde{A}d} \left(\sqrt{\frac{d^T \tilde{A}^T \tilde{A}d}{d^T y}} \tilde{y} - \tilde{A}^T \tilde{A}d \right)^T . \tag{679}$$

kde

$$\tilde{A} = P(J_+ + L), \qquad \tilde{y} = y - \frac{J_+ f_+ (J_+ f_+)^T d}{f_+^T f_+}.$$

Důkaz (a) Nejprve ukážeme, že pokud $(J_+ + L_+)d = z$ a $L_+^T f_+ = 0$, je podmínka $(J_+ + L_+)^T z = y$ ekvivalentní podmínce $(J_+ + L_+)^T Pz = \tilde{y}$. Z $(J_+ + L_+)d = z$ a $L_+^T f_+ = 0$ totiž plyne $f_+^T J_+ d = f_+^T z$, takže

$$\begin{aligned}(J_+ + L_+)^T Pz - \tilde{y} &= J_+^T z - \frac{J_+^T f_+ f_+^T J_+ d}{f_+^T f_+} + L_+^T Pz - y + \frac{J_+^T f_+ f_+^T J_+ d}{f_+^T f_+} \\ &= J_+^T z + L_+^T Pz - y = (J_+ + L_+)^T z - y.\end{aligned}$$

Poznamenejme, že z rovností $(J_+ + L_+)d = z$ a $(J_+ + L_+)^T Pz = \tilde{y}$ plyne vztah $z^T Pz = d^T \tilde{y}$.
(b) Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned}L(L_+, u) &= \frac{1}{2} \|P(L_+ - L)\|_F^2 + u^T \left((J_+ + L_+)^T Pz - \tilde{y} \right) \\ &= \sum_{i=1}^{m} \left[\frac{1}{2} \left(l_i^+ - l_i \right)^T P \left(l_i^+ - l_i \right) + u_i z^T P l_i^+ \right] + u^T (J_+^T Pz - \tilde{y}),\end{aligned}$$

kde $L_+ = [l_1^+, \dots l_m^+]$ a $L = [l_1, \dots l_m]$. Postačitelnost je pak bezprostředním důsledkem konvexity Frobeniovy normy. Derivováním Langrangeovy funkce dostaneme

$$\frac{\partial L(L_+, u)}{\partial l_i^+} = P \left(l_i^+ - l_i \right) + u_i Pz.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $P(l_i^+ - l_i) + u_i Pz = 0$, $1 \le i \le m$, neboli

$$P(L_+ - L) = -Pzu^T.$$

Z rovnosti $(J_+ + L_+)^T Pz = \tilde{y}$ dostaneme $(L_+ - L)^T Pz = -z^T Pzu = \tilde{y} - \tilde{A}^T z$, takže

$$u = -\frac{\tilde{y} - \tilde{A}^T z}{z^T Pz},$$

což po dosazení do předchozí rovnosti dává

$$P(L_+ - L) = \frac{Pz(\tilde{y} - \tilde{A}^T Pz)^T}{z^T Pz} \tag{680}$$

(neboť $P^2 = P$ implikuje $P\tilde{A} = \tilde{A}$). Použijeme-li druhou podmínku z (677), dostaneme $P(L_+ - L)d = Pz - \tilde{A}d$, takže lze psát

$$\frac{Pz(\tilde{y} - \tilde{A}^T Pz)^T d}{z^T Pz} = Pz - \bar{A}d.$$

Z posledního vyjádření je zřejmé, že vektor Pz je rovnoběžný s vektorem $\tilde{A}d$, neboli $Pz = \lambda\tilde{A}d$. Použijeme-li vztah $z^T Pz = d^T\tilde{y}$ dokázaný v (a), můžeme psát

$$\lambda^2 d^T\tilde{A}^T\tilde{A}d = z^T Pz = d^T\tilde{y} \quad \Rightarrow \quad \lambda = \pm\sqrt{\frac{d^T\tilde{y}}{d^T\tilde{A}^T\tilde{A}d}},$$

což po dosazení do $Pz = \lambda\tilde{A}d$ a potom do (680) dokazuje tvrzení věty. □

Poznámka 265. Strukturované metody s proměnou metrikou pro minimalizaci součtu čtverců byly původně navrženy tak, že se matice $B_i = J_i^T J_i + C_i$ používaly a matice C_i aktualizovaly v každém iteračním kroku. To je však nevýhodné, neboť v úlohách s nulovým reziduem, potřebujeme, aby $C_i \to 0$ dostatečné rychle, zatímco při použití aktualizací (672) nebo (674) je tato konvergence obvykle příliš pomalá. Proto byly vyvíjeny různé škálovací strategie. Ukázalo se však že je výhodnější používat hybridní strategie tak jako v poznámce 262: Nechť $C_1 = 0$. Jestliže $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme $C_{i+1} = 0$. Jestliže $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, aktualizujeme matici C_i pomocí (672) nebo (674). V obou případech pokládáme

$$B_{i+1} = J_{i+1}^T J_{i+1} + C_{i+1}$$

(stejné úvahy se týkají strukturovaných metod s proměnnou metrikou používající matice $A_i = J_i + L_i$ a aktualizace (678) nebo (679).

Poznámka 266. Velmi zajímavou možnost automatického škálování matice C nabízejí totálně strukturované metody s proměnnou metrikou uvedené v práci [86]. V tomto případě se používá a aktualizuje matice aproximující výraz

$$T(x) = \sum_{k=1}^{m} \frac{f_k(x)}{\|f(x)\|} G_k(x).$$

Používáme tedy model $B = J^T J + \|f\|T$ (takže $C = \|f\|T$) a matici T_+ aktualizujeme tak aby matice $\tilde{B}_+ = J_+^T J_+ + \|f\|T_+$ splňovala kvazinewtonovskou podmínku $\tilde{B}_+ s = y$. Toho lze docílit tak, že aplikujeme aktualizaci (360) na matici $\bar{B} = J_+^T J_+ + \|f\|T$. Nakonec položíme $B_+ = J_+^T J_+ + \|f_+\|T_+$. Užitím vztahu (360) dostaneme

$$\begin{aligned} T_+ &= T + \frac{1}{\|f\|}\left(\frac{(y - \bar{B}d)v^T + v(y - \bar{B}d)^T}{d^T v} - \frac{(y - \bar{B}d)^T d}{d^T v}\frac{vv^T}{d^T v}\right) \\ &= T + \frac{(\tilde{z} - Td)v^T + v(\tilde{z} - Td)^T}{d^T v} - \frac{(\tilde{z} - Td)^T d}{d^T v}\frac{vv^T}{d^T v}, \end{aligned} \tag{681}$$

kde $\tilde{z} = z/\|f\| = (y - J_+^T J_+ d)/\|f\|$ a kde $v = d$ pro aktualizaci PSB, $v = y$ pro aktualizaci DFP a $v = y + (y^T d/d^T\bar{B}d)^{1/2}\bar{B}d$ pro aktualizaci BFGS. Metoda hodnosti 1 používá aktualizaci

$$T_+ = T + \frac{(\tilde{z} - Ts)(\tilde{z} - Td)^T}{d^T(\tilde{z} - Td)}.$$

8.3 Numerické porovnání metod pro minimalizaci součtu čtverců

Metody studované v této kapitole byly testovány pomocí 102 testovacích úloh s 200 proměnnými (TEST24 z oddílu 1.5). V následující tabulce jsou uvedeny výsledky testů odpovídající těmto metodám:

`VM1` - metoda BFGS (288) s aktualizací symetrické matice,
`VM2` - metoda BFGS (309) s aktualizací Choleského rozkladu,
`GN` - Gaussova-Newtonova metoda,
`GNVM1` - hybridní metoda uvedená v poznámce 262,
`GNVM2` - strukturovaná metoda BFGS podle (674),
`GNVM3` - strukturovaná metoda BFGS podle (681).

Tyto metody jsou realizovány buď jako metody spádových směrů LS (první část tabulky) nebo jako metody s lokálně omezeným krokem TR (druhá část tabulky). Jsou uvedeny celkové počty iterací `NIT`, funkčních hodnot `NFV` a gradientů `NFG`, jakož i počet selhání a celkový čas výpočtu.

metoda `LS`	`NIT`	`NFV`	`NFG`	selhání	čas
`VM1`	21140	23442	23442	1	6.64
`GN`	10658	19266	29922	6	98.20
`GNVM1`	5103	7321	12423	-	46.78
`GNVM2`	4109	7205	11314	1	32.44
`GNVM3`	4174	7245	11419	1	32.15
metoda `TR`	`NIT`	`NFV`	`NFG`	selhání	čas
`VM2`	23627	25910	23726	2	12.34
`GN`	5555	6185	5650	2	40.44
`GNVM1`	4382	4970	4483	-	44.28
`GNVM2`	3818	4346	3919	-	39.49
`GNVM3`	3513	3970	3615	-	36.53

Tabulka 7: TEST24 – 192 úloh

Poznámka 267. Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Gaussovu–Newtonovu metodu není vhodné realizovat jako metodu spádových směrů, neboť se často řeší soustavy rovnic se špatně podmíněnými maticemi.
- Gaussovu–Newtonovu metodu je možné značně vylepšit kombinováním s metodami s proměnnou metrikou a to buď pomocí jednoduché hybridní strategie (poznámka 262) nebo pomocí strukturovaných aktualizací (674) a (681). Tyto kombinované metody jsou velmi robustní (ve spojení s metodami s lokálně omezeným krokem nikdy neselhaly). Potřebují také nejméně iterací a vyčíslení hodnot minimalizované funkce a nejsou tak citlivé na způsob realizace (fungují dobře i jako metody spádových směrů).
- Metody s proměnou metrikou mají menší režii, neboť kvazinewtonovské aktualizace vyžadují $O(n^2)$ aritmetických operací, zatímco řešení soustav lineárních rovnic, použitá u modifikací Gaussovy-Newtonovy metody, vyžadují $O(n^3)$ aritmetických opreací. Výpočetní čas metod s proměnou metrikou je tedy obvykle nižší, jak lze vyčíst z uvedené tabulky. Rozdíl časů je zde poměrně velký, neboť pro $n = 200$ je $O(n^3)$ mnohem větší než $O(n^2)$. Metody s proměnnou metrikou jsou poměrně robustní, neboť dokážou vyřešit téměř všechny specializované úlohy. Také je patrné, že metody s proměnnou metrikou je výhodnější realizovat jako metody spádových směrů, neboť aktualizace symetrické matice vyžadují méně operací než aktualizace jejího Choleského rozkladu popsané v oddílu 4.7.

8.4 Řešení lineární úlohy nejmenších čtverců

Nechť J je matice která má m řádků a n sloupců, kde $m \geq n$. Lineární úlohou nejmenších čtverců rozumíme nalezení vektoru $s \in R^n$, který minimalizuje normu $\|Js + f\|$. Vyhovuje-li této úloze více vektorů, volíme ten, který má nejmenší normu. Lineární úlohu nejmenších čtverců budeme zapisovat ve tvaru $Js + f \approx 0$.

Věta 164. *Vektor s je řešením úlohy nejmenších čtverců $Js + f \approx 0$ právě tehdy, když $s = -J^\dagger f$ kde $J^\dagger$ je pseudoinverze matice J.*

Důkaz (a) Má-li matice J plnou hodnost, je vektor s řešením lineární úlohy nejmenších čtverců $Js+f \approx 0$ právě tehdy, když $s = -(J^TJ)^{-1}J^Tf$. Nutnost plyne z věty 4 a z vyjádření (650). Postačitelnost plyne z toho, že matice J^TJ je pozitivně definitní, takže kvadratická funkce $\|Js + f\|^2$ je konvexní a má tedy jediný stacionární bod, který je jejím globálním minimem. Podle poznámky 127 platí $(J^TJ)^{-1}J^T = J^\dagger$, takže $s = -J^\dagger f$.
(b) Má-li matice J hodnost $l < n$, můžeme psát $J = UV^T$, kde matice $U \in R^{m\times l}$ a $V \in R^{n\times l}$ mají plnou hodnost. Jelikož matice U má plnou hodnost, je podle (a) vektor s řešením úlohy $Js = UV^Ts + f \approx 0$ právě tehdy, když $V^Ts = -(U^TU)^{-1}U^Tf$. Vektor s můžeme jednoznačně vyjádřit ve tvaru $s = Vy+z$, kde z leží v ortogonálním doplňku podprostoru $\mathcal{L}(V)$ (takže $V^Tz = 0$). Pak platí $V^TVy = -(U^TU)^{-1}U^Tf$, neboli

$$y = -(V^TV)^{-1}(U^TU)^{-1}U^Tf \quad \Leftrightarrow \quad s = -V(V^TV)^{-1}(U^TU)^{-1}U^Tf + z.$$

Podle definice 37 se snadno přesvědčíme, že $V(V^TV)^{-1}(U^TU)^{-1}U^T = (UV^T)^\dagger = J^\dagger$. Vektor s je tedy řešením úlohy $Js + f \approx 0$ právě tehdy, když $s = -J^\dagger f + z$. Ale

$$\|s\| = (J^\dagger f)^TJ^\dagger f + 2z^TJ^\dagger f + z^Tz = \|J^\dagger f\|^2 + \|z\|^2$$

(neboť $z^TJ^\dagger = z^TV(V^TV)^{-1}(U^TU)^{-1}U^T = 0$), takže vektor s má minimální normu právě tehdy, když $z = 0$, což dává $s = -J^\dagger f$. □

Poznámka 268. V případě, že matice J má plnou hodnost, lze řešení lineární úlohy nejmenších čtverců získat řešením normální soustavy rovnic $J^TJs = -J^Tf$ pomocí Choleského rozkladu matice J^TJ. Tento přístup není příliš vhodný, neboť platí $\kappa(J^TJ) = \kappa^2(J)$, kde $\kappa(J)$ je spektrální číslo podmíněnosti matice J. Proto je výhodnější použít ortogonální rozklad matice J. Je-li matice Q ortogonální, platí $\|Q\| = 1$ a $\kappa(Q) = 1$ (připomeňme, že čtvercová matice je ortogonální, platí-li $Q^TQ = QQ^T = I$).

Abychom mohli popsat ortogonální rozklad matice J, budeme nejprve studovat ortogonální opreace realizované elementárními ortogonálními maticemi.

Definice 54. *Řekneme že matice $Q \in R^{2\times 2}$ je Givensovou maticí elementární rotace, jestliže platí*

$$Q = \begin{bmatrix} c, & -s \\ s, & c \end{bmatrix},$$

kde $c^2 + s^2 = 1$. Tato matice je ortogonální, neboť

$$Q^TQ = \begin{bmatrix} c, & s \\ -s, & c \end{bmatrix}\begin{bmatrix} c, & -s \\ s, & c \end{bmatrix} = \begin{bmatrix} c^2+s^2, & 0 \\ 0, & s^2+c^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Poznámka 269. Nechť $x \in R^2$ a $\|x\| = 1$, takže lze psát $x_1 = \cos\alpha$ a $x_2 = \sin\alpha$. Jelikož $c^2 + s^2 = 1$, můžeme položit $c = \cos\varphi$ a $s = \sin\varphi$. Pak platí

$$Qx = \begin{bmatrix} \cos\varphi, & -\sin\varphi \\ \sin\varphi, & \cos\varphi \end{bmatrix}\begin{bmatrix} \cos\alpha \\ \sin\alpha \end{bmatrix} = \begin{bmatrix} \cos\varphi\cos\alpha - \sin\varphi\sin\alpha \\ \sin\varphi\cos\alpha + \cos\varphi\sin\alpha \end{bmatrix} = \begin{bmatrix} \cos(\varphi+\alpha) \\ \sin(\varphi+\alpha) \end{bmatrix},$$

takže vektor $y = Qx$ má také jednotkovou normu a vznikne rotací vektoru x o úhel φ.

Givensovu matici můžeme použít k vynulování prvku vektoru.

Věta 165. *Nechť $x \in R^2$ a nechť $Q \in R^{2\times 2}$ je Givensova matice taková, že $c = x_1/\|x\|$ a $s = x_2/\|x\|$. Pak platí*

$$Q^T x = \begin{bmatrix} c, & s \\ -s, & c \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \|x\| \\ 0 \end{bmatrix}$$

Důkaz Platí

$$c^2 + s^2 = \frac{1}{\|x\|^2}(x_1^2 + x_2^2) = 1,$$

takže po dosazení dostaneme

$$Q^T x = \begin{bmatrix} c, & s \\ -s, & c \end{bmatrix} \begin{bmatrix} c\|x\| \\ s\|x\| \end{bmatrix} = \begin{bmatrix} \|x\| \\ 0 \end{bmatrix}.$$

□

Poznámka 270. Givensovy matice elementárních rotací můžeme definovat i v $R^{n\times n}$. V tomto případě se Givensova matice Q_{ij} liší od jednotkové matice řádu n pouze tím, že jednotková submatice řádu 2, obsahující elementy i-tého a j-tého řádku a sloupce, je nahrazena submaticí

$$\begin{bmatrix} c_{ij}, & -s_{ij} \\ s_{ij}, & c_{ij} \end{bmatrix},$$

kde $c_{ij}^2 + s_{ij}^2 = 1$.

Poznámka 271. Givensovy matice elementárních rotací z $R^{n\times n}$ můžeme použít k vynulování posledních $n-1$ prvků vektoru $x \in R^n$. Zvolíme-li vhodně prvky $c_{i,i+1}$, $s_{i,i+1}$, $1 \le i \le n-1$, platí

$$Q_{12}^T Q_{23}^T \dots Q_{n-1,n}^T x = \begin{bmatrix} \|x\| \\ 0 \\ \dots \\ 0 \end{bmatrix}.$$

Nulování prvků se provádí podle schematu

$$\begin{bmatrix} * \\ * \\ \boxed{*} \\ \boxed{*} \end{bmatrix} \to \begin{bmatrix} * \\ \boxed{*} \\ \boxed{*} \\ 0 \end{bmatrix} \to \begin{bmatrix} \boxed{*} \\ \boxed{*} \\ 0 \\ 0 \end{bmatrix} \to \begin{bmatrix} \|x\| \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Dostáváme postupně vektory, kterým ubývají nenulové prvky od n-tého po druhý řádek.

Poznámka 272. Další užitečná aplikace Givensových matic elementárních rotací z $R^{n\times n}$ je nulování poddiagonálních prvků Hessenbergovy matice řádu n. Zvolíme-li vhodně prvky $c_{i,i+1}$, $s_{i,i+1}$, $1 \le i \le n-1$, platí

$$Q_{n-1,n}^T \dots Q_{23}^T Q_{12}^T H = R$$

Nulování prvků se provádí podle schematu

$$\begin{bmatrix} \boxed{*} & \boxed{*} & \boxed{*} & \boxed{*} \\ \boxed{*} & \boxed{*} & \boxed{*} & \boxed{*} \\ 0 & * & * & * \\ 0 & 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & \boxed{*} & \boxed{*} & \boxed{*} \\ 0 & \boxed{*} & \boxed{*} & \boxed{*} \\ 0 & 0 & * & * \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & \boxed{*} & \boxed{*} \\ 0 & 0 & \boxed{*} & \boxed{*} \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix}$$

Dostáváme postupně matice, kterým ubývají nenulové poddiagonální prvky od druhého po n-tý řádek.

Kromě Givensových matic elementárních rotací se též používají Householderovy matice elementárních reflexí. Householderova matice se liší od jednotkové matice členem, který má hodnost 1.

Definice 55. *Řekneme, že matice $Q \in R^{n\times n}$ je Householderovou maticí elementární reflexe, jestliže platí*

$$Q = I - 2\frac{vv^T}{v^Tv},$$

kde $v \in R^n$. Tato matice je ortogonální, neboť

$$Q^TQ = \left(I - 2\frac{vv^T}{v^Tv}\right)\left(I - 2\frac{vv^T}{v^Tv}\right) = I - 4\frac{vv^T}{v^Tv} + 4\frac{vv^T}{v^Tv} = I.$$

Poznámka 273. Nechť $x \in R^n$ a $v \in R^n$. Nechť vektor y je ortogonální projekcí vektoru x do směru vektoru v, takže

$$y = \frac{vv^T}{v^Tv}x = \frac{v^Tx}{v^Tv}v.$$

Vynásobíme-li vektor x maticí Q, dostaneme

$$Qx = \left(I - 2\frac{vv^T}{v^Tv}\right)x = x - 2\frac{v^Tx}{v^Tv}v = x - 2y,$$

takže vektor Qx vznikne zrcadlením vektoru x podle nadroviny kolmé k vektoru v (procházející počátkem souřadnic).

Věta 166. *Nechť $x \in R^n$, $y \in R^n$ a $\|y\| = \|x\|$. Položme*

$$Q = I - 2\frac{(x-y)(x-y)^T}{(x-y)^T(x-y)}.$$

Pak matice Q je Householderovou maticí elementární reflexe a platí $y = Qx$ a $x = Qy$.

Důkaz Matice Q je Householderovou maticí elementární reflexe podle definice 55. Platí

$$\begin{aligned} Qx &= x - 2(x-y)\frac{(x-y)^Tx}{(x-y)^T(x-y)} = x - 2(x-y)\frac{\|x\|^2 - y^Tx}{\|x\|^2 + \|y\|^2 - 2y^Tx} \\ &= x - 2(x-y)\frac{\|x\|^2 - y^Tx}{2(\|x\|^2 - y^Tx)} = x - (x-y) = y. \end{aligned}$$

Vztah $x = Qy$ plyne z toho, že matice Q je symetrická a ortogonální. □

Poznámka 274. Householderovu matici můžeme použít k vynulování posledních $n-1$ prvků vektoru $x \in R^n$. Položíme-li ve větě 166 $y = -\sigma\|x\|e_1$, kde $\sigma = \pm 1$ a e_1 je první sloupec jednotkové matice, můžeme psát $Q = I - 2vv^T/(v^Tv)$, kde

$$\begin{aligned} v &= x - y = x + \sigma\|x\|e_1, \\ v^Tv &= \|x\|^2 + 2\sigma\|x\|x_1 + \|x\|^2 = 2\|x\|(\|x\| + \sigma x_1). \end{aligned}$$

Znaménko volíme tak, aby jmenovatel v^Tv byl co největší, tedy $\sigma = \mathrm{sgn}(x_1)$. Pak $v = x + \sigma\|x\|e_1$, $v^Tv = 2\|x\|(\|x\| + |x_1|)$ a pro libovolný vektor $z \in R^n$ platí

$$Qz = z - 2\frac{v^Tz}{v^Tv}v = z - \frac{z^Tx + \sigma\|x\|z_1}{\|x\|(\|x\| + |x_1|)}(x + \sigma\|x\|e_1).$$

Vykrátíme-li výraz na pravé straně této rovnosti číslem $\|x\|^2$, můžeme psát

$$Qz = z - \frac{z^T\tilde{x} + \sigma z_1}{1 + |\tilde{x}_1|}(\tilde{x} + \sigma e_1),$$

kde $\tilde{x} = x/\|x\|$.

Definice 56. *Ortogonálním rozkladem matice $J \in R^{m\times n}$, $m \geq n$, nazveme vyjádření*

$$J = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{682}$$

kde Q je ortogonální matice řádu m a R je horní trojúhelníková matice řádu n.

Poznámka 275. K nalezení ortogonálního rozkladu (682) lze s výhodou použít Householderovy matice elementárních reflexí. Zvolíme-li vhodně vektory $v_i \in R^m$ (mající prvních $i-1$ prvků nulových) a položíme-li $Q_i = I - 2v_iv_i^T/v_i^Tv_i$, $1 \leq i \leq n$, můžeme psát

$$Q_n \dots Q_2Q_1J = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

(matice Q_i vynuluje $m-i$ prvků i-tého sloupce průběžně upravované matice). Ortogonální rozklad se provádí podle schematu

$$\begin{bmatrix} \boxed{\begin{matrix} * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \end{matrix}} \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & \boxed{\begin{matrix} * & * \\ * & * \\ * & * \\ * & * \end{matrix}} \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & \boxed{\begin{matrix} * \\ * \\ * \end{matrix}} \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Pak platí (682), kde $Q = Q_1Q_2 \dots Q_n$. Jelikož Householderovy matice Q_i, $1 \leq i \leq n$, jsou symetrické a ortogonální, platí $Q^TQ = Q_n \dots Q_2Q_1Q_1Q_2 \dots Q_n = I$, takže matice Q je ortogonální.

Poznámka 276. Má-li matice J hodnost $l < n$, není trojúhelníková matice R regulární a rozklad (682) má tvar

$$J = Q \begin{bmatrix} R, & S \\ 0, & 0 \end{bmatrix},$$

kde R je horní trojúhelníková matice řádu l a $S \in R^{l\times(n-l)}$. V tomto případě je vhodné prvky matice S vynulovat, což lze opět provést pomocí Householderových matic. Dostaneme tak rozklad

$$J = Q \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \tilde{Q}^T,$$

kde $\tilde{Q} = \tilde{Q}_1\tilde{Q}_2 \dots \tilde{Q}_l$. Dodatečné ortogonální úpravy se provádí podle schematu

$$\begin{bmatrix} * & \boxed{\begin{matrix} * & * & * \\ * & * & * \end{matrix}} \\ 0 & \begin{matrix} 0 & 0 & 0 \\ 0 & 0 & 0 \end{matrix} \end{bmatrix} \to \begin{bmatrix} \boxed{*} & * & \boxed{\begin{matrix} * & * \\ 0 & 0 \end{matrix}} \\ 0 & * & \begin{matrix} 0 & 0 \\ 0 & 0 \end{matrix} \end{bmatrix} \to \left[\begin{array}{cc|cc} * & * & 0 & 0 \\ 0 & * & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Věta 167. *Nechť*

$$J = Q \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \tilde{Q}^T, \tag{683}$$

kde $Q \in R^{m\times m}$, $\tilde{Q} \in R^{n\times n}$ jsou ortogonální matice a R je regulární horní trojúhelníková matice řádu l, $l \leq n$. Pak vektor s je řešením úlohy nejmenších čtverců $Js + f \approx 0$ právě tehdy, když

$$s = \tilde{Q} \begin{bmatrix} -R^{-1}u \\ 0 \end{bmatrix},$$

kde u je vektor obsahující prvních l prvků matice Q^Tf.

Důkaz Položme

$$\tilde{Q}^T s = \begin{bmatrix} y \\ z \end{bmatrix}, \qquad Q^T f = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Pak podle (683) platí

$$Q^T J s = Q^T J \tilde{Q} \tilde{Q}^T s = Q^T J \tilde{Q} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix}$$

(využíváme toho, že matice Q a $\tilde{Q}$ jsou ortogonální). Norma $\|Js+f\|$ je minimální, pokud je $\|Q^T(Js+f)\|$ minimální (násobení ortogonální maticí zachovává normu), čili pokud je $\|Ry+u\|$ minimální. Matice R je regulární, takže $y = -R^{-1}u$. Jelikož

$$s = \tilde{Q}\tilde{Q}^T s = \tilde{Q} \begin{bmatrix} y \\ z \end{bmatrix},$$

platí $\|s\|^2 = \|y\|^2 + \|z\|^2$, takže $\|s\|$ je minimální, pokud $z = 0$. □

Poznámka 277. Popsali jsme základní myšlenky řešení lineárních úloh nejmenších čtverců pomocí ortogonálních rozkladů. Abychom získali efektivní algoritmy, je třeba vyřešit řadu praktických problémů. Předně je třeba ukládat informace o Householderových maticích Q_i, $1 \leq i \leq n$ a $\tilde{Q}_i$, $1 \leq i \leq l$, tedy vektory v_i, $1 \leq i \leq n$ a $\tilde{v}_i$, $1 \leq i \leq l$. Prvky těchto vektorů se obvykle ukládají na místa nově vznikajících nulových prvků v upravované matici. Také je třeba použít permutace sloupců. Těmto problémům se zde věnovat nebudeme. Velmi kvalitní algoritmy pro ortogonální rozklady matic jsou obsaženy s knihovně LAPACK [2].

Poznámka 278. Metody pro řešení lineárních úloh nejmenších čtverců používající ortogonální rozklady jsou přímými metodami, takže majdou požadované řešení po konečném počtu kroků (počítáme-li přesně). K řešení lineárních úloh nejmenších čtverců lze také použít iterační metody, které jsou podrobně popsány v oddílu 10.7.

9 Metody pro rozsáhlé husté úlohy

Rozsáhlé úlohy nemůžeme řešit metodami, které vyžadují uchovávání velkých hustých matic. V tomto oddílu se budeme zabývat metodami, které nepracují s aproximací Hessovy matice ani její inverze. Jsou to vektorové metody podobné metodám sdružených gradientů popsaným v oddílu 3, i když jsou poněkud složitější, a většinou metody sdružených gradientů předčí. Budeme se zabývat metodami s proměnnou metrikou založenými na omezeném počtu aktualizací nebo na aktualizaci obdélníkových matic s omezeným počtem sloupců. Tyto metody nevyužívají řídkost struktury optimalizační úlohy, takže je lze použít k minimalizaci funkcí s hustými Hessovými maticemi. Do tohoto oddílu zařadíme i jednu diferenční verzi Newtonovy metody.

9.1 Vektorové metody s proměnnou metrikou s omezenou pamětí

Definice 57. *Nechť $0 < \bar{m} < n$, $i \in N$ a $m = \min(\bar{m}, i-1)$. Řekneme, že základní optimalizační metoda je $\bar{m}$-krokovou metodou s proměnnou metrikou s omezenou pamětí, jestliže*

$$s_i = -H_i^i g_i,$$

kde matice H_i^i se získává z řídké pozitivně definitní (obvykle jednotkové) matice H_{i-m}^i pomocí m aktualizací

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + U_j^i M_j^i (U_j^i)^T \right),$$

$i-m \le j \le i-1$, kde matice $U_j^i = [d_j, H_j^i y_j]$ a M_j^i jsou voleny tak, aby byly splněny kvazinewtonovské podmínky $H_{j+1}^i y_j = \rho_j^i d_j$, $i-m \le j \le i-1$.

Poznámka 279. Aktualizace uvedené v definici 57 jsou stejné jako aktualizace použité v definici 36. Lze je tedy vyjádřit ve tvaru

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + \frac{\rho_j^i}{\gamma_j^i} \frac{1}{b_j} d_j d_j^T - \frac{1}{a_j^i} H_j^i y_j (H_j^i y_j)^T + \frac{\eta_j^i}{a_j^i} \left(\frac{a_j^i}{b_j} d_j - H_j^i y_j \right) \left(\frac{a_j^i}{b_j} d_j - H_j^i y_j \right)^T \right) \tag{684}$$

a invertovat tak, že platí

$$B_{j+1}^i = \frac{1}{\gamma_j^i} \left(B_j^i + \frac{\gamma_j^i}{\rho_j^i} \frac{1}{b_j} y_j y_j^T - \frac{1}{c_j^i} B_j^i d_j (B_j^i d_j)^T + \frac{\beta_j^i}{c_j^i} \left(\frac{c_j^i}{b_j} y_j - B_j^i d_j \right) \left(\frac{c_j^i}{b_j} y_j - B_j^i d_j \right)^T \right), \tag{685}$$

kde $B_j^i = (H_j^i)^{-1}$ a $a_j^i = y_j^T H_j^i y_j$, $b_j = y_j^T d_j$, $c_j^i = d_j^T B_j^i d_j$. Podstatné je to, že matice H_i^i vznikne z počáteční řídké matice H_{i-m}^i pomocí nejvýše $\bar{m}$ aktualizací, takže stačí ukládat omezený počet vektorů a provádět omezený počet operací. V dalším výkladu budeme předpokládat, že buď $H_{i-m}^i = I$ a $\gamma_{i-m} = y_{i-1}^T d_{i-1} / y_{i-1}^T y_{i-1}$ nebo $H_{i-m}^i = (y_{i-1}^T d_{i-1} / y_{i-1}^T y_{i-1})\, I$ a $\gamma_{i-m} = 1$ (vhodnost této volby potvrzují numerické experimenty).

Zvláště výhodná je metoda BFGS s omezenou pamětí s aktualizací

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + \left(\frac{\rho_j^i}{\gamma_j^i} + \frac{a_j^i}{b_j} \right) \frac{1}{b_j} d_j d_j^T - \frac{1}{b_j} \left(H_j^i y_j d_j^T + d_j (H_j^i y_j)^T \right) \right), \tag{686}$$

$i-m \le j \le i-1$, pro kterou platí tato věta [94].

Věta 168. *Nechť x_i, $i \in N$, je posloupnost generovaná $\bar{m}$-krokovou metodou BFGS s omezenou pamětí s $H_{i-m}^i = I$, $i \in N$, a s přesným výběrem délky kroku (takže $s_i^T g_{i+1} = 0$, $i \in N$) aplikovaná na ryze konvexní kvadratickou funkci*

$$Q(x) = \frac{1}{2}(x - x^*)^T G (x - x^*). \tag{687}$$

Pak platí

$$s_i = -\left(\prod_{j=i-m}^{i-1} \gamma_j^i\right)\left(g_i - \frac{y_{i-1}^T g_i}{y_{i-1}^T d_{i-1}} d_{i-1}\right) \tag{688}$$

pro $1 \leq i \leq n$ *(směrové vektory* s_i, $1 \leq i \leq n$, *jsou rovnoběžné se směrovými vektory generovanými metodou sdružených gradientů).*

Důkaz Pro $1 \leq i \leq \bar{m}$ jsou směrové vektory získané $\bar{m}$-krokovou metodou BFGS shodné se směrovými vektory získanými standardní metodou BFGS, takže pro tyto indexy platí (688) (důsledek 7 a jeho důkaz). Důkaz pro $i > \bar{m}$ provedeme indukcí. Předpokládejme, že (688) platí pro všechny indexy $1 \leq i < l$, kde $\bar{m} < l \leq n$ a položme $i = l$. Pak podle věty 40 lze pro $1 \leq j \leq i-1$ psát

$$d_j^T g_i = 0, \qquad g_j^T g_i = 0. \tag{689}$$

Ukážeme nejprve sestupnou indukcí, že pro libovolný index $i - m \leq k \leq i - 1$ platí

$$H_i^i g_i = \left(\prod_{j=k}^{i-1} \gamma_j^i\right)\left(H_k^i g_i - \sum_{j=k}^{i-1} \frac{y_j^T H_k^i g_i}{y_j^T d_j} d_j\right). \tag{690}$$

Platí to zřejmě pro $k = i - 1$, neboť z (686) a z $d_{i-1}^T g_i = 0$ (přesný výběr délky kroku) plyne, že

$$H_i^i g_i = \gamma_{i-1}^i \left(H_{i-1}^i g_i - \frac{y_{i-1}^T H_{i-1}^i g_i}{y_{i-1}^T d_{i-1}} d_{i-1}\right).$$

Nyní snížíme k o jedničku. Použijeme-li (690), (686) a rovnost $d_{k-1}^T g_i = 0$, která plyne z (689), můžeme psát

$$\begin{aligned} H_i^i g_i &= \left(\prod_{j=k}^{i-1} \gamma_j^i\right)\left(\gamma_{k-1}^i \left(H_{k-1}^i g_i - \frac{y_{k-1}^T H_{k-1}^i g_i}{y_{k-1}^T d_{k-1}} d_{k-1}\right) - \gamma_{k-1}^i \sum_{j=k}^{i-1} \frac{y_j^T H_{k-1}^i g_i}{y_j^T d_j} d_j\right) \\ &= \left(\prod_{j=k-1}^{i-1} \gamma_j^i\right)\left(H_{k-1}^i g_i - \sum_{j=k-1}^{i-1} \frac{y_j^T H_{k-1}^i g_i}{y_j^T d_j} d_j\right). \end{aligned} \tag{691}$$

neboť podle (689) pro $k \leq j \leq i-1$ platí $y_j^T d_{k-1} = (g_{j+1} - g_j)^T d_{k-1} = 0$, takže

$$y_j^T H_k^i g_i = \gamma_{k-1}^i y_j^T \left(H_{k-1}^i g_i - \frac{y_{k-1}^T H_{k-1}^i g_i}{y_{k-1}^T d_{k-1}} d_{k-1}\right) = \gamma_{k-1}^i y_j^T H_{k-1}^i g_i.$$

Jelikož vztah (691) je ekvivalentní s (690), kde k je nahraženo $k-1$, je sestupný indukční krok ukončen. Můžeme tedy psát

$$s_i = -H_i^i g_i = -\left(\prod_{j=i-m}^{i-1} \gamma_j^i\right)\left(g_i - \sum_{j=i-m}^{i-1} \frac{y_j^T g_i}{y_j^T d_j} d_j\right) \tag{692}$$

(neboť $H_{i-m}^k = I$). Podle (689) však pro $j < i-1$ platí $y_j^T g_i = (g_{j+1} - g_j)^T g_i = 0$, takže (692) přejde na (688), čímž je hlavní indukční krok dokončen. □

Důsledek 24. *(Kvadratické ukončení). Nechť jsou splněny předpoklady věty 168. Pak existuje index* $k \leq n$ *takový, že* $g_{k+1} = 0$ *a* $x_{k+1} = x^*$.

Důkaz Podle věty 168 jsou směrové vektory generované $\bar{m}$-krokovou metodou s proměnnou metrikou s omezenou pamětí rovnoběžné s vektory generovanými metodou sdružených gradientů. Podle věty 40 tedy existuje index $k \leq n$ takový, že $g_{k+1} = 0$ a $x_{k+1} = x^*$ (při přesném výběru délky kroku nezáleží na normě směrového vektoru). □

Poznámka 280. Věta 168 a důsledek 24 neplatí pro všechny metody s proměnnou metrikou s omezenou pamětí. Dá se ukázat, že metoda DFP s omezenou pamětí vlastnost kvadratického ukončení nemá. Potíž je v tom, že se k aktualizaci matice H_j^i nepoužívá vektor $d_j = -\alpha_j H_j^i g_j$, nýbrž vektor $d_j = -\alpha_j H_j^j g_j$ získaný pomocí jiné matice. Z tohoto důvodu nejsou všechny metody s proměnnou metrikou s omezenou pamětí a s přesným výběrem délky kroku ekvivalentní (neplatí analogie věty 75).

Jednu z dalších výhod metody BFGS s omezenou pamětí ukazuje tato věta [94].

Věta 169. *Nechť jsou splněny předpoklady věty 168. Pak pro* $1 \le i \le n$ *a* $i-m \le j \le i-1$ *platí*

$$H_i^i y_j = \omega_j^{i-1} \frac{\rho_j^i}{\gamma_j^i} d_j, \quad kde \quad \omega_j^{k-1} = \left(\prod_{l=j}^{k-1} \gamma_l^i \right)$$

(je splněno m kvazinewtonovských podmínek).

Důkaz Poznamenejme, že výraz ω_j^{k-1} by měl obsahovat ještě index i, ale to v tomto důkazu pomineme. Důkaz provedeme indukcí. Předpokládejme, že pro nějaký index $i-m \le k \le i-1$ platí

$$H_k^i y_j = \omega_j^{k-1} \frac{\rho_j^i}{\gamma_j^i} d_j, \qquad i-m \le j \le k-1 \tag{693}$$

(platí to pro $k = i-m$, neboť v tomto případě neexistuje index j splňující nerovnost $i-m \le j \le k-1$). Podle věty 168 jsou směrové vektory s_i, $1 \le i \le n$, G–ortogonální, takže pro $i-m \le j \le k-1$ lze psát

$$d_k^T y_j = 0, \qquad y_k^T d_j = 0. \tag{694}$$

Použijeme-li (686), (693) a (694), dostaneme

$$\begin{aligned} H_{k+1}^i y_j &= \gamma_k^i \left(H_k^i y_j + \left(\frac{a_k^i}{b_k} + \frac{\rho_k^i}{\gamma_k^i} \right) \frac{1}{b_k} d_k d_k^T y_j - \frac{1}{b_k} \left(H_k^i y_k d_k^T y_j + d_k y_k^T H_k^i y_j \right) \right) \\ &= \gamma_k^i \omega_j^{k-1} \frac{\rho_j^i}{\gamma_j^i} \left(d_j + \frac{1}{b_k} y_k^T d_j d_k \right) = \omega_j^k \frac{\rho_j^i}{\gamma_j^i} d_j \end{aligned}$$

pro $i-m \le j \le k-1$ a jelikož podle definice 57 platí $H_{k+1}^i y_k = \rho_k d_k = \omega_k^k (\rho_k^i / \gamma_k^i) d_k$, můžeme psát $H_{k+1}^i y_j = \omega_j^k (\rho_j^i / \gamma_j^i) d_k$ pro $i-m \le j \le k$. Platí tedy (693) pro index o jedničku vyšší. □

Poznámka 281. Numerické testy ukazují, že je výhodné použít pouze počáteční škálování a to zahrnout do výběru matice H_{i-m}^i (obvykle pokládáme $H_{i-m}^i = (y_{i-1}^T d_{i-1} / y_{i-1}^T y_{i-1})\, I$, $i \in N$). Proto budeme předpokládat, že $\gamma_j^i = \gamma_j = 1$, $i-m \le j \le i-1$. Dále budeme předpokládat, že $\rho_j^i = \rho_j$ a $\eta_j^i = \eta_j$, $i-m \le j \le i-1$. To je vcelku logické, neboť korekční parametr ρ_j^i se odvozuje z vlastností minimalizované funkce v okolí bodu x_j, takže index i je irelevantní, a parametr η_j^i se určuje pomocí parametru ρ_j^i.

Nyní dokážeme globální konvergenci metod s proměnnou metrikou s omezenou pamětí. Tak jako v oddílu 4.5 budeme předpokládat, že funkce $F : R^n \to R$ vyhovuje předpokladům F1, F4, F5. Pak (podobně jako v důkazu lemmatu 48) dostaneme

$$\underline{G} \le \frac{y^T y}{y^T d} = \frac{\|y\|^2}{b} \le \overline{G}, \qquad \frac{1}{\overline{G}} \le \frac{d^T d}{y^T d} = \frac{\|d\|^2}{b} \le \frac{1}{\underline{G}}, \tag{695}$$

Lemma 75. *Nechť funkce* F *splňuje předpoklady F1, F4, F5. Nechť pro* $i \in N$ *platí*

$$H_{i-m}^i = \frac{y_{i-1}^T d_{i-1}}{y_{i-1}^T y_{i-1}} I \quad \Longleftrightarrow \quad B_{i-m}^i = \frac{y_{i-1}^T y_{i-1}}{y_{i-1}^T d_{i-1}} I. \tag{696}$$

Pak existují konstanty $0 < K_0 < 1 < C_0$ *takové, že pro* $i \in N$ *platí* $\operatorname{Tr} H_{i-m}^i \le C_0$, $\operatorname{Tr} B_{i-m}^i \le C_0$ *a* $\det B_{i-m}^i \ge K_0$.

Důkaz Použijeme-li nerovnosti (695) spolu s definicí matic H^i_{i-m}, B^i_{i-m} a označíme-li $C_0 = n\max(\overline{G}, 1/\underline{G})$, $K_0 = \underline{G}^n$, dostaneme

$$\begin{aligned}
\operatorname{Tr} H^i_{i-m} &= \frac{y^T_{i-1}d_{i-1}}{y^T_{i-1}d_{i-1}} \operatorname{Tr} I \le \frac{n}{\underline{G}} \le C_0, \\
\operatorname{Tr} B^i_{i-m} &= \frac{y^T_{i-1}y_{i-1}}{y^T_{i-1}d_{i-1}} \operatorname{Tr} I \le n\overline{G} \le C_0, \qquad \det B^i_{i-m} = \left(\frac{y^T_{i-1}y_{i-1}}{y^T_{i-1}d_{i-1}}\right)^n \det I \ge \underline{G}^n = K_0.
\end{aligned}$$

□

Lemma 76. *Nechť funkce F splňuje předpoklady F1, F4, F5. Uvažujme $\bar{m}$-krokovou metodu s proměnnou metrikou s omezenou pamětí (definice 57) s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Pak existují-li konstanty $0 < K < 1 < C$ takové, že pro $i \in N$ platí $\operatorname{Tr} B^i_i \le C$ a $\det B^i_i \ge K$, jsou směrové vektory $s_i = -H^i_i g_i$ stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$.*

Důkaz Jsou-li splněny předpoklady lemmatu 76, můžeme psát

$$\kappa(B^i_i) = \|B^i_i\|\|(B^i_i)\|^{-1} = \frac{\overline{\lambda}(B^i_i)}{\underline{\lambda}(B^i_i)} \le \left(\frac{\overline{\lambda}(B^i_i)}{\underline{\lambda}(B^i_i)}\right)^n \le \frac{(\operatorname{Tr} B^i_i)^n}{\det B^i_i} \le \frac{C^n}{K} \triangleq \overline{\kappa}, \tag{697}$$

takže směrové vektory $s_i = -H^i_i g_i$ jsou podle poznámky 30 stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$.

Věta 170. *Uvažujme $\bar{m}$-krokovou metodu s proměnnou metrikou s omezenou pamětí (definice 57) s výběrem délky kroku splňujícím slabou Wolfeho podmínku takovou, že pro $i \in N$ a $j \in N$ platí (696) a $\gamma_j = 1$, $\underline{\varrho} \le \varrho_j \le \overline{\varrho}$, $0 \le \eta_j \le \overline{\eta}$ (poznámka 281). Nechť funkce $F : \mathcal{D} \to R$ splňuje předpoklady F1, F4, F5. Pak směrové vektory $s_i = -H^i_i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$.*

Důkaz Využijeme toho, že se provádí pouze $m \le \bar{m}$ aktualizací (684), které jsou formálně shodné s aktualizacemi vyšetřovanými v oddílu 4.5. Můžeme tedy použít (685) a postupovat stejně jako v důkazu lemmatu 48. Použijeme-li (413) pro $i - n \le j \le i - 1$, můžeme podle lemmatu 75 psát

$$\|B^i_{j+1}\| \le \operatorname{Tr} B^i_{j+1} \le \overline{K}(\operatorname{Tr} B^i_j + 1) \le \overline{K}^{\bar{m}}(\operatorname{Tr} B^i_{i-m} + 1) \le \overline{K}^{\bar{m}}(C_0 + 1) \triangleq C.$$

Podobně použitím (**??**) a (414) dostaneme

$$\frac{\det B^i_{j+1}}{\det B^i_j} \ge \underline{K}\frac{y^T_j d_j}{d^T_j B^i_j d_j} \ge \frac{\underline{K}}{\|B^i_j\|}\frac{y^T_j d_j}{d^T_j d_j} \ge \frac{\underline{K}\,\underline{G}}{C},$$

takže podle lemmatu 76 platí

$$\det B^i_i \ge \left(\frac{\underline{K}\,\underline{G}}{C}\right)^{\bar{m}} \det B^i_{i-m} \ge \left(\frac{\underline{K}\,\underline{G}}{C}\right)^{\bar{m}} K_0 \triangleq K.$$

Jsou tedy splněny předpoklady lemmatu 76, takže směrové vektory $s_i = -H^i_i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$. □

Metody s proměnnou metrikou s omezenou pamětí lze realizovat přirozeným způsoben tak, že kromě vektorů d_j, y_j, $i - m \le j \le i - 1$ počítáme a ukládáme vektory $H^i_j y_j$, $i - m \le j \le i - 1$. To vyžaduje ukládání dalších m vektorů dimenze n a celkem $O(m^2 n)$ aritmetických operací. Pro některé metody s proměnnou metrikou však existují realizace, které nepotřebují ukládat další vektory dimenze n a počet aritmetických operací je pouze $O(mn)$.

Nejprve se budeme zabývat vektorovou reprezentací metody BFGS, studovanou v práci [125]. Tato reprezentace je založená na pseudosoučinovém tvaru (291), podle kterého pro metodu BFGS platí

$$H^i_{j+1} = V^T_j H^i_j V_j + \frac{\rho_j}{b_j} d_j d^T_j, \qquad V_j = I - \frac{1}{b_j} y_j d^T_j, \tag{698}$$

kde $y_j = g_{j+1} - g_j$, $d_j = x_{j+1} - x_j$ a $b_j = y^T_j d_i$ pro $i - m \le j \le i - 1$.

Věta 171. *Nechť H^i_{j+1} je matice získaná v j-tém kroku metody BFGS. Pak platí*

$$H^i_{j+1} = \left(\prod_{k=i-m}^{j} V_k\right)^T H^i_{i-m} \left(\prod_{k=i-m}^{j} V_k\right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k\right).$$

Důkaz (Indukcí) Pro $j = i - m$ to bezprostředně plyne z (698). Indukční krok vypadá takto

$$\begin{aligned} H^i_{j+1} &= V_j^T H^i_j V_j + \frac{\rho_j}{b_j} d_j d_j^T = V_j^T \left(\prod_{k=i-m}^{j-1} V_k\right)^T H^i_{i-m} \left(\prod_{k=i-m}^{j-1} V_k\right) V_j + \\ &\quad + \sum_{l=i-m}^{j-1} \frac{\rho_l}{b_l} V_j^T \left(\prod_{k=l+1}^{j-1} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j-1} V_k\right) V_j + \frac{\rho_j}{b_j} d_j d_j^T \\ &= \left(\prod_{k=i-m}^{j} V_k\right)^T H^i_{i-m} \left(\prod_{k=i-m}^{j} V_k\right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k\right). \end{aligned} \tag{699}$$

□

Důsledek 25. *Nechť jsou splněny předpoklady věty 171. Pak vektor $s_i = -H^i_i g_i$ lze získat pomocí dvou rekurentních vztahů uvedených v práci [119] (Strangovy rekurence). Nejprve se položí $u_i = -g_i$ a zpětnou rekurencí*

$$\sigma_j = \frac{d_j^T u_{j+1}}{b_j}, \qquad u_j = u_{j+1} - \sigma_j y_j \tag{700}$$

se počítají čísla σ_j a vektory u_j, $i-1 \geq j \geq i-m$. Potom se položí $v_{i-m} = H^i_{i-m} u_{i-m}$ a přímou rekurencí

$$v_{j+1} = v_j + \left(\rho_j \sigma_j - \frac{y_j^T v_j}{b_j}\right) d_j \tag{701}$$

se počítají vektory v_{j+1}, $i - m \leq j \leq i - 1$. Nakonec se položí $s_i = v_i$.

Důkaz Položíme li $u_i = -g_i$ a

$$u_j = -\left(\prod_{k=j}^{i-1} V_k\right) g_i, \quad i - 1 \geq j \geq i - m,$$

vidíme, že platí

$$u_j = V_j u_{j+1} = \left(I - \frac{1}{b_j} y_j d_j^T\right) u_{j+1} = u_{j+1} - \frac{d_j^T u_{j+1}}{b_j} y_j,$$

což je právě rekurence (700). Položíme-li $v_{i-m} = H^i_{i-m} u_{i-m}$ a

$$v_{j+1} = \left(\prod_{k=i-m}^{j} V_k\right)^T H^i_{i-m} u_{i-m} + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T u_{l+1}, \quad i - m \leq j \leq i - 1,$$

vidíme, že platí

$$v_{j+1} = V_j^T v_j + \frac{\rho_j}{b_j} d_j d_j^T u_{j+1} = \left(I - \frac{1}{b_j} d_j y_j^T\right) v_j + \rho_j \sigma^i_j d_j = v_j + \left(\rho_j \sigma_j - \frac{1}{b_j} y_j^T v_j\right) d_j,$$

což je právě rekurence (701). □

Poznámka 282. Tvrzení věty 171 ukazuje, že matici H_i^i můžeme určit z matice H_{i-m}^i (která je řídká) pomocí vektorů d_j, y_j, $i-m \le j \le i-1$. Matici H_i^i nemusíme konstruovat explicitně. Podle důsledku 25 stačí počítat vektor $s_i = -H_i^i g_i$, pomocí rekurentních vztahů (700)–(701). V těchto vztazích je třeba uchovávat čísla σ_j, $i-m \le j \le i-1$. Vektory u_j, v_j, $i-m \le j \le i-1$ mohou být uloženy na stejném místě jako vektor $s_i = -H_i^i g_i$. Pro $m = \bar{m}$, což je maximální možná hodnota, potřebujeme uchovávat $2\bar{m}+3$ vektorů (d_j, y_j, $i-\bar{m} \le j \le i-1$, a 3 vektory x_i, g_i, s_i pro základní optimalizační metodu) a použijeme zhruba $4mn$ operací násobení a sčítání.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 17. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \le \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$ a určíme směrový vektor s_i. Pokud $m = 0$, položíme $s_i := -g_i$, v opačném případě vypočteme s_i pomocí rekurentích vztahů (700)–(701).

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$ a $y_i := g_{i+1} - g_i$ a zvolíme hodnotu parametru ρ_i (obvykle $\rho_i = 1$).

Krok 4 Uložíme vektory d_i, y_i a čísla b_i, ρ_i do pracovního pole. Pokud $m = \bar{m}$, odstraníme vektory d_{i-m}, y_{i-m} a čísla b_{i-m}, ρ_{i-m} z pracovního pole. Zvětšíme i o 1 a přejdeme na krok 2.

Postup, který jsme právě popsali a který tvoří základ algoritmu 17, můžeme použít pouze pro metodu BFGS. Pro ostatní metody z Broydenovy třídy vyvstanou tři problémy.

(1) Strangovy rekurence lze použít pouze pro metodu BFGS, kdy mají příslušné aktualizace tvar (698), takže je můžeme realizovat pomocí rekurentních vztahů vyžadujících zhruba $4mn$ aritmetických operací.
(2) Obecné aktualizace z Broydenovy třídy používají vektory $H_j^i y_j$, $i-m \le j \le i-1$, které nejsou k disposici (bylo by je třeba počítat rekurentně podobným způsobem, jakým se počítá směrový vektor $s_i = -H_i^i g_i$).
(3) I kdybychom vektory $H_j^i y_j$, $i-m \le j \le i-1$, nějak nahradili (například vektory $H_j^j y_j$, $i-m \le j \le i-1$, které se používají v předchozích iteračních krocích), je třeba spočítat nový vektor $H_i^i y_i$, což vyžaduje dalších zhruba $4mn$ aritmetických operací.

Probereme nyní některé možnosti, jak lze tyto potíže obejít. Budeme přitom používat výsledky uvedené v práci [165].

Teoreticky bychom mohli místo (698) použít obecný pseudosoučinový tvar (291). Metody získané tímto způsobem nejsou příliš vhodné, neboť je třeba ukládat m vektorů navíc. Proto je účelné vyjádřit Broydenovu třídu (286) ve tvaru formálně shodném s metodou BFGS, neboli nalézt vektor $\hat{d} = d + \lambda Hy$ tak, aby platilo

$$\frac{1}{\gamma} H_+ = H + \hat{U}\hat{M}\hat{U}^T = H + [\hat{d}, Hy] \begin{bmatrix} \hat{m}_1, & \hat{m}_2 \\ \hat{m}_2, & 0 \end{bmatrix} [\hat{d}, Hy]^T. \tag{702}$$

Lemma 77. *Nechť $UMU^T = \hat{U}\hat{M}\hat{U}^T$, kde $U = [d, Hy]$ a $\hat{U} = [d + \lambda Hy, Hy]$. Pak platí $\det M = \det \hat{M}$.*

Důkaz Zřejmě

$$\begin{aligned} \det(\hat{U}^T\hat{U}) &= (d+\lambda Hy)^T(d+\lambda Hy)(Hy)^T Hy - \left((d+\lambda Hy)^T Hy\right)^2 \\ &= \left(d^T d + 2\lambda d^T Hy + \lambda^2 (Hy)^T Hy\right)(Hy)^T Hy - \left(d^T Hy + \lambda (Hy)^T Hy\right)^2 \\ &= d^T d (Hy)^T Hy - (d^T Hy)^2 = \det(U^T U). \end{aligned}$$

Podle důsledku 9 má matice $\hat{U}\hat{M}\hat{U}^T$ stejná nenulová vlastní čísla jako matice $\hat{U}^T\hat{U}\hat{M}$ a jejich součin je roven číslu $\det(\hat{U}^T\hat{U})\det\hat{M}$. Toto číslo se musí rovnat číslu $\det(U^TU)\det M$ a jelikož $\det(\hat{U}^T\hat{U}) = \det(U^TU)$, platí $\det\hat{M} = \det M$. □

Věta 172. *Aktualizaci z Broydenovy třídy (286), pro kterou platí $\mu \geq 0$, můžeme vyjádřit ve tvaru*

$$\frac{1}{\gamma}H_+ \quad = \quad H + \frac{1}{\hat{b}}\left(\frac{a}{\hat{b}} + \frac{\hat{\rho}}{\gamma}\right)\hat{d}\hat{d}^T - \frac{1}{\hat{b}}\left(Hy\hat{d}^T + \hat{d}(Hy)^T\right), \tag{703}$$

kde

$$\hat{d} = d + \lambda Hy, \quad \lambda = \frac{m_2 + \sqrt{\mu}}{m_1}, \quad \frac{1}{\hat{b}} = \sqrt{\mu}, \quad \frac{\hat{\rho}}{\hat{b}} = \eta\frac{\rho}{b}, \tag{704}$$

přičemž m_1, m_2 jsou odpovídající prvky matice M definované vztahy (283) a $\mu = -\det M$ je výraz určený vzorcem (285).

Důkaz Podle (702) a lemmatu 77 platí $\hat{m}_2^2 = -\det\hat{M} = -\det M = \mu$, kde μ je výraz určený vztahem (285). Protože je vhodné, aby platilo $\hat{b} = -1/\hat{m}_2 > 0$, volíme znaménko tak, že $\hat{m}_2 = -\sqrt{\mu}$. Porovnáme-li koeficienty u dd^T a $Hyd^T + d(Hy)^T$ ve výrazech (271) a (702), dostameme $\hat{m}_1 = m_1$ a $\hat{m}_2 + \lambda\hat{m}_1 = m_2$, což spolu s $\hat{m}_2 = -\sqrt{\mu}$ dává

$$\lambda = \frac{m_2 - \hat{m}_2}{\hat{m}_1} = \frac{m_2 + \sqrt{\mu}}{m_1}.$$

Použijeme-li (283) a (703), můžeme psát

$$m_1 = \hat{m}_1 = \frac{1}{\hat{b}}\left(\frac{a}{\hat{b}} + \frac{\hat{\rho}}{\gamma}\right) = a\mu + \frac{\hat{\rho}}{\gamma\hat{b}} = \frac{1}{b}\left(\eta\frac{a}{b} + (1-\eta)\frac{\rho}{\gamma}\right) + \frac{\hat{\rho}}{\gamma\hat{b}} = m_1 - \eta\frac{\rho}{\gamma b} + \frac{\hat{\rho}}{\gamma\hat{b}},$$

což implikuje rovnost $\hat{\rho}/\hat{b} = \eta\rho/b$. Pokud $\mu = 0$ (metoda hodnosti 1), platí $1/\hat{b} = 0$, $\lambda = m_2/m_1 = -\gamma/\rho$, $\eta = \rho b/(\rho b + \gamma a)$ a $\hat{\rho}/\hat{b} = \eta\rho/b$. □

Poznámka 283. Pokud $\gamma = 1$, můžeme vzorec (703) zapsat ve tvaru

$$H_+ = \hat{V}^T H\hat{V} + \frac{\hat{\rho}}{\hat{b}}\hat{d}\hat{d}^T, \qquad \hat{V} = I - \frac{1}{\hat{b}}y\hat{d}^T, \tag{705}$$

Tento pseudosoučinový vzorec je velmi vhodný pro implementaci obecných metod s proměnnou metrikou s omezenou pamětí. Má však jednu nevýhodu, neplatí $\hat{b} = y^T\hat{d}$, takže matice $\hat{V}$ není maticí projekce (neplatí $\hat{V}^2 = \hat{V}$). Abychom získali pseudosoučinový vzorec, kde matice $\hat{V}$ je maticí projekce, použijeme další transformaci $\hat{y} = y + \omega B\hat{d}$, kde $B = H^{-1}$. Podobně jako v důkazu lematu 77 se dá ukázat, že tato transformace nemění hodnotu determinantu matice $\hat{M}$ a zachovává vlastnosti metody BFGS ($\hat{m}_3 = 0$). Odtud plyne, že se hodnota prvku $\hat{m}_2$ nemění, takže číslo ω lze zvolit tak, aby platilo $\hat{y}^T\hat{d} = \hat{b} = 1/\sqrt{\mu}$. Jelikož $\hat{y}^T\hat{d} = y^T\hat{d} + \omega\hat{d}^TB\hat{d}$, volíme $\omega = (1/\sqrt{\mu} - y^T\hat{d})/\hat{d}^TB\hat{d}$. Dostaneme tak aktualizaci

$$H_+ = \hat{V}^T H\hat{V} + \frac{\hat{\rho}}{\hat{b}}\hat{d}\hat{d}^T, \qquad \hat{V} = I - \frac{1}{\hat{b}}\hat{y}\,\hat{d}^T, \tag{706}$$

kde $\hat{d} = d + \lambda Hy$, $\hat{y} = y + \omega B\hat{d}$, $\hat{b} = \hat{y}^T\hat{d}$,

$$\omega = \frac{1/\sqrt{\mu} - y^T\hat{d}}{\hat{c}}, \qquad \hat{\rho} = \eta\hat{b} + \omega^2\frac{\hat{c}}{\hat{b}}, \tag{707}$$

kde $\hat{c} = \hat{d}^TB\hat{d} = (\alpha g - \lambda y)^T\hat{d}$ (předpokládáme, že $s = -Hg$, takže $Bd = \alpha Bs = -\alpha g$).

Poznámka 284. Pseudosoučinový vzorec (705) bychom mohli použít k realizaci libovolné rozložitelné metody s proměnnou metrikou s omezenou pamětí z Broydenovy třídy (kdy $\mu \geq 0$) pomocí Strangových rekurencí ve kterých bychom vektory d_j, $i-m \leq j \leq i-1$, nahradili vektory $\hat{d}_j = d_j + \lambda_j H_j^i y_j$. Jak již bylo poznamenáno, vektory $H_j^i y_j$, $i-m \leq j \leq i-1$, nelze získat jednoduchým výpočtem, takže místo nich používáme vektory $H_j^j y_j$, které známe z předchozích iteračních kroků. Touto úpravou však přijdeme o některé výhodné teoretické vlastnosti (není splněna kvazinewtonovská podmínka a neplatí věta 168 ani její důsledek 24). Globální konvergence však zůstane zachována, vyhovuje-li výběr parametru η_i, $i \in N$, těmto předpokladům:

(1) Jsou splněny nerovnosti $\|\hat{d}_i\| \leq \Delta \|d_i\|$, kde $\Delta > 0$ je vhodně zvolená (velká) konstanta, která nezávisí na indexu i.
(2) Platí $\underline{\eta} \leq \eta_i \leq \overline{\eta}$ a $0 < (\mu_i/\eta_i) b_i^2 \leq \overline{\eta}$, kde $0 < \underline{\eta} < 1 < \overline{\eta}$.

Platnost prvního předpokladu lze zajistit tak, že položíme $\eta_i = 1$ (metoda BFGS), není-li pro zvolenou hodnotu parametru η_i splněna požadovaná nerovnost (pak $\|\hat{d}_i\| = \|d_i\|$). Platnost druhého předpokladu lze zajistit použitím následujícího lemmatu

Lemma 78. *Nechť $0 < \underline{\eta} < 1 < \overline{\eta}$ a*

$$\begin{aligned} \underline{\eta}_i &= \max\left(\underline{\eta}, \frac{\rho_i b_i}{\rho_i b_i + (\overline{\eta}-1)a_i}\right), \\ \overline{\eta}_i &= \overline{\eta}, \qquad\qquad \rho_i b_i \leq a_i, \\ \overline{\eta}_i &= \min(\overline{\eta}, \eta_i^{R1}), \qquad \rho_i b_i > a_i. \end{aligned}$$

Pak jestliže $\underline{\eta}_i \leq \eta_i < \overline{\eta}_i$, platí $\underline{\eta} \leq \eta_i \leq \overline{\eta}$ a $0 < (\mu_i/\eta_i) b_i^2 \leq \overline{\eta}$.

Důkaz Použijeme-li (285), dostaneme

$$\frac{\mu_i}{\eta_i} b_i^2 = \frac{a_i - \rho_i b_i}{a_i} + \frac{\rho_i b_i}{a_i} \frac{1}{\eta_i}. \tag{708}$$

Podle poznámky 140 platí $(\mu_i/\eta_i) b_i^2 > 0$, pokud $\rho_i b_i \leq a_i$, nebo pokud $\rho_i b_i > a_i$ a $\vartheta_1 < \eta_i^{R1}$. Z vyjádření (708) plyne, že nerovnost $(\mu_i/\eta_i) b_i^2 \leq \overline{\eta}$ je splněna, pokud

$$\frac{a_i - \rho_i b_i}{a_i} + \frac{\rho_i b_i}{a_i} \frac{1}{\eta_i} \leq \overline{\eta},$$

neboli

$$\eta_i \geq \frac{\rho_i b_i}{\rho_i b_i + (\overline{\eta}-1)a_i}.$$

□

K důkazu globální konvergence metod, založených na zobecněných pseudosoučinnových vzorcích (705) a (706) nelze použít postup uvedený v důkazu věty 170 (neplatí $H_j^j y_j = H_j^i y_j$, $i-m \leq j \leq i-1$). Lze však použít následující vlastnost zobecněných pseudosoučinnových vztahů.

Lemma 79. *Nechť H je pozitivně definitní matice, $u \in R^n$, $v \in R^n$, a $\vartheta > 0$. Pak matice*

$$H_+ = \tau^2 \vartheta u u^T + (I - \tau u v^T) H (I - \tau v u^T) \tag{709}$$

je pozitivně definitní a platí

$$\mathrm{Tr}(H_+) \leq \tau^2 \vartheta \|u\|^2 + \mathrm{Tr}(H)(1 + |\tau| \|u\| \|v\|)^2, \tag{710}$$

$$\mathrm{Tr}(H_+^{-1}) \leq \mathrm{Tr}(H^{-1}) + \frac{1}{\vartheta} \|v\|^2. \tag{711}$$

Důkaz (a) Z vyjádření (709) plyne, že matice H_+ je pozitivně definitní, pokud $\vartheta > 0$, neboť pak pro $\tau^2 > 0$ a $w^T u \neq 0$ platí $w^T H_+ w \geq \tau^2 \vartheta (w^T u)^2 > 0$, a pro $\tau^2 = 0$ nebo $w^T u = 0$ platí $w^T H_+ w = w^T H w > 0$ (pokud $\|w\| > 0$), neboť matice H je pozitivně definitní.

(b) Vztah (709) můžeme zapsat ve tvaru

$$H_+ = H + \tau^2(\vartheta + v^T H v)uu^T - \tau(Hvu^T + uv^T H). \tag{712}$$

Použijeme-li (712), dostaneme

$$\begin{aligned}
\operatorname{Tr}(H_+) &= \operatorname{Tr}(H) + \tau^2(\vartheta + v^T H v)u^T u - 2\tau u^T H v \\
&\leq \operatorname{Tr}(H) + \tau^2(\vartheta + \operatorname{Tr}(H)\|v\|^2)\|u\|^2 + 2\,|\tau|\sqrt{u^T H u v^T H v} \\
&\leq \operatorname{Tr}(H) + \tau^2(\vartheta + \operatorname{Tr}(H)\|v\|^2)\|u\|^2 + 2\,|\tau|\operatorname{Tr}(H)\|u\|\|v\| \\
&\leq \tau^2\vartheta\|u\|^2 + \operatorname{Tr}(H)(1 + |\tau|\|u\|\|v\|)^2.
\end{aligned}$$

(c) Ukážeme, že pro dva vektory $\bar{u} \in R^n$, $\bar{v} \in R^n$ platí

$$\left(I + (\bar{u} - \bar{v})(\bar{u} - \bar{v})^T - \bar{v}\bar{v}^T\right)^{-1} = I + \frac{\bar{v}\bar{v}^T}{1 - \|\bar{v}\|^2} - \frac{(\bar{u} - \theta\bar{v})(\bar{u} - \theta\bar{v})^T}{\|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2)}, \qquad \theta = \frac{1 - \bar{u}^T\bar{v}}{1 - \|\bar{v}\|^2} \tag{713}$$

(pokud jsou oba jmenovatele nenulové). Označme $W = I - \bar{v}\bar{v}^T$ a $w = \bar{u} - \bar{v}$. Pak podle lemmatu 31 (e) platí

$$W^{-1} = I + \frac{\bar{v}\bar{v}^T}{1 - \bar{v}\bar{v}^T}, \qquad (W + ww^T)^{-1} = W^{-1} - \frac{W^{-1}ww^TW^{-1}}{1 + w^TW^{-1}w}. \tag{714}$$

První rovnost v (714) odpovídá prvnímu členu v (713). Druhou rovnost dále upravíme. Zřejmě

$$W^{-1}w = w + \frac{\bar{v}^T w}{1 - \|\bar{v}\|^2}\bar{v} = \bar{u} - \left(1 - \frac{(\bar{u} - \bar{v})^T\bar{v}}{1 - \|\bar{v}\|^2}\right)\bar{v} = \bar{u} - \frac{1 - \bar{u}^T\bar{v}}{1 - \|\bar{v}\|^2}\bar{v} = \bar{u} - \theta\bar{v},$$

což dává

$$\begin{aligned}
1 + w^TW^{-1}w &= 1 + w^T\bar{u} - \theta w^T\bar{v} = 1 + \|\bar{u}\|^2 - \bar{u}^T\bar{v} - \theta(\bar{u} - \bar{v})^T\bar{v} \\
&= \|\bar{u}\|^2 + \theta(1 - \|\bar{v}\|^2 - \bar{u}^T\bar{v} + \|\bar{v}\|^2) = \|\bar{u}\|^2 + \theta(1 - \bar{u}^T\bar{v}) = \|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2).
\end{aligned}$$

Dosadíme-li oba tyto vztahy do druhé rovnosti v (714), dostaneme druhý člen v (713).

(d) Pro $\tau = 0$ je nerovnost (711) triviální. Nechť $\tau \neq 0$ a $\omega = \tau(\vartheta + v^T H v)$ (takže $\tau\omega > 0$). Pak vzorec (709) můžeme zapsat ve tvaru

$$\begin{aligned}
H_+ &= H + \frac{\tau}{\omega}\left((\omega u - Hv)(\omega u - Hv)^T - Hvv^TH\right) \\
&= H^{1/2}\left(I + (\bar{u} - \bar{v})(\bar{u} - \bar{v})^T - \bar{v}\bar{v}^T\right)H^{1/2},
\end{aligned} \tag{715}$$

kde $\bar{u} = \sqrt{\tau\omega}H^{-1/2}u$ a $\bar{v} = \sqrt{\tau/\omega}H^{1/2}v$. Protože

$$1 - \|\bar{v}\|^2 = 1 - \frac{\tau}{\omega}v^THv = 1 - \frac{\tau}{\omega}\left(\frac{\omega}{\tau} - \vartheta\right) = \frac{\tau}{\omega}\vartheta > 0, \tag{716}$$

můžeme podle (713), (715), (716) psát

$$\operatorname{Tr}(H_+^{-1}) = \operatorname{Tr}(H^{-1}) + \frac{\bar{v}^TH^{-1}\bar{v}}{1 - \|\bar{v}\|^2} - \frac{(\bar{u} - \theta\bar{v})^TH^{-1}(\bar{u} - \theta\bar{v})}{\|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2)} \leq \operatorname{Tr}(H^{-1}) + \frac{\bar{v}^TH^{-1}\bar{v}}{1 - \|\bar{v}\|^2} = \operatorname{Tr}(H^{-1}) + \frac{v^Tv}{\vartheta}.$$

□

Věta 173. *Uvažujme metodu s proměnnou metrikou s omezenou pamětí (703)–(705) s výběrem délky kroku splňujícím slabou Wolfeho podmínku, která vyhovuje předpokladům uvedeným v poznámce 284. Nechť funkce F splňuje předpoklady F1, F4, F5. Pak směrové vektory $s_i = -H_i^i g_i$ jsou stejnoměrně spádové a platí* $\lim_{i\to\infty} \|g_i\| = 0$.

Důkaz (a) Je zřejmé, že vztah (705) má tvar (709), kde $u = \hat{d}$, $v = y$, $\tau = 1/\hat{b} = \sqrt{\mu}$, $\tau^2\vartheta = \hat{\rho}/\hat{b} = \eta\rho/b$ a $1/\vartheta = \mu b/(\eta\rho)$. Jelikož podle lemmatu 78 pro $i - n \le j \le i - 1$ platí

$$|\tau_j| = \sqrt{\mu_j} = \sqrt{\frac{\mu_j b_j^2}{\eta_j}}\frac{\sqrt{\eta_j}}{b_j} \le \frac{\overline{\eta}}{b_j},$$

můžeme psát

$$\begin{aligned}
\operatorname{Tr} H_{j+1}^i &\le \eta_j\rho_j\frac{\|\hat{d}_j\|^2}{b_j} + \operatorname{Tr} H_j^i\left(1 + \overline{\eta}\frac{\|\bar{d}_j\|\|y_j\|}{b_j}\right)^2 \le \overline{\eta}\,\overline{\rho}\Delta^2\frac{\|d_j\|^2}{b_j} + \operatorname{Tr} H_j^i\left(1 + \overline{\eta}\Delta\frac{\|d_j\|\|y_j\|}{b_j}\right)^2 \\
&\le \frac{\overline{\eta}\,\overline{\rho}\Delta^2}{\underline{G}} + \operatorname{Tr} H_j^i\left(1 + \overline{\eta}\Delta\sqrt{\frac{\overline{G}}{\underline{G}}}\right)^2 \triangleq C_1 + C_2\operatorname{Tr} H_j^i \\
\operatorname{Tr} B_{j+1}^i &\le \operatorname{Tr} B_j^i + \frac{\mu_j b_j^2}{\eta_j\rho_j}\frac{\|y_j\|^2}{b_j} \le \operatorname{Tr} B_j^i + \frac{\overline{\eta}}{\underline{\rho}}\overline{G} \triangleq \operatorname{Tr} B_j^i + C_3.
\end{aligned}$$

(b) Podle (a) a lemmatu 75 platí

$$\operatorname{Tr} H_i^i \le \max(C_0, C_1)(1 + C_2 + C_2^2 + \cdots + C_2^{\overline{m}}) \triangleq \overline{C}, \qquad \operatorname{Tr} B_i^i \le C_0 + \overline{m}C_3 \triangleq C,$$

takže

$$\kappa(H_i^i) = \|H_i^i\|\|B_i^i\| \le \operatorname{Tr} H_i^i \operatorname{Tr} B_i^i \le \overline{C}C \triangleq \overline{\kappa}, \tag{717}$$

takže směrové vektory $s_i = -H_i^i g_i$ jsou podle poznámky 30 stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$. □

Používáme-li pseudosoučinový tvar (705), je třeba kromě směrových vektorů $s_i = -H_i g_i$, $i \in N$, počítat vektory $H_i y_i$, $i \in N$, což zvyšuje počet aritmetických operací. K odstranění této nevýhody je vhodné počítat směrový vektor podle některého ze vzorců (452) nebo (458) a Strangovy rekurence použít pouze pro výpočet vektoru $H_i y_i$. Problém je v tom, že pro takto získané směrové vektory platí $s_i = -H_i^i g_i$ pouze tehdy, když $s_{i-1} = -H_{i-1}^i g_{i-1}$. To obecně splněno není, neboť matice H_{i-1}^i se liší od matice H_{i-1}^{i-1}, použité pro výpočet směrového vektoru s_{i-1}.

Poznámka 285. Kdybychom použili vzorec (452) v případě, že $d \ne -\alpha Hg$, mohlo by se stát, že by získaný směrový vektor nebyl spádový. Proto je výhodnější použít vzorec

$$s_+ = -\frac{d^T g_+}{b}d - \frac{b + \eta\tau}{b + \tau}V^T p, \qquad V = I - \frac{1}{b}yd^T, \tag{718}$$

kde $p = HVg_+$ a $\tau = \max(\alpha\, p^T y, b)$, který je modifikací vzorce (458) a který zaručuje, že $g_+^T s_+ < 0$ (poznámka 180). Vektor $p = HVg_+$ můžeme (tak jako v důkazu věty 108) použít k výpočtu vektoru Hy. Položíme

$$Hy = \frac{b}{c}(d + \alpha p), \qquad a = y^T Hy = \frac{b}{c}(b + \alpha y^T p), \qquad b = d^T y, \qquad c = -\alpha d^T g. \tag{719}$$

Poznámka 286. K výpočtu vektoru $p_i = H_i V_i g_{i+1}$ lze použít modifikované Strangovy rekurence (odvozené z aktualizace (705)), kde vystupují veličiny se stříškou. Nejprve položíme

$$u_i = V_i g_{i+1} = g_{i+1} - \frac{d_{i-1}^T g_{i+1}}{y_{i-1}^T d_{i-1}} y_{i-1}$$

a zpětnou rekurencí

$$\sigma_j = \frac{\hat{d}_j^T u_{j+1}}{\hat{b}_j}, \qquad u_j = u_{j+1} - \sigma_j^i y_j \tag{720}$$

spočítáme čísla σ_j^i a vektory u_j, $i-1 \geq j \geq i-m$. Potom položíme $v_{i-m} = H_{i-m}^i u_{i-m}$ a přímou rekurencí

$$v_{j+1} = v_j + \left(\hat{\rho}_j \sigma_j - \frac{y_j^T v_j}{\hat{b}_j}\right) \hat{d}_j, \tag{721}$$

spočítáme vektory v_{j+1}, $i-m \leq j \leq i-1$. Nakonec položíme $p_i = v_i$.

9.2 Modifikované vektorové metody s proměnnou metrikou s omezenou pamětí

Modifikovaná vektorová metoda s proměnnou metrikou s omezenou pamětí, studovaná v práci [164], je založena na Strangových rekurencích (700), (701), kde se místo vektorů d_j, y_j používají transformované vektory $\bar{d}_j$, $\bar{y}_j$, $i-m \leq j \leq i-1$. Platí tedy

$$H_{j+1}^i = \bar{V}_j^T H_j^i \bar{V}_j + \frac{1}{\bar{b}_j} \bar{d}_j \bar{d}_j^T, \qquad \bar{V}_j = I - \frac{1}{\bar{b}_j} \bar{y}_j \bar{d}_j^T \tag{722}$$

(předpokládáme že $\bar{\rho}_j = 1$) pro $i-m \leq j \leq i-1$, kde $H_{i-m}^i = (d_i^T y_i / y_i^T y_i) I$. Přitom $\bar{d}_1 = d_1$, $\bar{y}_1 = y_1$ a $\bar{d}_i = d_i - \lambda_i \bar{d}_{i-1}$, $\bar{y}_i = y_i - \omega_i \bar{y}_{i-1}$, $i > 1$. Podstata této metody spočívá v transformaci vektorů d_i, y_i na $\bar{d}_i$, $\bar{y}_i$ pomocí vektorů $\bar{d}_{i-1}$, $\bar{y}_{i-1}$ získaných v předchozím iteračním kroku. Využívá se toho, že matice H_i^{i+1}, konstruovaná podle (722), splňuje kvazinewtonovskou podmínku $H_i^{i+1} \bar{y}_{i-1} = \bar{d}_{i-1}$. Jak bylo poznamenáno na konci oddílu 4.2, nelze pomocí aktualizací z Broydenovy třídy sestrojit symetrickou matici H_{i+1}^{i+1} splňující současně dvě kvazinewtonovské podmínky $H_{i+1}^{i+1} y_{i-1} = d_{i-1}$, $H_{i+1}^{i+1} y_i = d_i$ (k tomu je potřeba, aby platilo $d_i^T y_{i-1} = d_{i-1}^T y_i$). Ukážeme, že vhodnou volbou parametrů λ_i, ω_i lze docílit toho, že platí $\bar{d}_i^T \bar{y}_{i-1} = 0$, $\bar{d}_{i-1}^T \bar{y}_i = 0$, což stačí k tomu, aby modifikovaná aktualizace splňovala dvě kvazinewtonovské podmínky $H_{i+1}^{i+1} \bar{y}_{i-1} = \bar{d}_{i-1}$, $H_{i+1}^{i+1} \bar{y}_i = \bar{d}_i$ (věta 174) a aby vektory $\bar{d}_i$, $\bar{d}_{i-1}$ byly dvojnásobně konjugované (věta 175). Pro zjednodušení popisu vyšetřované aktualizace budeme horní index $i+1$ a dolní index i vynechávat a dolní indexy $i-1$, $i+1$ nahradíme symboly $-$, $+$. Pak místo $H_i^{i+1} \bar{y}_{i-1} = \bar{d}_{i-1}$, $H_{i+1}^{i+1} \bar{y}_i = \bar{d}_i$ píšeme $H \bar{y}_- = \bar{d}_-$, $H_+ \bar{y} = \bar{d}$, kde

$$H_+ = \bar{V}^T H \bar{V} + \frac{1}{\bar{b}} \bar{d}\, \bar{d}^T, \qquad \bar{V} = I - \frac{1}{\bar{b}} \bar{y}\, \bar{d}^T \tag{723}$$

a $\bar{d} = d - \lambda \bar{d}_-$, $\bar{y} = y - \omega \bar{y}_-$. Používáme přitom označení $\bar{a}_- = \bar{y}_-^T H_- \bar{y}_-$, $\bar{b}_- = \bar{y}_-^T \bar{d}_-$, $\bar{c}_- = \bar{d}_-^T B_- \bar{d}_-$, kde $B_- = H_-^{-1}$, a $\bar{a} = \bar{y}^T H \bar{y}$, $\bar{b} = \bar{y}^T \bar{d}$, $\bar{c} = \bar{d}^T B \bar{d}$, kde $B = H^{-1}$.

Věta 174. *Nechť H je symetrická pozitivně definitní matice taková, že $H\bar{y}_- = \bar{d}_-$, a H_+ je matice získaná aktualizací (723), kde $\bar{b} > 0$. Pak jestliže $\bar{d}^T \bar{y}_- = 0$, $\bar{d}_-^T \bar{y} = 0$, neboli*

$$\lambda = d^T \bar{y}_- / \bar{b}_-, \qquad \omega = \bar{d}_-^T y / \bar{b}_-, \tag{724}$$

platí $H_+ \bar{y}_- = \bar{d}_-$.

Důkaz Jelikož (723) odpovídá aktualizaci BFGS, můžeme podle (288) a (309) psát

$$H_+ = H + \left(1 + \frac{\bar{a}}{\bar{b}}\right) \frac{\bar{d}\bar{d}^T}{\bar{b}} - \frac{H\bar{y}\bar{d}^T + \bar{d}(H\bar{y})^T}{\bar{b}}, \qquad B_+ = B + \frac{\bar{y}\bar{y}^T}{\bar{b}} - \frac{B\bar{d}(B\bar{d})^T}{\bar{c}}. \tag{725}$$

Použitím těchto rovností dostaneme

$$\bar{y}_-^T H_+ \bar{y}_- = \bar{d}_-^T \bar{y}_- + \left(1 + \frac{\bar{a}}{\bar{b}}\right) \frac{(\bar{d}^T \bar{y}_-)^2}{\bar{b}} - 2 \frac{\bar{d}^T \bar{y}_- \bar{d}_-^T \bar{y}}{\bar{b}} \tag{726}$$

$$\bar{d}_-^T B_+ \bar{d}_- = \bar{d}_-^T \bar{y}_- + \frac{(\bar{d}_-^T \bar{y})^2}{\bar{b}} - \frac{(\bar{d}^T \bar{y}_-)^2}{\bar{c}}, \tag{727}$$

takže

$$\begin{aligned}(H_+\bar{y}_- - \bar{d}_-)^T B_+(H_+\bar{y}_- - \bar{d}_-) &= \bar{y}_-^T H_+\bar{y}_- + \bar{d}_-^T B_+\bar{d}_- - 2\bar{d}_-^T\bar{y}_- \\ &= \frac{1}{\bar{\bar{b}}}\left((\bar{d}_-^T\bar{y} - \bar{d}^T\bar{y}_-)^2 + (\bar{d}^T\bar{y}_-)^2\left(\frac{\bar{\bar{a}}}{\bar{\bar{b}}} - \frac{\bar{\bar{b}}}{\bar{c}}\right)\right). \qquad (728)\end{aligned}$$

Pokud $\bar{d}_-^T\bar{y} = 0$ a $\bar{d}^T\bar{y}_- = 0$, je tento výraz nulový, takže $H_+\bar{y}_- = \bar{d}_-$ (jelikož $\bar{b} > 0$, je matice B_+ pozitivně definitní). Z rovnic $\bar{d}_-^T\bar{y} = 0$, $\bar{d}^T\bar{y}_- = 0$ a z vyjádření $\bar{d} = d - \lambda\bar{d}_-$, $\bar{y} = y - \omega\bar{y}_-$ dostaneme vztahy pro parametry λ a ω. □

Splnění dvou kvazinewtonovských podmínek má úzký vztah ke dvojnásobné konjugovanosti.

Věta 175. *Nechť jsou splněny předpoklady věty 174. Jsou-li vektory $\bar{d}$, $H\bar{y}$ lineárně nezávislé, je podmínka $H_+\bar{y}_- = \bar{d}_-$ splněna právě tehdy, když $\bar{d}^T B\bar{d}_- = 0$ a $\bar{d}^T B_+\bar{d}_- = 0$.*

Důkaz Podle předpokladu platí $\bar{y}_- = B\bar{d}_-$ a použitím (725) dostaneme

$$\begin{aligned}H_+\bar{y}_- &= H_+B\bar{d}_- = HB\bar{d}_- + \left(1 + \frac{\bar{\bar{a}}}{\bar{\bar{b}}}\right)\frac{\bar{d}^T B\bar{d}_-}{\bar{\bar{b}}}\bar{d} - \frac{\bar{d}^T B\bar{d}_-}{\bar{\bar{b}}}H\bar{y} - \frac{\bar{y}^T HB\bar{d}_-}{\bar{\bar{b}}}\bar{d} \\ &= \bar{d}_- + \left(\frac{\bar{d}^T B\bar{d}_- - \bar{y}^T\bar{d}_-}{\bar{\bar{b}}} + \frac{\bar{\bar{a}}}{\bar{\bar{b}}}\frac{\bar{d}^T B\bar{d}_-}{\bar{\bar{b}}}\right)d - \frac{\bar{d}^T B\bar{d}_-}{\bar{\bar{b}}}H\bar{y},\end{aligned}$$

takže $H_+\bar{y}_- = \bar{d}_-$ platí právě tehdy, když $\bar{d}^T B\bar{d}_- = 0$ a $\bar{y}^T\bar{d}_- = \bar{y}^T H_+B_+\bar{d}_- = \bar{d}^T B_+\bar{d}_- = 0$ (neboť $H_+\bar{y} = \bar{d}$ podle (723)). □

Z vyjádření (728) je patrné, že rozhodující vliv na velikost tohoto výrazu má nulovost čísla $\bar{d}^T\bar{y}_-$, neboť rozdíl $\bar{a}/\bar{b} - \bar{\bar{b}}/\bar{c}$, který je podle Schwarzovy nerovnosti nezáporný (poznámka 111), může být velmi velký. Proto budeme vždy volit $\lambda = d^T\bar{y}_-/\bar{b}_-$.

Věta 176. *Nechť jsou splněny předpoklady věty 174 a $\lambda = d^T\bar{y}_-/\bar{b}_-$. Pak hodnota $\omega = \bar{d}_-^T y/\bar{b}_-$ minimalizuje spektrální číslo podmíněnosti matice $B^{1/2}H_+B^{1/2}$.*

Důkaz Jestliže $\lambda = d^T\bar{y}_-/\bar{b}_-$, platí $\bar{d}^T\bar{y}_- = 0$, takže $\bar{b} = \bar{d}^T(y - \omega\bar{y}_-) = \bar{d}^T y$. Čísla $\bar{b}$, $\bar{c}$ tedy nezávisejí na hodnotě parametru ω. Podle lemmatu 32 má matice $B^{1/2}H_+B^{1/2}$ $n-2$ jednotkových vlastních čísel a zbylá dvě vlastní čísla $\underline{\lambda}$, $\overline{\lambda}$ jsou řešením kvadratické rovnice $\lambda^2 - \bar{\sigma}\lambda + \bar{\delta} = 0$, přičemž pro metodu BFGS platí

$$\bar{\sigma} = \frac{\bar{a}\bar{c}}{\bar{\bar{b}}^2} + \frac{\bar{c}}{\bar{\bar{b}}}, \qquad \bar{\delta} = \frac{\bar{c}}{\bar{\bar{b}}}$$

(používáme tabulku uvedenou v poznámce 124, kam dosazujeme hodnoty s pruhem). Podle lemmatu 44 (b) pro metodu BFGS platí $0 < \underline{\lambda} \leq 1 \leq \overline{\lambda}$, (neboť $\bar{\mu} = 1/\bar{\bar{b}}^2 \geq 0$). Spektrální číslo podmíněnosti matice $B^{1/2}H_+B^{1/2}$ je tedy minimální, je-li podíl $\overline{\lambda}/\underline{\lambda}$ minimální. Z důkazu lemmatu 44 plyne, že tento podíl je minimální, je-li výraz $\bar{\sigma}/(2\sqrt{\bar{\delta}})$ minimální. Jelikož čísla $\bar{b}$ a $\bar{c}$ nezávisejí na parametru ω, je tento výraz minimální, je-li číslo $\bar{a}$ minimální. Jelikož

$$\bar{a} = \bar{y}^T H\bar{y} = y^T Hy - 2\omega\bar{y}_-^T Hy + \omega^2\bar{y}_-^T H\bar{y}_- = y^T Hy - 2\omega\bar{d}_-^T y + \omega^2\bar{d}_-^T\bar{y}_-$$

je číslo $\bar{a}$ minimální, pokud $\omega = \bar{d}_-^T y/\bar{b}_-$. □

Ukážeme, že pokud $H_+\bar{y}_- = \bar{d}_-$, neboli $\lambda = d^T\bar{y}_-/\bar{b}_-$ a $\omega = \bar{d}_-^T y/\bar{b}_-$, je matice H_+ v jistém smyslu blíže k matici H než k matici H_-.

Věta 177. *Nechť jsou splněny předpoklady věty 174. Pak jestliže $H_+\bar{y}_- = \bar{d}_-$, platí*

$$\|H_+^{-1/2}HH_+^{-1/2} - I\|_F^2 \leq \|H_+^{-1/2}H_-H_+^{-1/2} - I\|_F^2 - \left(1 - \frac{\bar{a}_-}{\bar{\bar{b}}_-}\right)^2.$$

Důkaz Položme $\tilde{G}_- = H_+^{-1}$, $R_-^{-1} = \tilde{G}_-^{1/2} H_- \tilde{G}_-^{1/2}$, $(R')^{-1} = \tilde{G}_-^{1/2} H \tilde{G}_-^{1/2}$ a $\bar{z}_- = \tilde{G}_-^{1/2} \bar{d}_- = \tilde{G}_-^{-1/2} \bar{y}_-$. Pak $\tilde{G}_- \bar{d}_- = \bar{y}_-$ a podle poznámky 173 platí

$$\begin{aligned}
\|(R')^{-1} - I\|_F^2 &= \|R_-^{-1} - I\|_F^2 - \left(1 - \frac{\bar{z}_-^T R_-^{-1} \bar{z}_-}{\bar{z}_-^T \bar{z}_-}\right)^2 - 2\left(\frac{\bar{z}_-^T R_-^{-2} \bar{z}_-}{\bar{z}_-^T \bar{z}_-} - \left(\frac{\bar{z}_-^T R_-^{-1} \bar{z}_-}{\bar{z}_-^T \bar{z}_-}\right)^2\right) \\
&\leq \|R_-^{-1} - I\|_F^2 - \left(1 - \frac{\bar{z}_-^T R_-^{-1} \bar{z}_-}{\bar{z}_-^T \bar{z}_-}\right)^2 = \|R_-^{-1} - I\|_F^2 - \left(1 - \frac{\bar{a}_-}{\bar{b}_-}\right)^2
\end{aligned}$$

(používáme (427) s $\eta = 1$). □

Zvolíme-li neoptimální hodnotu parametru ω, není splněna kvazinewtonovská podmínka $H_+ \bar{y}_- = \bar{d}_-$. Vhodnou volbou parametru ω však můžeme zmenšit porušení standardní kvazinewtonovské podmínky $H_+ y = d$, neboť to je podle následující věty závislé na rozdílu hodnot ω a λ.

Věta 178. *Nechť jsou splněny předpoklady věty 174. Pak jestliže $\lambda = d^T \bar{y}_- / \bar{b}_-$, platí*

$$(H_+ y - d)^T B_+ (H_+ y - d) = (\omega - \lambda)^2 \bar{b}_- + \lambda^2 (\bar{d}_-^T \bar{y})^2 / \bar{b}. \tag{729}$$

Pokud navíc $\omega = \bar{d}_-^T y / \bar{b}_-$, je druhý člen v (729) nulový.

Důkaz Jestliže $\lambda = d^T \bar{y}_- / \bar{b}_-$, platí $\bar{d}^T \bar{y}_- = 0$, což po dosazení do (726) a (727) dává $\bar{y}_-^T H_+ \bar{y}_- = \bar{b}_-$ a $\bar{d}_-^T B_+ \bar{d}_- = \bar{b}_- + (\bar{d}_-^T \bar{y})^2 / \bar{b}$. Dále podle (723) platí

$$H_+ y - d = H_+ \bar{y} + \omega H_+ \bar{y}_- - d = \bar{d} + \omega H_+ \bar{y}_- - d = \omega H_+ \bar{y}_- - \lambda \bar{d}_-$$

Použijeme-li získané identity, můžeme psát

$$\begin{aligned}
(H_+ y - d)^T B_+ (H_+ y - d) &= (\omega H_+ \bar{y}_- - \lambda \bar{d}_-)^T B_+ (\omega H_+ \bar{y}_- - \lambda \bar{d}_-) \\
&= \omega^2 \bar{y}_-^T H_+ \bar{y}_- - 2\lambda\omega \bar{d}_-^T \bar{y}_- + \lambda^2 \bar{d}_-^T B_+ \bar{d}_- \\
&= (\omega - \lambda)^2 \bar{b}_- + \lambda^2 (\bar{d}_-^T \bar{y})^2 / \bar{b}
\end{aligned}$$

Jestliže $\omega = \bar{d}_-^T y / \bar{b}_-$, platí $\bar{d}_-^T \bar{y} = 0$, takže druhý člen v (729) odpadne. □

Z vyjádření (729) vyplývá, že podmínka $H_+ y = d$ může být splněna pouze tehdy, když $d^T \bar{y}_- = \bar{d}_-^T y$ (pak $\omega = \lambda$). Je však možné vynulovat výraz $y^T (H_+ y - d)$.

Věta 179. *Nechť jsou splněny předpoklady věty 174 a $\lambda = d^T \bar{y}_- / \bar{b}_-$. Pak pokud*

$$\omega = \frac{\sqrt{d^T \bar{y}_- \bar{d}_-^T y}}{\bar{b}_-}, \tag{730}$$

platí $y^T (H_+ y - d) = 0$ (výraz (730) je geometrický střed optimálních hodnot (724)).

Důkaz Jelikož $\lambda = d^T \bar{y}_- / \bar{b}_-$, platí $\bar{d}^T \bar{y}_- = 0$ a z důkazu věty 178 víme, že $\bar{y}_-^T H_+ \bar{y}_- = \bar{b}_-$. Použijeme-li tyto vztahy spolu s (723), dostaneme

$$\begin{aligned}
y^T (H_+ y - d) &= (\bar{y} + \omega \bar{y}_-)^T (\bar{d} + \omega H_+ \bar{y}_-) - (\bar{y} + \omega \bar{y}_-)^T \bar{d} - \lambda y^T \bar{d}_- \\
&= \bar{y}^T \bar{d} + \omega^2 \bar{b}_- - \bar{y}^T \bar{d} - \lambda y^T \bar{d}_- = \omega^2 \bar{b}_- - \frac{d^T \bar{y}_- \bar{d}_-^T y}{\bar{b}_-}.
\end{aligned}$$

Tento výraz je nulový, pokud $\omega^2 = y^T \bar{d}_- \bar{d}_-^T y / \bar{b}_-^2$, neboli pokud platí (730). □

Poznámka 287. Ve všech dokazovaných větách se předpokládá, že $\bar{b} > 0$. Pokud $\lambda = d^T\bar{y}_-/\bar{b}_-$, platí

$$\bar{b} = (d - \lambda\bar{d}_-^T)(y - \omega\bar{y}_-) = b - \frac{d^T\bar{y}_-\bar{d}_-^T y}{\bar{b}_-} \tag{731}$$

(takže $\bar{b}$ nezávisí na ω). Není však obecně zaručeno, že $\bar{b} > 0$. Z tohoto důvodu i z důvodu numerické efektivity je třeba transformaci odmítnout a položit $\bar{d} = d$, $\bar{y} = y$ (neboli $\lambda = 0$, $\omega = 0$), pokud $d^T\bar{y}_-\bar{d}_-^T y > (1 - \delta_1)b\bar{b}_-$, (neboli $\bar{b} < \delta_1 b$ podle (731)), kde $0 < \delta_1 < 1$, nebo pokud se čísla λ a ω příliš liší, konkrétně pokud platí $\lambda\omega < 0$ nebo $|\lambda - \omega| > \bar{b}_-/b$. Abychom dokázali globální konvergenci, odmítneme transformaci také tehdy, když $\|\bar{d}\| > \Delta\|d\|$ nebo $\|\bar{y}\| > \Delta\|y\|$, kde $\Delta > 1$. Ukazuje se, že je výhodnější použít hodnotu (730), pokud $d^T\bar{y}_-\bar{d}_-^T y > (1 - \delta_2)b\bar{b}_-$ (neboli $\bar{b} < \delta_2 b$ podle (731)). Hodnota (730) splňuje podmínku $\omega^2 \le b/\bar{b}_-$, neboť z $\bar{b} > 0$ a (731) plyne, že $\omega^2 = d^T\bar{y}_-\bar{d}_-^T y/\bar{b}_-^2 < b/\bar{b}_-$.

Ukážeme nyní, že modifikovaná vektorová metoda s proměnnou metrikou s omezenou pamětí je v jistém smyslu optimální, aplikujeme-li ji na ryze konvexní kvadratickou funkci s pozitivně definitní Hessovou maticí G.

Lemma 80. *Nechť G a H jsou symetrické pozitivně definitní matice takové, že $y = Gd$, $\bar{y}_- = G\bar{d}_-$ a $H\bar{y}_- = \bar{d}_-$. Pak jsou-li vektory d a $\bar{d}_-$ lineárně nezávislé a vybíráme-li parametry λ a ω podle (724), platí $\omega = \lambda$ a $\bar{b} > 0$. Nechť H_+ je matice určená podle (723). Pak hodnota $\|G^{1/2}H_+G^{1/2} - I\|_F$ nabývá svého minima, jsou-li parametry λ a $\omega = \lambda$ určeny podle vzorce (724).*

Důkaz (a) Podle předpokladu a vztahu (724) platí

$$\omega = \bar{d}_-^T y/\bar{b}_- = \bar{d}_-^T Gd/\bar{b}_- = \bar{y}_-^T d/\bar{b}_- = \lambda,$$

takže $G\bar{d} = Gd - \lambda G\bar{d}_- = y - \omega\bar{y}_- = \bar{y}$ a tedy $\bar{b} = \bar{d}^T\bar{y} = \bar{d}^T G\bar{d} > 0$ (matice G je pozitivně definitní).

(b) Označme $\bar{z} = G^{1/2}\bar{d} = G^{-1/2}\bar{y}$, $z = G^{1/2}d = G^{-1/2}y$, $\bar{z}_- = G^{1/2}\bar{d}_- = G^{-1/2}\bar{y}_-$ (takže $\bar{z} = z - \lambda\bar{z}_-$) a $R_+^{-1} = G^{1/2}H_+G^{1/2}$, $R^{-1} = G^{1/2}HG^{1/2}$. Pak podle (723) platí

$$R_+^{-1} = \bar{P}R^{-1}\bar{P} + \frac{\bar{z}\bar{z}^T}{\bar{z}^T\bar{z}} = I + \bar{P}(R^{-1} - I)\bar{P}, \qquad \bar{P} = I - \frac{\bar{z}\bar{z}^T}{\bar{z}^T\bar{z}}, \tag{732}$$

neboť $\bar{b} = \bar{d}^T\bar{y} = \bar{d}^T G\bar{d} = \bar{z}^T\bar{z}$ a $\bar{P}^2 = \bar{P}$. Je zřejmé, že číslo $\bar{b} = \|\bar{z}\|^2 = \|z\|^2 - 2\lambda z^T\bar{z}_- + \lambda^2\|\bar{z}_-\|^2$ je minimální, pokud $\lambda = z^T\bar{z}_-/\|\bar{z}_-\|^2 = d^T\bar{y}_-/\bar{b}_-$. Použijeme-li (732) a označíme-li $M = R^{-1} - I$, můžeme psát

$$\begin{aligned}
\|R_+^{-1} - I\|_F^2 &= \|\bar{P}(R^{-1} - I)\bar{P}\|_F^2 = \operatorname{Tr}(\bar{P}M\bar{P}\bar{P}M\bar{P}) = \operatorname{Tr}(\bar{P}M\bar{P}M) \\
&= \operatorname{Tr}\left(\left(M - \frac{\bar{z}\bar{z}^T M}{\bar{z}^T\bar{z}}\right)\left(M - \frac{\bar{z}\bar{z}^T M}{\bar{z}^T\bar{z}}\right)\right) \\
&= \|R^{-1} - I\|_F^2 - 2\frac{\bar{z}^T M^2\bar{z}}{\bar{z}^T\bar{z}} + \left(\frac{\bar{z}^T M\bar{z}}{\bar{z}^T\bar{z}}\right)^2
\end{aligned} \tag{733}$$

(využíváme toho, že matice $\bar{P}$, M jsou symetrické, $\bar{P}^2 = \bar{P}$, a že pro libovolné čtvercové matice A, B a vektory u, v platí $\|A\|_F^2 = \operatorname{Tr}(A^TA)$, $\operatorname{Tr}(A + B) = \operatorname{Tr}A + \operatorname{Tr}B$, $\operatorname{Tr}(AB) = \operatorname{Tr}(BA)$ a $\operatorname{Tr}(uv^T) = v^Tu$). Vztah $H\bar{y}_- = \bar{d}_-$ lze přepsat ve tvaru $G^{1/2}(H\bar{y}_- - \bar{d}_-) = R^{-1}\bar{z}_- - \bar{z}_- = M\bar{z}_- = 0$, takže $M\bar{z} = Mz$, což dává $\bar{z}^T M^2\bar{z} = z^T M^2 z$ a $\bar{z}^T M\bar{z} = z^T Mz$ (tato čísla tedy nezávisejí na parametru λ). Položíme-li $\xi = 1/\bar{b}$, můžeme výraz (733) přepsat ve tvaru $\|R_+^{-1} - I\|_F^2 = \|R^{-1} - I\|_F^2 - 2z^T M^2 z\xi + (z^T Mz)^2\xi^2$, takže

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\|R_+^{-1} - I\|_F^2 = 2(z^T Mz)^2 - 2z^T M^2 z\xi = 2\left(\frac{(\bar{z}^T M\bar{z})^2}{\bar{z}^T\bar{z}} - \bar{z}^T M^2\bar{z}\right) \le 0$$

(používáme Schwarzovu nerovnost aplikovanou na vektory $\bar{z}$ a $M\bar{z}$). Norma $\|(R_+)^{-1} - I\|_F$ je tedy minimální, je-li ξ maximalní, neboli je-li $\bar{b}$ minimální, což jak jsme dokázali nastane pro $\lambda = d^T\bar{y}_-/\bar{b}_-$. □

Věta 180. *Předpokládejme, že minimalizovaná funkce má tvar (687), kde G je symetrická pozitivně definitní matice. Nechť H je symetrická pozitivně definitní matice taková, že $H\bar{y}_- = \bar{d}_-$, a H_+ je matice určená podle (723). Pak jsou-li vektory d a $\bar{d}_-$ lineárně nezávislé a vybíráme-li parametry λ a ω podle (724), platí $H_+\bar{y} = \bar{d}$, $H_+y = d$, $H_+\bar{y}_- = \bar{d}_-$ a hodnota $\|G^{1/2}H_+G^{1/2} - I\|_F$ je minimální.*

Důkaz (a) Dokážeme indukcí, že pro $i \in N$ platí $\bar{y}_i = G\bar{d}_i$, takže matice G splňuje předpoklady lemmatu 80. Předpokládejme, že $\bar{y}_{i-1} = G\bar{d}_{i-1}$ (platí to pro $i = 2$, neboť $\bar{d}_1 = d_1$ a $\bar{y}_1 = y_1$). Pak lze psát $\bar{d}_{i-1}^T y_i = \bar{d}_{i-1}^T G d_i = \bar{y}_{i-1}^T d_i$, což spolu s (724) dává $\omega_i = \lambda_i$, takže $\bar{y}_i = y_i - \lambda\bar{y}_{i-1} = Gd_i - \lambda G\bar{d}_{i-1} = G\bar{d}_i$.

(b) Rovnosti $H_+\bar{y} = \bar{d}$, $H_+\bar{y}_- = \bar{d}_-$ a $H_+y = d$ plynou z (723), (728) a (729). Zbytek tvrzení plyne z lemmatu 80. □

Následující věta ukazuje, že modifikovaná metoda s proměnnou metrikou s omezenou pamětí, aplikovaná na kvadratickou funkci, v jistém smyslu generuje konjugované směrové vektory bez přesného výběru délky kroku.

Věta 181. *Aplikujeme-li modifikovanou metodu s proměnnou metrikou s omezenou pamětí, používající aktualizaci (723), na ryze konvexní kvadratickou funkci s pozitivně definitní Hessovou maticí G, přičemž $d_{k+1} = s_{k+1} = -H_{k+1}g_{k+1}$ (jednotková délka kroku) pro nějaký index $k \in N$, pak pro $1 \le i \le m$ platí*

$$H_{k+i}\,\bar{y}_k = \bar{d}_k, \tag{734}$$

$$\bar{d}_k^T G\bar{d}_{k+i} = 0, \tag{735}$$

$$\bar{d}_k^T g_{k+i+1} = 0. \tag{736}$$

Důkaz Důkaz provedeme indukcí. Pro $i = 1$ plyne (734) bezprostředně z věty 174. Jelikož pro kvadratickou funkci platí $G\bar{d}_k = \bar{y}_k$, můžeme psát

$$\bar{d}_k^T G\bar{d}_{k+1} = \bar{y}_k^T\bar{d}_{k+1} = \bar{y}_k^T\left(d_{k+1} - \frac{d_{k+1}^T\bar{y}_k}{\bar{y}_k^T\bar{d}_k}\bar{d}_k\right) = 0$$

a

$$\bar{d}_k^T g_{k+2} = \bar{d}_k^T(y_{k+1} + g_{k+1}) = \bar{y}_k^T(d_{k+1} - H_{k+1}g_{k+1}) = 0,$$

neboť $\bar{d}_k^T y_{k+1} = \bar{d}_k^T G d_{k+1} = \bar{y}_k^T d_{k+1}$, $\bar{d}_k = H_{k+i}\bar{y}_k$ a z $\alpha_{k+1} = 1$ plyne $H_{k+1}g_{k+1} = d_{k+1}$. Předpokládejme nyní, že (735)–(736) platí pro všechny indexy $1 \le j \le i-1$, kde $2 \le i \le m$.

(a) Podle (735) pro $1 \le j \le i-1$ platí $\bar{y}_k^T\bar{d}_{k+j} = \bar{d}_k^T\bar{y}_{k+j} = 0$, takže

$$\bar{V}_{k+j}^T\bar{y}_k = \left(I - \frac{\bar{y}_{k+j}^T\bar{d}_{k+j}}{\bar{d}_{k+j}^T\bar{y}_{k+j}}\right)\bar{y}_k = \bar{y}_k.$$

a

$$\bar{V}_{k+j}\bar{d}_k = \left(I - \frac{\bar{d}_{k+j}^T\bar{y}_{k+j}}{\bar{y}_{k+j}^T\bar{d}_{k+j}}\right)\bar{d}_k = \bar{d}_k.$$

Položíme-li $H_{k+i-m}^{k+i} = \gamma_{k+i-m}I$, můžeme podle (699) psát

$$H_{k+i}\bar{y}_k = \gamma_{k+i-m}\left(\prod_{j=i-m}^{i-1}\bar{V}_{k+j}\right)^T\left(\prod_{j=i-m}^{i-1}\bar{V}_{k+j}\right)\bar{y}_k + \sum_{l=i-m}^{i-1}\left(\prod_{j=l+1}^{i-1}\bar{V}_{k+j}\right)^T\frac{\bar{d}_{k+l}\bar{d}_{k+l}^T}{\bar{d}_{k+l}\bar{y}_{k+l}^T}\left(\prod_{j=l+1}^{i-1}\bar{V}_{k+j}\right)\bar{y}_k.$$

Jestliže $i - m \le 0$, neboli $i \le k + m$, je první člen tohoto vyjádření nulový, neboť pro $i - 1 \ge j \ge 1$ platí $\bar{V}_{k+j}^T\bar{y}_k = \bar{y}_k$ a pak $\bar{V}_k^T\bar{y}_k = 0$. Ze stejného důvodu jsou nulové všechny členy s $l < 0$. Člen s $l = 0$ je

roven $\bar{d}_k$, neboť pro $i-1 \geq j \geq 1$ platí $\bar{V}_{k+j}^T \bar{y}_k = \bar{y}_k$, pak $(\bar{d}_k \bar{d}_k^T / \bar{d}_k^T \bar{y}_k)\bar{y}_k = \bar{d}_k$ a pro $1 \leq j \leq i-1$ platí $\bar{V}_{k+j}\bar{d}_k = \bar{d}_k$. Členy s $l > 0$ jsou nulové, neboť podle (735) pro $1 \leq l \leq i-1$ platí $\bar{d}_{k+l}^T \bar{y}_k = 0$. Můžeme tedy psát $H_{k+i}\bar{y}_k = \bar{d}_k$.

(b) Použijeme-li (736) a (a), dostaneme

$$\bar{y}_k^T d_{k+i} = -\alpha_{k+i} \bar{y}_k^T H_{k+i} g_{k+i} = \bar{d}_k^T g_{k+i} = 0,$$

což spolu s (735) dává

$$\bar{d}_k^T G \bar{d}_{k+i} = \bar{y}_k^T \bar{d}_{k+i} = \bar{y}_k^T d_{k+i} - \lambda_{k+i} \bar{y}_k^T \bar{d}_{k+i-1} = 0.$$

(c) Použijeme-li (736) a (b), můžeme psát

$$\bar{d}_k^T g_{k+i+1} = \bar{d}_k^T y_{k+i} + \bar{d}_k^T g_{k+i} = \bar{d}_k^T (\bar{y}_{k+i} + \lambda_{k+i} \bar{y}_{k+i-1}) = 0.$$

□

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 18. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\delta_1 = 10^{-6}$, $\delta_2 = 10^{-2}$, $\Delta = 1000$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \leq \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$ a určíme směrový vektor s_i: Pokud $m = 0$, položíme $s_i := -g_i$, v opačném případě vypočteme s_i pomocí rekurentích vztahů (700)–(701), kde místo veličin ρ_j, b_j, d_j, y_j používáme veličiny s pruhem (přitom $\bar{\rho}_j = 1$).

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$, $y_i := g_{i+1} - g_i$ a $b_i := d_i^T y_i$. Jestliže $m = 0$, položíme $\bar{d}_i := d_i$, $\bar{y}_i := y_i$, $\bar{b}_i := b_i$ a přejdeme na krok 6.

Krok 4 Položíme $\lambda_i := d_i^T \bar{y}_{i-1} / \bar{b}_{i-1}$, $\omega_i := y_i^T \bar{d}_{i-1} / \bar{b}_{i-1}$ a $\bar{b}_i := b_i - \lambda_i \omega_i \bar{b}_{i-1}$. Pokud $\lambda_i \omega_i < 0$ nebo $|\lambda_i - \omega_i| > \bar{b}_{i-1}/b_i$ nebo $\bar{b}_i < \delta_1 b_i$ nebo $\|\bar{d}_{i-1}\| > \Delta \|d_{i-1}\|$ nebo $\|\bar{y}_{i-1}\| > \Delta \|y_{i-1}\|$, položíme $\bar{d}_i := d_i$, $\bar{y}_i := y_i$, $\bar{b}_i := b_i$ a přejdeme na krok 5. Pokud $\bar{b}_i < \delta_2 b_i$, položíme $\omega_i := \sqrt{\lambda_i \omega_i}$.

Krok 5 Položíme $\bar{d}_i := d_i - \lambda_i \bar{d}_{i-1}$, $\bar{y}_i := y_i - \omega_i \bar{y}_{i-1}$ a $\bar{b}_i := \bar{d}_i^T \bar{y}_i$.

Krok 6 Uložíme vektory $\bar{d}_i$, $\bar{y}_i$ a číslo $\bar{b}_i$ do pracovního pole. Pokud $m = \bar{m}$, odstraníme vektory $\bar{d}_{i-m}$, $\bar{y}_{i-m}$ a číslo $\bar{b}_{i-m}$ z pracovního pole. Zvětšíme i o 1 a přejdeme na krok 2.

Nyní dokážeme globální konvergenci algoritmu 18. Tak jako v oddílu 9.1 budeme předpokládat, že funkce $F : R^n \to R$ vyhovuje předpokladům F1, F4, F5, takže platí (695).

Lemma 81. *Uvažujme aktualizaci tvaru*

$$H_+ = H + \left(1 + \frac{\bar{a}}{\bar{b}}\right) \frac{1}{\bar{b}} \bar{d}\bar{d}^T - \frac{1}{\bar{b}} \left(H\bar{y}\bar{d}^T + \bar{d}(H\bar{y})^T\right), \qquad B_+ = B + \frac{1}{\bar{b}} \bar{y}\bar{y}^T - \frac{1}{\bar{c}} B\bar{d}(B\bar{d})^T. \tag{737}$$

kde $B = H^{-1}$ je pozitivně definitní matice, $\bar{d} \in R^n$, $\bar{y} \in R^n$, $\bar{a} = \bar{y}^T H \bar{y} > 0$, $\bar{b} = \bar{y}^T \bar{d} > 0$, $\bar{c} = \bar{d}^T B \bar{d} > 0$. Pak matice $B_+ = H_+^{-1}$ je pozitivně definitntí a platí

$$\operatorname{Tr} B_+ \leq \operatorname{Tr} B + \frac{\bar{y}^T \bar{y}}{\bar{y}^T \bar{d}}, \qquad \det B_+ = \det B \ \frac{\bar{y}^T \bar{d}}{\bar{d}^T B \bar{d}}. \tag{738}$$

Důkaz Pozitivntí definitnost matice $H_+ = B_+^{-1}$ je vlastností aktualizace BFGS (poznámka 117). Vztahy (738) plynou bezprostředně z (412) a (414), kam dosazujeme veličiny s pruhem, přižemž $\bar{\gamma} = 1$, $\bar{\rho} = 1$ a $\bar{\beta} = 0$. □

Lemma 82. *Nechť H_i^i je matice určená rekurentně podle (722) a nechť pro $i \in N$ platí*

$$\operatorname{Tr} B_{i-m}^i \le C_0, \qquad \frac{\|\bar{y}_i\|^2}{\bar{b}_i} \le C_1, \tag{739}$$

$$\det B_{i-m}^i \ge K_0, \qquad \frac{\bar{b}_i}{\|\bar{d}_i\|^2} \ge K_1, \tag{740}$$

kde $0 < K_0 < 1 < C_0$, $0 < K_1 < 1 < C_1$. Pak existují čísla $0 < K < 1 < C$ taková, že $\operatorname{Tr} B_i^i \le C$ a $\det B_i^i \ge K$.

Důkaz Jelikož aktualizace (722) odpovídají metodě BFGS, můžeme použít vzorce (737), takže podle (738) a (739) pro $n - m \le j \le i$ platí

$$\operatorname{Tr} B_j^i \le \operatorname{Tr} B_{i-m}^i + \sum_{k=i-m}^{j-1} \frac{\bar{y}_k^T \bar{y}_k}{\bar{y}_k^T \bar{d}_k} \le \; C_0 + \bar{m} C_1 \triangleq C, \tag{741}$$

což pro $j = i$ dává $\operatorname{Tr} B_i^i \le C$. Použijeme-li (741), můžeme pro $n - m \le k \le i$ psát

$$\bar{d}_k^T B_k^i \bar{d}_k \le \|B_k^i\| \|\bar{d}_k\|^2 \le \operatorname{Tr} B_k^i \; \|\bar{d}_k\|^2 \le C \|\bar{d}_k\|^2,$$

(neboť pro pozitivně definitní matici B_k^i platí $\|B_k^i\| \le \operatorname{Tr} B_k^i$), což spolu s (738) a (740) pro $n - m < j \le i$ dává

$$\det B_j^i = \det B_{i-m}^i \; \prod_{k=i-m}^{j-1} \frac{\bar{y}_k^T \bar{d}_k}{\bar{d}_k^T B_k \bar{d}_k} \ge \bar{m} K_0 K_1 / C \triangleq K,$$

takže pro $j = i$ platí $\det B_i^i \ge K$. □

Věta 182. *Uvažujme metodu s proměnnou metrikou s omezenou pamětí, realizovanou algoritmem 18 s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Nechť funkce F splňuje předpoklady F1, F4, F5. Pak směrové vektory $s_i = -H_i^i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$.*

Důkaz (a) Nechť $i \in N$ a $\lambda_i = d_i^T \bar{y}_{i-1} / \bar{b}_{i-1} \ne 0$. Jelikož $\bar{b}_{i-1} \ge \delta_1 b_{i-1}$ a $\|\bar{d}_{i-1}\| \le \Delta \|d_{i-1}\|$, (krok 4 algoritmu 18), můžeme podle důsledku 11 s použitím (695) psát

$$\begin{aligned} \|\bar{d}_i\| &= \|d_i - \lambda_i \bar{d}_{i-1}\| = \left\| \left(I - \frac{\bar{d}_{i-1} \bar{y}_{i-1}^T}{\bar{b}_{i-1}} \right) \right\| \|d_i\| \le \frac{\|\bar{d}_{i-1}\| \|\bar{y}_{i-1}\|}{\bar{b}_{i-1}} \|d_i\| \\ &\le \frac{\Delta^2}{\delta_1} \frac{\|d_{i-1}\| \|y_{i-1}\|}{b_{i-1}} \|d_i\| \le \frac{\Delta^2}{\delta_1} \sqrt{\frac{\overline{G}}{\underline{G}}} \|d_i\| \triangleq \overline{C} \|d_i\|. \end{aligned}$$

Pokud $\lambda_i = 0$, platí $\bar{d}_i = d_i$, takže opět $\|\bar{d}_i\| \le \overline{C} \|d_i\|$ (neboť $\overline{C} > 1$). S použitím (695) tedy dostaneme

$$\frac{\|\bar{d}_i\|^2}{\bar{b}_i} \le \frac{\overline{C}^2}{\delta_1} \frac{\|d_i\|^2}{b_i} \le \frac{\overline{C}^2}{\delta_1 \underline{G}} \triangleq \frac{1}{K_1}.$$

Jestliže $\omega_i = d_i^T \bar{y}_{i-1} / \bar{b}_{i-1} \ne 0$ nebo $\omega_i = 0$, můžeme postupovat stejným způsobem jako při odhadu normy $\|d_i\|$. Odvodíme tak nerovnost $\|\bar{y}_i\| \le \overline{C} \|y_i\|$, která s použitím (695) dává

$$\frac{\|\bar{y}_i\|^2}{\bar{b}_i} \le \frac{\overline{C}^2}{\delta_1} \frac{\|y_i\|^2}{b_i} \le \frac{\overline{C}^2 \overline{G}}{\delta_1} \le \frac{\overline{G}}{\delta_1^2} \max(\overline{C}^2, 4\Delta^2) \triangleq C_1.$$

Nechť $\omega_i = \sqrt{d_i^T \bar{y}_{i-1} \bar{d}_{i-1}^T y_i} \,/\, \bar{b}_{i-1}$. Jelikož podle lemmatu 3 pro libovolné dva vektory $u \in R^n$, $v \in R^n$ platí $\|u + v\|^2 \le 2(\|u\|^2 + \|v\|^2)$, jelikož $\bar{b}_i \ge \delta_1 b_i$, $\bar{b}_{i-1} \ge \delta_1 b_{i-1}$, $\|\bar{y}_i\| \le \Delta \|y_i\|$, $\|\bar{y}_{i-1}\| \le \Delta \|y_{i-1}\|$ (krok 4 algoritmu 18) a jelikož podle poznámky 287 platí $\omega_i^2 \le b_i / \bar{b}_{i-1}$, můžeme psát

$$\frac{\|\bar{y}_i\|^2}{\bar{b}_i} = \frac{\|y_i - \omega_i \bar{y}_{i-1}\|^2}{\bar{b}_i} \le \frac{2}{\bar{b}_i} (\|y_i\|^2 + \omega_i^2 \|\bar{y}_{i-1}\|^2) \le \frac{2\Delta^2}{\delta_1^2} \left(\frac{\|y_i\|^2}{b_i} + \frac{\|y_{i-1}\|^2}{b_{i-1}} \right) \le \frac{4\Delta^2 \overline{G}}{\delta_1^2} \le C_1$$

(neboť $0 < \delta_1 < 1 < \Delta$).

(b) Podle (a) a lemmatu 75 jsou splněny předpoklady lemmatu 82, takže existují čísla $0 < K < 1 < C$ taková, že $\mathrm{Tr}\, B_i^i \le C$ a $\det B_i^i \ge K$. Jsou tedy splněny předpoklady lemmatu 76, takže vektory $s_i = -H_i^i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$. □

Poznámka 288. V důkazu globální konvergence potřebujeme, aby pro $i \in N$ platilo $\|\bar{d}_i\| \le \Delta\|d_i\|$, $\|\bar{y}_i| \le \Delta\|y_i\|$. Tento test vyžaduje dva skalární součiny ($2n$ aritmetických operací) navíc. Proto ho raději vynecháváme a globální konvergenci zajišťujeme podle poznámky 32 (účinnost metody se tím nezhorší a výpočet se urychlí).

9.3 Maticové metody s proměnnou metrikou s omezenou pamětí

Strangovy rekurence (700)–(701) jsou nejstarší a nejjednodušší realizací metody BFGS s omezenou pamětí. Pro některé aplikace jsou výhodnější maticové reprezentace, studované v práci [18], které nyní odvodíme. Abychom se při popisu těchto realizací vyhnuli dvojímu indexování, budeme bez újmy na obecnosti předpokládat, že $i \le \bar{m}$. Pak matice H_j^i, $1 \le j \le i$, nezávisejí na horním indexu, který můžeme vynechat. Případ, kdy $i > \bar{m}$, probereme později.

Lemma 83. *Nechť $N = -M^{-1}$, kde M je matice vystupující ve větě 78 s $\gamma = 1$ a $\rho = 1$. Pak platí*

$$N = \begin{bmatrix} \dfrac{\eta ab}{\eta a + (1-\eta)b} - b, & \dfrac{\eta ab}{\eta a + (1-\eta)b} \\ \dfrac{\eta ab}{\eta a + (1-\eta)b}, & \dfrac{\eta ab}{\eta a + (1-\eta)b} + a \end{bmatrix}. \tag{742}$$

Důkaz Z vyjádření matice M (věta 78) plyne

$$N = -M^{-1} = -\frac{1}{\det M} \begin{bmatrix} \dfrac{\eta - 1}{a}, & \dfrac{\eta}{b} \\ \dfrac{\eta}{b}, & \dfrac{1}{b}\left(\eta\dfrac{a}{b} + 1\right) \end{bmatrix}.$$

Dosadíme-li za $-\det M$ vztah μ definovaný v poznámce 109 (s $\gamma = 1$ a $\rho = 1$), dostaneme po úpravě tvrzení lemmatu. □

Poznámka 289. Pro metodu DFP je $\eta = 0$, takže

$$N = \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}. \tag{743}$$

Pro metodu BFGS je $\eta = 1$, takže

$$N = \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix}. \tag{744}$$

Lemma 84. *Nechť B a $\beta - b^T B^{-1} b$ jsou čtvercové regulární matice. Pak platí*

$$[A, a] \begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} [A, a]^T = AB^{-1}A^T + (a - AB^{-1}b)(\beta - b^T B^{-1} b)^{-1}(a - AB^{-1}b)^T. \tag{745}$$

Důkaz Vynásobením se snadno přesvědčíme, že platí

$$\left[\begin{array}{ll} B, & b \\ b^T, & \beta \end{array}\right]^{-1} = \left[\begin{array}{ll} B^{-1} + B^{-1}b(\beta - b^TB^{-1}b)^{-1}b^TB^{-1}, & -B^{-1}b(\beta - b^TB^{-1}b)^{-1} \\ -(\beta - b^TB^{-1}b)^{-1}b^TB^{-1}, & (\beta - b^TB^{-1}b)^{-1} \end{array}\right]. \tag{746}$$

Zbytek tvrzení snadno ověříme dosazením tohoto vyjádření do výchozího vzorce a následným roznásobením. □

V dalším výkladu budeme předpokládat, že H_1 je symetrická pozitivně definitní matice a že pro libovolný index $1 \leq j \leq i$, kde $1 \leq i \leq \bar{m}$, platí

$$H_{j+1} = H_j - [d_j,\ H_jy_j]\, N_j^{-1}\, [d_j,\ H_jy_j]^T, \tag{747}$$

kde N_j je matice řádu 2 specifikující konkrétní metodu s proměnnou metrikou. Budeme se snažit nalézt vyjádření

$$H_{i+1} = H_1 - [D_i,\ H_1Y_i]\, \bar{N}_i^{-1}\, [D_i,\ H_1Y_i]^T, \tag{748}$$

kde $D_i = [d_1, \ldots, d_i]$, $Y_i = [y_1, \ldots, y_i]$ a kde $\bar{N}_i$ je explicitně definovaná symetrická matice řádu $2i$. Budeme přitom používat označení R_i pro horní trojúhelníkovou matici řádu i takovou, že $(R_i)_{kl} = d_k^Ty_l$, $k \leq l$, a $(R_i)_{kl} = 0$, $k > l$, a C_i pro diagonální matici řádu i takovou, že $(C_i)_{kk} = d_k^Ty_k$. Abychom zjednodušili zápis budeme v důkazech často indexy $i-1$ a i vynechávat a index $i+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $H = H_i$, $N = N_i$, $\bar{N} = \bar{N}_{i-1}$, $D = D_{i-1}$, $Y = Y_{i-1}$, $R = R_{i-1}$, $C = C_{i-1}$, takže $D_+ = D_i = [D, d]$, $Y_+ = Y_i = [Y, y]$ a

$$R_+ = R_i = \left[\begin{array}{ll} R, & D^Ty \\ 0, & d^Ty \end{array}\right], \qquad R_+ - C_+ = R_i - C_i = \left[\begin{array}{ll} R - C, & D^Ty \\ 0, & 0 \end{array}\right].$$

Důkazy budeme provádět matematickou indukcí. Budeme předpokládat, že existuje explicitně definovaná matice $\bar{N}$ taková, že

$$H = H_1 - [D,\ H_1Y]\, \bar{N}^{-1}\, [D,\ H_1Y]^T. \tag{749}$$

Pro stejným způsobem konstruovanou matici $\bar{N}_+$ ukážeme, že matice

$$H_+ = H_1 - [D_+,\ H_1Y_+]\, \bar{N}_+^{-1}\, [D_+,\ H_1Y_+]^T \tag{750}$$

je totožná s maticí

$$H_+ = H - [d,\ Hy]\, N^{-1}\, [d,\ Hy]^T. \tag{751}$$

Tento způsob dokazování není příliš konstruktivní, nicméně v případě metod DFP, BFGS a R1 se dá struktura matice $\bar{N}$ odhadnout ze struktury matice N.

Poznámka 290. Dosadíme-li (749) do (751), dostaneme

$$\begin{aligned} H_+ &= H - \Big[d,\ H_1y - [D,\ H_1Y]\, \bar{N}^{-1}\, [D,\ H_1Y]^T y\Big] \cdot \\ &\quad N^{-1} \Big[d,\ H_1y - [D,\ H_1Y]\, \bar{N}^{-1}\, [D,\ H_1Y]^T y\Big]^T \\ &= H - \left([d,\ H_1y] - [D,\ H_1Y]\, \bar{N}^{-1} \left[\begin{array}{ll} 0, & D^Ty \\ 0, & Y^TH_1y \end{array}\right]\right) \cdot \\ &\quad N^{-1} \left([d,\ H_1y] - [D,\ H_1Y]\, \bar{N}^{-1} \left[\begin{array}{ll} 0, & D^Ty \\ 0, & Y^TH_1y \end{array}\right]\right)^T. \end{aligned} \tag{752}$$

Abychom dokázali, že postupná aplikace vzorců (749), (751) je ekvivalentní vzorci (750), budeme se snažit vyjádřit matici (750) ve tvaru (752).

Následující věty uvádějí maticové reprezentace nejznámějších metod s proměnnou metrikou s omezenou pamětí.

Věta 183. *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le j \le i$, kde $1 \le i \le \bar{m}$, platí (747), kde matice N_j je určená vztahem (743) (metoda DFP). Pak lze psát*

$$H_{i+1} = H_1 - [D_i,\ H_1Y_i] \begin{bmatrix} -C_i, & R_i - C_i \\ (R_i - C_i)^T, & Y_i^T H_1 Y_i \end{bmatrix}^{-1} [D_i,\ H_1Y_i]^T . \tag{753}$$

Důkaz Pro $i = 1$ je vztah (753) ekvivalentní se (287) (s (751) kde matice N je určena podle (743)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (753) platí pro všechny indexy menší než k, kde $0 < k \le \bar{m}$. Pro index $i = k$ můžeme vzorec (753), který má tvar (750), zapsat (po permutaci) ve tvaru

$$H_+ = H_1 - [D,\ H_1Y,\ d,\ H_1y] \begin{bmatrix} -C, & R - C, & 0, & D^Ty \\ (R-C)^T, & Y^TH_1Y, & 0, & Y^TH_1y \\ 0, & 0, & -d^Ty, & 0 \\ y^TD, & y^TH_1Y, & 0, & y^TH_1y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^TH_1 \\ d^T \\ y^TH_1 \end{bmatrix} .$$

Použijeme-li lemma 84 a označíme-li

$$\bar{N} = \begin{bmatrix} -C, & R - C \\ (R-C)^T, & Y^TH_1Y \end{bmatrix} ,$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D,\ H_1Y]\,\bar{N}^{-1}\,[D,\ H_1Y]^T - \\ &\quad \left([d,\ H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \left(\begin{bmatrix} -d^Ty, & 0 \\ 0, & y^TH_1y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^TD, & y^TH_1Y \end{bmatrix} \bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^{-1} \cdot \\ &\quad \left([d,\ H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T \\ &= H - \left([d, H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \begin{bmatrix} -d^Ty, & 0 \\ 0, & y^THy \end{bmatrix}^{-1} \left([d, H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T , \end{aligned}$$

což je právě vztah (752) s maticí N určenou podle (743). □

Věta 184. *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le j \le i$, kde $1 \le i \le \bar{m}$, platí (747), kde matice N_j je určená vztahem (744) (metoda BFGS). Pak lze psát*

$$H_{i+1} = H_1 - [D_i,\ H_1Y_i] \begin{bmatrix} 0, & R_i \\ R_i^T, & C_i + Y_i^TH_1Y_i \end{bmatrix}^{-1} [D_i,\ H_1Y_i]^T . \tag{754}$$

Důkaz Pro $i = 1$ je vztah (754) ekvivalentní se (288) (s (751) kde matice N je určena podle (744)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (754) platí pro všechny indexy menší než k, kde $0 < k \le \bar{m}$. Pro index $i = k$ můžeme vzorec (754), který má tvar (750), zapsat (po permutaci) ve tvaru

$$H_+ = H_1 - [D,\ H_1Y,\ d,\ H_1y] \begin{bmatrix} 0, & R, & 0, & D^Ty \\ R^T, & C + Y^TH_1Y, & 0, & Y^TH_1y \\ 0, & 0, & 0, & d^Ty \\ y^TD, & y^TH_1Y, & d^Ty, & d^Ty + y^TH_1y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^TH_1 \\ d^T \\ y^TH_1 \end{bmatrix} .$$

Použijeme-li lemma 84 a označíme-li

$$\bar{N} = \begin{bmatrix} 0, & R \\ R^T, & C + Y^T H_1 Y \end{bmatrix},$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D,\ H_1 Y]\, \bar{N}^{-1}\, [D,\ H_1 Y]^T - \\ &\quad \left([d,\ H_1 y] - [D, H_1 Y]\, \bar{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ &\quad \left(\begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \bar{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^{-1} \cdot \\ &\quad \left([d,\ H_1 y] - [D, H_1 Y]\, \bar{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^T \\ &= H - \left([d, H_1 y] - [D,\ H_1 Y]\, \bar{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ &\quad \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix}^{-1} \left([d, H_1 y] - [D,\ H_1 Y]\, \bar{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^T, \end{aligned}$$

což je právě vztah (752) s maticí N určenou podle (744). □

Věta 185. *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le j \le i$, kde $1 \le i \le \bar{m}$, platí*

$$H_{j+1} = H_j + (d_j - H_j y_j)(d_j^T y_j - y_j^T H_j y_j)^{-1}(d_j - H_j y_j)^T \tag{755}$$

(metoda hodnosti 1). Pak lze psát

$$H_{i+1} = H_1 + (D_i - H_1 Y_i)(R_i + R_i^T - C_i - Y_i^T H_1 Y_i)^{-1}(D_i - H_1 Y_i)^T. \tag{756}$$

Důkaz Pro $i = j = 1$ je vztah (756) ekvivalentní s (755). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (756) platí pro všechny indexy menší než k, kde $0 < k \le \bar{m}$. Pro index $i = k$ můžeme (756) zapsat ve tvaru

$$H_+ = H_1 + [D - H_1 Y,\ d - H_1 y] \begin{bmatrix} R + R^T - C - Y^T H_1 Y, & D^T y - Y^T H_1 y \\ y^T D - y^T H_1 Y, & d^T y - y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T - Y^T H_1 \\ d^T - y^T H_1 \end{bmatrix}.$$

Použijeme-li lemma 84 a označíme-li

$$\bar{N} = R + R^T - C - Y^T H_1 Y,$$

dostaneme

$$\begin{aligned} H_+ &= H_1 + (D - H_1 Y)\, \bar{N}^{-1}\, (D - H_1 Y)^T + \\ &\quad \left((D - H_1 Y)\, \bar{N}^{-1}\, (D - H_1 Y)^T\, y - (d - H_1 y) \right) \cdot \\ &\quad \left(d^T y - y^T H_1 y - y^T\, (D - H_1 Y)\, \bar{N}^{-1}\, (D - H_1 Y)^T\, y \right)^{-1} \cdot \\ &\quad \left((D - H_1 Y)\, \bar{N}^{-1}\, (D - H_1 Y)^T\, y - (d - H_1 y) \right)^T \\ &= H + (d - Hy) \left(d^T y - y^T H y \right)^{-1} (d - Hy)^T, \end{aligned}$$

což je právě vztah (755). □

Poznámka 291. Podobná maticová vyjádření můžeme odvodit pro matici $B = H^{-1}$. Lze k tomu použít dualitu (poznámka 119). Jelikož přitom dojde k výměně $D_i \to Y_i$, $Y_i \to D_i$, je třeba horní polovinu matice $D_i^T Y_i$ nahradit horní polovinou matice $Y_i^T D_i$, neboli transponovanou dolní polovinou matice $D_i^T Y_i$. Proto místo horní trojúhelníkové matice R_i použijeme dolní trojúhelníkovou matici L_i takovou, že $(L_i)_{kl} = 0$, $k < l$, a $(L_i)_{kl} = d_k^T y_l$, $k \geq l$. Pro metodu DFP dostaneme

$$B_{i+1} = B_1 - [Y_i, B_1 D_i] \begin{bmatrix} 0, & L_i^T \\ L_i, & C_i + D_i^T B_1 D_i \end{bmatrix}^{-1} [Y_i, B_1 D_i]^T . \tag{757}$$

Pro metodu BFGS dostaneme

$$B_{i+1} = B_1 - [Y_i, B_1 D_i] \begin{bmatrix} -C_i, & (L_i - C_i)^T \\ L_i - C_i, & D_i^T B_1 D_i \end{bmatrix}^{-1} [Y_i, B_1 D_i]^T . \tag{758}$$

Pro metodu hodnosti 1 dostaneme

$$B_{i+1} = B_1 + (Y_i - B_1 D_i) \left(L_i + L_i^T - C_i - D_i^T B_1 D_i \right)^{-1} (Y_i - B_1 D_i)^T . \tag{759}$$

Nyní ukážeme, jak lze popsaná maticová vyjádření upravit pro použití v metodách s proměnnou metrikou s omezenou pamětí. Omezíme se přitom na metodu BFGS, která je z popsaných metod nejefektivnější. Matici (754) můžeme po dosazení $H_1 = \gamma_i I$, kde $\gamma_i = d_i^T y_i / y_i^T y_i$, zapsat ve tvaru

$$H_{i+1} = \gamma_i I + [D_i, \gamma_i Y_i] \begin{bmatrix} (R_i^{-1})^T (C_i + \gamma_i Y_i^T Y_i) R_i^{-1}, & -(R_i^{-1})^T \\ -R_i^{-1}, & 0 \end{bmatrix} [D_i, \gamma_i Y_i]^T . \tag{760}$$

Lze se o tom přesvědčit vynásobením použité matice maticí $\bar{N}_i$ z (754) (kde $H_1 = \gamma_i I$). Tento vzorec je velmi výhodný, neboť se v něm invertuje pouze horní trojúhelníková matice řádu m (takže při výpočtu směrového vetoru řešíme pouze soustavy rovnic s horní trojúhelníkovou maticí R_i řádu m). Vzorec (760) můžeme zapsat ve tvaru

$$H_{i+1} = (R_i^{-1} D_i^T)^T C_i R_i^{-1} D_i^T + \gamma_i (I - Y_i R_i^{-1} D_i^T)^T (I - Y_i R_i^{-1} D_i^T). \tag{761}$$

jak se snadno přesvědčíne roznásobením a porovnáním obou vztahů. Směrový vektor lze tedy vypočítat podle vzorce

$$\begin{aligned} s_{i+1} = -H_{i+1} g_{i+1} \;=\; & -\gamma_i \, g_{i+1} \;-\; D_i (R_i^T)^{-1} \left[(C_i + \gamma_i Y_i^T Y_i) R_i^{-1} D_i^T g_{i+1} - \gamma_i Y_i^T g_{i+1} \right] \\ & + Y_i \left[\gamma_i R_i^{-1} D_i^T g_{i+1} \right], \end{aligned} \tag{762}$$

Nejprve vynásobíme gradient g_{i+1} zleva maticemi D_i^T a Y_i^T ($2mn$ operací), získané vektory upravujeme pomocí matic dimenze m (hranaté závorky) a výsledné vektory vynásobíme zleva maticemi D_i a Y_i ($2mn$ operací).

Matici (758) můžeme po dosazení $B_1 = (1/\gamma_i) I$ také upravit. Využijeme toho, že platí

$$\begin{bmatrix} -C_i, & (L_i - C_i)^T \\ L_i - C_i, & \dfrac{1}{\gamma_i} D_i^T D_i \end{bmatrix} = \begin{bmatrix} C_i^{1/2}, & 0 \\ -(L_i - C_i) C_i^{-1/2}, & \overline{L}_i \end{bmatrix} \begin{bmatrix} -C_i^{1/2}, & 0 \\ (L_i - C_i) C_i^{-1/2}, & \overline{L}_i \end{bmatrix}^T ,$$

kde

$$\overline{L}_i \overline{L}_i^T = (L_i - C_i)^T C_i^{-1} (L_i - C_i) + \frac{1}{\gamma_i} D_i^T D_i \tag{763}$$

(lze se o tom přesvědčit prostým vynásobením). Použijeme-li vzorec pro inverzi dolní blokově trojúhelníkové matice

$$\begin{bmatrix} A & 0 \\ B & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}, & 0 \\ -C^{-1} B A^{-1}, & C^{-1} \end{bmatrix},$$

jehož správnost lze opět ověřit vynásobením, dostaneme

$$\begin{bmatrix} -C_i, & (L_i - C_i)^T \\ L_i - C_i, & \frac{1}{\gamma_i} D_i^T D_i \end{bmatrix}^{-1} = \begin{bmatrix} -C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix}^T \begin{bmatrix} C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix},$$

takže

$$B_{i+1} = \frac{1}{\gamma_i} I - \left[Y_i, \frac{1}{\gamma_i} D_i \right] \begin{bmatrix} -C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix}^T \begin{bmatrix} C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix} \left[Y_i, \frac{1}{\gamma_i} D_i \right]^T \quad (764)$$

(opět se řeší pouze soustavy rovnic s dolní trojúhelníkovou maticí $\overline{L}_i$ řádu m). Matice $\overline{L}_i$ se získává Choleského rozkladem matice (763). Poznamenejme, že metoda BFGS založená na maticovém vyjádření (760) není numericky efektivnější než metoda BFGS používající Strangovy rekurence. Vzorce (758) a (764) jsou však velmi užitečné, neboť je lze použít tam, kde je nutné pracovat s maticí B.

Ukážeme nyní, jak se konstruuje matice H_{i+1} v obecném případě, kdy může platit $i > \bar{m}$. Budeme předpokládat, že $H_1 = \gamma_i I$ a používat vzorec (760), ve kterém $D_i = [d_{i-m+1}, \ldots, d_i]$, $Y_i = [y_{i-m+1}, \ldots, y_i]$, C_i obsahuje diagonálu matice $D_i^T Y_i$ a R_i obsahuje horní polovinu matice $D_i^T Y_i$. Matice D_i, Y_i vzniknou z matic D_{i-1}, Y_{i-1} přidáním nových sloupců d_i, y_i, a pokud $i > \bar{m}$, ubráním starých sloupců d_{i-m}, y_{i-m}. Podobně jednoduše získáme matice $D_i^T Y_i$, $Y_i^T Y_i$ z matic $D_{i-1}^T Y_{i-1}$, $Y_{i-1}^T Y_{i-1}$ a tudíž i matice C_i, R_i z matic C_{i-1}, R_{i-1}. Tím máme k dispozici všechny matice potřebné k výpočtu matice H_{i+1}.

Poznámka 292. Metoda používající vzorec (760) vyžaduje zhruba $6mn$ operací násobení a sčítání v každém iteračním kroku ($2mn$ na výpočet nových sloupců matic $D_i^T Y_i$, $Y_i^T Y_i$ a $4mn$ na výpočet směrového vektoru s_{i+1} podle vzorce (760)). Zhruba $2(m-1)n$ operací však lze ušetřit, pokud při určování matice H_{i+1} počítáme a ukládáme vektory $D_i^T g_{i+1}$, $Y_i^T g_{i+1}$ místo vektorů $D_i^T y_i$, $Y_i^T y_i$. Prvních $m-1$ prvků vektorů $D_i^T y_i$, $Y_i^T y_i$ pak určujeme z již spočtených hodnot podle vzorců $d_j^T y_i = d_j^T g_{i+1} - d_j^T g_i$, $y_j^T y_i = y_j^T g_{i+1} - y_j^T g_i$, $i - m + 1 \leq j \leq i - 1$, takže je nutné spočítat pouze dva skalární součiny $d_i^T y_i$, $y_i^T y_i$. Vektory $D_i^T g_{i+1}$, $Y_i^T g_{i+1}$ lze pak použít k výpočtu směrového vektoru $s_{i+1} = -H^{i+1} g_{i+1}$ (viz (760)), takže odpadne $2mn$ operací násobení a sčítání.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 19. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \leq \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$ a určíme směrový vektor s_i. Pokud $m = 0$, položíme $s_i := -g_i$, v opačném případě položíme $s_i = -H_i g_i$, kde H_i je matice určená vztahem (760) (kde vystupuje i místo $i+1$ a $i-1$ místo i).

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$ a $y_i := g_{i+1} - g_i$.

Krok 4 Sestrojíme matice D_i, Y_i a R_i, C_i z matic D_{i-1}, Y_{i-1} a R_{i-1}, C_{i-1} (staré sloupce ubíráme pouze tehdy, pokud $i > \bar{m}$). Zvětšíme i o 1 a přejdeme na krok 2.

Explicitní vzorce (753), (754), (756) nelze zobecnit. Není znám obecný explicitní vzorec použitelný pro libovolnou metodu z Broydenovy třídy. Je však možné použít rekurzivní postup popsaný v práci [115]. Vyjdeme z toho, že vzorec (760) můžeme po permutaci zapsat ve tvaru

$$H_{i+1} = H_1 + \bar{U}_i \bar{M}_i \bar{U}_i^T$$

kde $\bar{U}_i = [d_1, H_1 y_1, \ldots d_i, H_1 y_i]$ a $\bar{M}_i$ je matice řádu $2i$ (předpokládáme opět, že $i \leq \bar{m}$). Výhodou tohoto vzorce je, že matice $\bar{M}_i$ je zadána v explicitním tvaru, který lze použít i pro $i > \bar{m}$. Explicitní tvar

matice $\tilde{M}_i$ byl odvozen pouze pro metody DFP, BFGS a R1 (vzorce (753), (754) a (756)). Pro obecnou metodu z Broydenovy třídy takové explicitní vyjádření neznáme. Ukážeme však, že matici $\tilde{M}_i$ lze spočítat rekurentně, přičemž počet potřebných aritmetických operací je řádově stejný jako u metod DFP, BFGS a R1. Předpokládejme, že matice H_{i+1} vznikne z matice $H_1 = \lambda_i I$ pomocí i aktualizací tvaru

$$H_{j+1} = H_j + U_j M_j U_j^T, \qquad 1 \le j \le i, \tag{765}$$

kde $U_j = [d_j, H_j y_j]$ a

$$M_j = \begin{bmatrix} m_{1,j}, & m_{2,j} \\ m_{2,j}, & m_{3,j} \end{bmatrix}.$$

Budeme hledat vyjádření

$$H_{i+1} = H_1 + \bar{U}_i \bar{M}_i \bar{U}_i^T, \tag{766}$$

kde $\bar{U}_i = [d_1, H_1 y_1, \ldots, d_i, H_1 y_i]$ a $\bar{M}_i$ je symetrická matice řádu $2i$. Poznamenejme, že volba $H_1 = \lambda_i I$ (kde obvykle $\lambda_i = d_i^T y_i / y_i^T y_i$) je podstatná pro konstrukci efektivního algoritmu.

Věta 186. *Nechť H_{i+1} je matice získaná z matice H_1 pomocí i aktualizací (765). Pak platí (766), kde $\bar{M}_i$ je matice určená rekurentně tak že $\bar{M}_1 = M_1$ a*

$$\bar{M}_j = \begin{bmatrix} \bar{M}_{j-1} + m_{3,j}\, z_{j-1} z_{j-1}^T, & m_{2,j}\, z_{j-1}, & m_{3,j}\, z_{j-1} \\ m_{2,j}\, z_{j-1}^T, & m_{1,j}, & m_{2,j} \\ m_{3,j}\, z_{j-1}^T, & m_{2,j}, & m_{3,j} \end{bmatrix}, \qquad 2 \le j \le i, \tag{767}$$

kde

$$z_{j-1} = \bar{M}_{j-1} \bar{r}_{j-1}, \qquad \bar{r}_{j-1} = \bar{U}_{j-1}^T y_j. \tag{768}$$

Důkaz Důkaz provedeme indukcí. Předpokládejme, že pro nějaký index $2 \le j < i$ platí

$$H_j = H_1 + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T \tag{769}$$

(podle (765) to platí pro $j = 2$ neboť $\bar{U}_1 = U_1$ a $\bar{M}_1 = M_1$). Dosadíme-li (769) do (765) a přihlédneme-li k tomu, že podle (768) a (769) platí

$$U_j = [d_j, H_j y_j] = \left[d_j,\, H_1 y_j + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T y_j\right] = \left[d_j,\, H_1 y_j + \bar{U}_{j-1} z_{j-1}\right],$$

můžeme psát

$$\begin{aligned}
H_{j+1} &= H_1 + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T + \left[d_j,\, H_1 y_j + \bar{U}_{j-1} z_{j-1}\right] M_j \left[d_j,\, H_1 y_j + \bar{U}_{j-1} z_{j-1}\right]^T \\
&= H_1 + \bar{U}_{j-1}\, \bar{M}_{j-1}\, \bar{U}_{j-1}^T + m_{1,j}\, d_j d_j^T \\
&\quad + m_{2,j} \left(d_j (H_1 y_j)^T + H_1 y_j d_j^T\right) + m_{2,j} \left(d_j\, (\bar{U}_{j-1} z_{j-1})^T + \bar{U}_{j-1} z_{j-1}\, d_j^T\right) \\
&\quad + m_{3,j}\, H_1 y_j (H_1 y_j)^T + m_{3,j} \left(H_1 y_j\, (\bar{U}_{j-1} z_{j-1})^T + \bar{U}_{j-1} z_{j-1}\, (H_1 y_j)^T\right) \\
&\quad + m_{3,j}\, \bar{U}_{j-1} z_{j-1} z_{j-1}^T \bar{U}_{j-1}^T \\
&= H_1 + \left[\bar{U}_{j-1},\, d_j,\, H_1 y_j\right] \begin{bmatrix} \bar{M}_{j-1} + m_{3,j}\, z_{j-1} z_{j-1}^T, & m_{2,j}\, z_{j-1}, & m_{3,j}\, z_{j-1} \\ m_{2,j}\, z_{j-1}^T, & m_{1,j}, & m_{2,j} \\ m_{3,j}\, z_{j-1}^T, & m_{2,j}, & m_{3,j} \end{bmatrix} \left[\bar{U}_{j-1},\, d_j,\, H_1 y_j\right]^T \\
&= H_1 + \bar{U}_j \bar{M}_j \bar{U}_j^T,
\end{aligned}$$

čímž je indukční krok dokončen. □

Poznámka 293. Porovnáme-li (765) s (283), vidíme, že platí

$$m_{1,j} = \frac{1}{b_j} \left(\eta_j \frac{a_j}{b_j} + 1\right), \qquad m_{2,j} = -\frac{\eta_j}{b_j}, \qquad m_{3.j} = \frac{\eta_j - 1}{a_j}, \tag{770}$$

kde $a_j = y_j^T H_j y_j$ a $b_j = y_j^T d_j$. Použijeme-li (768) a (769), dostaneme

$$a_j = y_j^T H_j y_j = y_j^T (H_1 y_j + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T y_j) = y_j^T H_1 y_j + \bar{r}_{j-1}^T z_{j-1},$$

takže hodnotu a_j, potřebnou pro výpočet diagonálních prvků matice M_j podle (770), lze určit pomocí vektorů $\bar{r}_{j-1}$ a z_{j-1}.

Ukážeme nyní jak se konstruuje matice $H_{i+1} = \lambda_i I + \bar{U}_i \bar{M}_i \bar{U}_i^T$ v obecném případě, kdy může platit $i > \bar{m}$. Nechť $m = \min(\bar{m}, i)$ a $S_i = \mathrm{diag}(1, \lambda_i, \ldots, 1, \lambda_i)$ (kde $\lambda_i > 0$) je diagonální matice řádu $2m$. Označme

$$\check{U}_{i-1} = [d_{i-m+1}, y_{i-m+1}, \ldots, d_{i-1}, y_{i-1}], \quad \check{R}_{i-1} = \begin{bmatrix} d_{i-m+1}^T y_{i-m+1}, & \ldots & d_{i-m+1}^T y_{i-1} \\ y_{i-m+1}^T y_{i-m+1}, & \ldots & y_{i-m+1}^T y_{i-1} \\ \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots \\ 0, & \ldots & d_{i-1}^T y_{i-1} \\ 0, & \ldots & y_{i-1}^T y_{i-1} \end{bmatrix} \tag{771}$$

(tyto matice jsou prázdné, pokud $i = 1$) a

$$\hat{U}_i = [\check{U}_{i-1}, d_i, y_i], \qquad \hat{R}_i = \begin{bmatrix} \check{R}_{i-1}, & \check{U}_{i-1}^T y_i \\ 0, & d_i^T y_i \\ 0, & y_i^T y_i \end{bmatrix}, \tag{772}$$

takže $\check{R}_{i-1}$ a $\hat{R}_i$ jsou blokově horní trojúhelníkové matice, jejichž každý blok obsahuje dva řádky a jeden sloupec. Nechť $k = j - (i - m)$, takže $k = 1$, pokud $j = i - m + 1$ a $k = m$, pokud $j = i$. Pak $\bar{U}_i = S_i \hat{U}_i$ a matici $\bar{M}_i \triangleq \hat{M}_i^i$ (používáme další horní index) určíme rekurentně tak že položíme

$$\hat{M}_{i-m+1}^i = \begin{bmatrix} m_{1,i-m+1}^i, & m_{2,i-m+1}^i \\ m_{2,i-m+1}^i, & m_{3,i-m+1}^i \end{bmatrix} \tag{773}$$

a pro $i - m + 1 \le j \le i - 1$ spočteme vektor $z_j^i = \hat{M}_j^i S_j^i \hat{r}_j^i$, kde S_j^i je hlavní podmatice řádu $2k$ matice S_i a $\hat{r}_j^i$ je vektor dimenze $2k$ obsahující prvních $2k$ prvků k-tého sloupce matice $\check{R}_{i-1}$, a položíme

$$\hat{M}_{j+1}^i = \begin{bmatrix} \hat{M}_j^i + m_{3,j+1}^i z_j^i (z_j^i)^T, & m_{2,j+1}^i z_j^i, & m_{3,j+1}^i z_j^i \\ m_{2,j+1}^i (z_j^i)^T, & m_{1,j+1}^i, & m_{2,j+1}^i \\ m_{3,j+1}^i (z_j^i)^T, & m_{2,j+1}^i, & m_{3,j+1}^i \end{bmatrix}. \tag{774}$$

Použijeme-li matice $\hat{U}_i$ a $\hat{M}_i^i$, můžeme směrový vektor s_{i+1} určit podle vzorce

$$s_{i+1} = -H_{i+1} g_{i+1} = -\lambda_i g_{i+1} - \bar{U}_i \bar{M}_i \bar{U}_i^T g_{i+1} = -\lambda_i g_{i+1} - \hat{U}_i S_i \hat{M}_i^i S_i \hat{U}_i^T g_{i+1}. \tag{775}$$

V tomto případě spotřebujeme zhruba $6mn$ opreací násobení a sčítání ($2mn$ na určení posledního sloupce matice $\hat{R}_i$ a $4mn$ na výpočet vektoru s_{i+1} podle vzorce (775)) a zhruba $2mn$ hodnot je třeba ukládat v paměti počítače. Matice $\check{U}_i$ a $\check{R}_i$, používané v dalším iteračním kroku, lze snadno získat z matic $\hat{U}_i$ a $\hat{R}_i$. Jestliže $i < \bar{m}$, pak $\check{U}_i = \hat{U}_i$ a $\check{R}_i = \hat{R}_i$. Jestliže $i \ge \bar{m}$, pak $\check{U}_i$ a $\check{R}_i$ vznikne z $\hat{U}_i$ a $\hat{R}_i$ odstraněním řádků a sloupců, jejichž prvky se určují pomocí vektorů s indexem $i - m + 1$. Můžeme tedy psát

$$[d_{i-m+1}, y_{i-m+1}, \check{U}_i] = \hat{U}_i, \quad \begin{bmatrix} d_{i-m+1}^T y_{i-m+1}, & [d_{i-m+1}^T y_{i-m+2}, \ldots, d_{i-m+1}^T y_i] \\ y_{i-m+1}^T y_{i-m+1}, & [y_{i-m+1}^T y_{i-m+2}, \ldots, y_{i-m+1}^T y_i] \\ 0, & \check{R}_i \end{bmatrix} = \hat{R}_i. \tag{776}$$

Popsaný postup lze upravit tak, že se ušetří zhruba $2mn$ opreací násobení a sčítání. Z vyjádření (774) je patrné, že ke konstrukci matice $\hat{M}_i^i$ nepotřebujeme poslední sloupec $\hat{r}_i$ matrice $\hat{R}_i$. Proto je možné

počítat místo vektoru $\hat{r}_i = \hat{U}_i^T y_i$ vektor $\hat{v}_i = \hat{U}_i^T g_{i+1}$, který se pak použije k určení směrového vektoru s_{i+1} podle vzorce

$$s_{i+1} = -\lambda_i g_{i+1} - \hat{U}_i S_i \hat{M}_i^i S_i \hat{v}_i. \tag{777}$$

Po spočtení směrového vektoru s_{i+1} lze určit prvních $2(m-1)$ prvků vektoru $\hat{r}_i$ podle vzorce

$$\check{U}_{i-1}^T y_i = \check{U}_{i-1}^T g_{i+1} - \check{U}_{i-1}^T g_i, \tag{778}$$

kde vektor $\check{U}_{i-1}^T g_{i+1}$ obsahuje prvních $2(m-1)$ prvků vektoru $\hat{v}_i$ (viz (772)) a vektor $\check{U}_{i-1}^T g_i$ obsahuje posledních $2(m-1)$ prvků vektoru $\hat{v}_{i-1}$ (vektor $\hat{v}_{i-1}$ známe z předchozího iteračního kroku). Poslední dva prvky $d_i^T y_i$ a $y_i^T y_i$ vektoru $\hat{r}_i$ se počítají zvlášť, neboť je potřebujeme k určení škálovacího parametru λ_i.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 20. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$.

Step 1 Nechť $\check{U}_0 \in R^{n\times 0}$ a $\check{R}_0 \in R^{0\times 0}$ jsou prázdné matice. Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $s_1 := -g_1$ and $i := 1$.

Step 2 Pokud $\|g_i\| \le \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$.

Step 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$ a $y_i := g_{i+1} - g_i$ a vypočteme hodnoty $d_i^T y_i$, $y_i^T y_i$ a $\lambda_i := d_i^T y_i / y_i^T y_i$ definující diagonální matici $S_i := \mathrm{diag}(1, \lambda_i, \ldots, 1, \lambda_i)$ řádu $2m$.

Step 4 Určíme matici $\hat{M}_{i-m+1}^i$ podle vzorce (773). Položíme $\hat{U}_i := [\check{U}_{i-1}, d_i, y_i]$ a $\hat{v}_i := \hat{U}_i^T g_{i+1}$. Položíme $j := i - m + 1$ a $k = 1$.

Step 5 Pokud $j = i$ přejdeme na krok 7.

Step 6 Zvolíme hodnotu parametru η_j^i vystupujícího v (770). Položíme $z_j^i := \hat{M}_j^i S_j^i \check{r}_j^i$, kde S_j^i je hlavní submatice řádu $2k$ matice S_i a $\check{r}_j^i$ je vektor dimanze $2k$ obsahující prvních $2k$ prvků k-tého sloupce matice $\check{R}_{i-1}$, vypočteme matici $\hat{M}_{j+1}^i$ podle vzorce (774), položíme $j := j+1$, $k := k+1$ a přejdeme na krok 5.

Step 7 Vypočteme směrový vektor s_{i+1} podle vzorce (777). Určíme vektor $\check{U}_{i-1}^T y_i$ podle vzorce (778) a matici $\hat{R}_i$ podle vzorce (772).

Step 8 Pokud $i < \bar{m}$, položíme $\check{U}_i := \hat{U}_i$ a $\check{R}_i := \hat{R}_i$. V opačném případě určíme matice $\check{U}_i$ a $\check{R}_i$ podle vzorce (776). Položíme $i := i + 1$ a přejdeme na krok 2.

Způsobem, který jsme právě popsali, lze realizovat i aktualizace z Davidonovy třídy metod s proměnnou metrikou popsané v oddílu 4.8. V tom případě předpokládáme, že matice H_{i+1} a vektor u_{i+1} vzniknou z matice $H_i = \lambda_i I$ a vektoru u_1 pomocí aktualizací tvaru

$$H_{j+1} = H_j + U_j M_j U_j^T, \qquad u_{j+1} = \left(I - \frac{u_j y_j^T}{y_j^T u_j}\right)(d_j - H_j y_j),, \qquad 1 \le j \le i, \tag{779}$$

kde $U_j = [u_j,\ d_j - H_j y_j]$ a

$$M_j = \begin{bmatrix} m_{1,j}, & m_{2,j} \\ m_{2,j}, & m_{3,j} \end{bmatrix}.$$

Budeme hledat vyjádření

$$H_{i+1} = H_1 + \bar{U}_i \bar{M}_i \bar{U}_i^T, \tag{780}$$

kde $\bar{U}_i = [u_1, d_1 - H_1 y_1, \ldots, u_i, d_i - H_1 y_i]$ a $\bar{M}_i$ je symetrická matice řádu $2i$.

Věta 187. *Nechť H_{i+1} je matice získaná z matice H_1 pomocí i aktualizací (779). Pak platí (780), kde $\bar{M}_i$ je matice určená rekurentně tak že $\bar{M}_1 = M_1$ a*

$$\bar{M}_j = \begin{bmatrix} \bar{M}_{j-1} + m_{3,j}\, z_{j-1} z_{j-1}^T, & m_{2,j}\, z_{j-1}, & m_{3,j}\, z_{j-1} \\ m_{2,j}\, z_{j-1}^T, & m_{1,j}, & m_{2,j} \\ m_{3,j}\, z_{j-1}^T, & m_{2,j}, & m_{3,j} \end{bmatrix}, \qquad 2 \le j \le i, \tag{781}$$

kde

$$z_{j-1} = \bar{M}_{j-1} \bar{r}_{j-1}, \qquad \bar{r}_{j-1} = \bar{U}_{j-1}^T y_j. \tag{782}$$

Proof Důkaz provedeme indukcí. Předpokládejme, že pro nějaký index $2 \le j < i$ platí

$$H_j = H_1 + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T \tag{783}$$

(podle (779) to platí pro $j = 2$ neboť $\bar{U}_1 = U_1$ a $\bar{M}_1 = M_1$). Dosadíme-li (783) do (779) a přihlédneme-li k tomu, že podle (768) a (769) platí

$$U_j = [u_j, d_j - H_j y_j] = \left[u_j,\, d_j - H_1 y_j + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T y_j\right] = \left[u_j\, v_j + \bar{U}_{j-1} z_{j-1}\right],$$

kde $v_j = d_j - H_1 y_j$, můžeme psát

$$\begin{aligned} H_{j+1} &= H_1 + \bar{U}_{j-1} \bar{M}_{j-1} \bar{U}_{j-1}^T + \left[u_j,\, v_j + \bar{U}_{j-1} z_{j-1}\right] M_j \left[u_j,\, v_j + \bar{U}_{j-1} z_{j-1}\right]^T \\ &= H_1 + \bar{U}_{j-1}\, \bar{M}_{j-1}\, \bar{U}_{j-1}^T + m_{1,j}\, u_j u_j^T \\ &\quad + m_{2,j} \left(u_j v_j^T + v_j d_j^T\right) + m_{2,j} \left(u_j \left(\bar{U}_{j-1} z_{j-1}\right)^T + \bar{U}_{j-1} z_{j-1}\, u_j^T\right) \\ &\quad + m_{3,j}\, v v_j^T + m_{3,j} \left(v_j \left(\bar{U}_{j-1} z_{j-1}\right)^T + \bar{U}_{j-1} z_{j-1}\, v_j^T\right) \\ &\quad + m_{3,j}\, \bar{U}_{j-1} z_{j-1} z_{j-1}^T \bar{U}_{j-1}^T \\ &= H_1 + \left[\bar{U}_{j-1},\, u_j,\, v_j\right] \begin{bmatrix} \bar{M}_{j-1} + m_{3,j}\, z_{j-1} z_{j-1}^T, & m_{2,j}\, z_{j-1}, & m_{3,j}\, z_{j-1} \\ m_{2,j}\, z_{j-1}^T, & m_{1,j}, & m_{2,j} \\ m_{3,j}\, z_{j-1}^T, & m_{2,j}, & m_{3,j} \end{bmatrix} \left[\bar{U}_{j-1},\, u_j,\, v_j\right]^T \\ &= H_1 + \bar{U}_j \bar{M}_j \bar{U}_j^T, \end{aligned}$$

9.4 Modifikované maticové metody s proměnnou metrikou s omezenou pamětí

Modifikovaná maticová metoda s proměnnou metrikou s omezenou pamětí, studovaná v práci [166], je založena na skutečnosti, že vhodnou transformací vektorů d_i, y_i, $i > 1$, lze zajistit splnění více kvazinewtonovských podmínek. Na rozdíl od metody popsané v oddílu 9.2 se k této transformaci používá více předchozích vektorů. Z tohoto důvodu není výhodné používat Strangovy rekurence. Místo toho se používají maticové vzorce (761) a (762), ve kterých vystupují transformované vektory. Budeme tedy předpokládat, že

$$H_{i+1} = (\bar{R}_i^{-1} \bar{D}_i^T)^T \bar{C}_i \bar{R}_i^{-1} \bar{D}_i^T + \gamma_i (I - \bar{Y}_i \bar{R}_i^{-1} \bar{D}_i^T)^T (I - \bar{Y}_i \bar{R}_i^{-1} \bar{D}_i^T) \tag{784}$$

a směrový vektor budeme počítat podle vzorce

$$\begin{aligned} s_{i+1} = -H_{i+1} g_{i+1} \;=\; & -\gamma_i g_{i+1} - \bar{D}_i (\bar{R}_i^{-1})^T \left[(\bar{C}_i + \gamma_i \bar{Y}_i^T \bar{Y}_i) \bar{R}_i^{-1} \bar{D}_i^T g_{i+1} - \gamma_i \bar{Y}_i^T g_{i+1}\right] \\ & + \gamma_i \bar{Y}_i \left(\bar{R}_i^{-1} \bar{D}_i^T g_{i+1}\right), \end{aligned} \tag{785}$$

kde matice $\bar{D}_i$, $\bar{Y}_i$, $\bar{R}_i$, $\bar{C}_i$ se liší od matic D_i, Y_i, R_i, C_i (uvedených v oddílu 9.3) tím, že jsou definovány pomocí transformovaných vektorů $\bar{d}_j$, $\bar{y}_j$, $i - m \le j \le i - 1$.

Abychom se vyhnuli dvojímu indexování, budeme tak jako v oddílu 9.3 nejprve předpokládat, že $i \le \bar{m}$. Pak matice H_j^i, $1 \le j \le i$, nezávisejí na horním indexu, který můžeme vynechat. Případ, kdy $i > \bar{m}$ probereme později. Transformace, které použijeme, mají obecně tvar $\bar{d}_1 = d_1$, $\bar{y}_1 = y_1$ a $\bar{d}_i = d_i - \bar{D}_{i-1} \bar{\lambda}_i$, $\bar{y}_i = y_i - \bar{Y}_{i-1} \bar{\omega}_i$, $i > 1$, kde $\bar{\lambda}_i$ a $\bar{\omega}_i$ jsou vektory dimenze $i - 1$. Jelikož numerické testy dokládají, že není

výhodné používat k transformacím všechny sloupce matic $\bar{D}_{i-1}$ a $\bar{Y}_{i-1}$, budeme některé (odpovídající si) prvky vektorů $\bar{\lambda}_i$ a $\bar{\omega}_i$ nulovat. Abychom zjednodušili označování, definujeme množinu $\hat{\mathcal{I}}_{i-1}$, obsahující indexy nenulových prvků vektorů $\bar{\lambda}_i$ a $\bar{\omega}_i$. Dále označíme symboly $\hat{D}_{i-1}$, $\hat{Y}_{i-1}$, $\hat{R}_{i-1}$, $\hat{C}_{i-1}$ a $\hat{\lambda}_i$, $\hat{\omega}_i$ matice a vektory, které vzniknou z $\bar{D}_{i-1}$, $\bar{Y}_{i-1}$, $\bar{R}_{i-1}$, $\bar{C}_{i-1}$ a $\bar{\lambda}_i$, $\bar{\omega}_i$ vyškrtnutím řádků a sloupců, jejichž indexy nepatří do množiny $\hat{\mathcal{I}}_{i-1}$. Pak lze transformace zapsat ve tvaru

$$\bar{d}_i = d_i - \bar{D}_{i-1}\bar{\lambda}_i = d_i - \hat{D}_{i-1}\hat{\lambda}_i \qquad \bar{y}_i = y_i - \bar{Y}_{i-1}\bar{\omega}_i = y_i - \hat{Y}_{i-1}\hat{\omega}_i. \tag{786}$$

Princip popisované metody spočívá v tom, že se vektory $\bar{d}_i$, $\bar{y}_i$, $i > 1$, konstruují induktivně tak, aby platilo $H_{i+1}\hat{Y}_i = \hat{D}_i$, kde $\hat{\mathcal{I}}_i \setminus \{i\} \subset \hat{\mathcal{I}}_{i-1}$ (pokud $i \leq \bar{m}$), přičemž předpokládáme, že matice H_i splňuje podmínku $H_i\hat{Y}_{i-1} = \hat{D}_{i-1}$. Abychom zjednodušili označování, budeme tak jako v oddílu 9.3 indexy $i-1$ a i často vynechávat a index $i+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $\hat{\mathcal{I}} = \hat{\mathcal{I}}_{i-1}$, $\hat{D} = \hat{D}_{i-1}$, $\hat{Y} = \hat{Y}_{i-1}$, $\hat{\lambda} = \hat{\lambda}_i$, $\hat{\omega} = \hat{\omega}_i$, $H = H_i$, $\bar{D} = \bar{D}_{i-1}$, $\bar{Y} = \bar{Y}_{i-1}$, $\bar{R} = \bar{R}_{i-1}$, $\bar{C} = \bar{C}_{i-1}$ a $H_+ = H_{i+1}$, $\bar{D}_+ = \bar{D}_i = [\bar{D}, \bar{d}]$, $\bar{Y}_+ = \bar{Y}_i = [\bar{Y}, \bar{y}]$, takže například

$$H = (\bar{R}^{-1}\bar{D}^T)^T\bar{C}\bar{R}^{-1}\bar{D}^T + (I - \bar{Y}\bar{R}^{-1}\bar{D}^T)^T(I - \bar{Y}\bar{R}^{-1}\bar{D}^T), \tag{787}$$

$$H_+ = \bar{V}^T H\bar{V} + \frac{1}{\bar{b}}\bar{d}\,\bar{d}^T, \qquad \bar{V} = I - \frac{1}{\bar{b}}\bar{y}\,\bar{d}^T, \tag{788}$$

$$\bar{d} = d - \hat{D}\hat{\lambda}, \qquad \bar{y} = y - \hat{Y}\hat{\omega}. \tag{789}$$

Nejprve ukážeme základní vlastnosti použitých transformací a jejich vliv na splnění více kvazinewtonovských podmínek.

Věta 188. *Nechť H je symetrická pozitivně definitní matice taková, že $H\hat{Y} = \hat{D}$, H_+ je matice určená podle vzorce (788), kde $\bar{b} > 0$. Pak H_+ je pozitivně definitní a platí*

$$(H_+\bar{y}_j - \bar{d}_j)^T B_+(H_+\bar{y}_j - \bar{d}_j) = \frac{1}{\bar{b}}\left(\bar{d}_j^T\bar{y} - \bar{d}^T\bar{y}_j)^2 + \left(\frac{\bar{a}}{\bar{b}} - \frac{\bar{b}}{\bar{c}}\right)(\bar{d}^T\bar{y}_j)^2\right) \qquad \forall j \in \hat{\mathcal{I}}, \tag{790}$$

kde $\bar{a} = \bar{y}^T H\bar{y}$, $\bar{c} = \bar{d}^T B\bar{d}$ a $B = H^{-1}$, $B_+ = H_+^{-1}$, takže $\bar{a}/\bar{b} \geq \bar{b}/\bar{c}$ a $\bar{a}/\bar{b} = \bar{b}/\bar{c}$ právě tehdy, jsou-li vektory $\bar{d}$ a $H\bar{y}$ lineárně závislé. Jestliže navíc

$$\hat{D}^T\bar{y} = 0, \qquad \hat{Y}^T\bar{d} = 0, \tag{791}$$

je splněna kvazinewtonovská podmínka $H_+\hat{Y} = \hat{D}$ a platí

$$(H_+y - d)^T B_+(H_+y - d) = (\hat{\lambda} - \hat{\omega})^T\hat{D}^T\hat{Y}(\hat{\lambda} - \hat{\omega}). \tag{792}$$

Důkaz (a) Jelikož (788) odpovídá aktualizaci BFGS, můžeme podle (288) a (309) psát

$$H_+ = H + \left(1 + \frac{\bar{a}}{\bar{b}}\right)\frac{\bar{d}\bar{d}^T}{\bar{b}} - \frac{H\bar{y}\bar{d}^T + \bar{d}(H\bar{y})^T}{\bar{b}}, \qquad B_+ = B + \frac{\bar{y}\bar{y}^T}{\bar{b}} - \frac{B\bar{d}(B\bar{d})^T}{\bar{c}}. \tag{793}$$

Použitím těchto rovností pro $j \in \hat{\mathcal{I}}$ dostaneme

$$\bar{y}_j^T H_+\bar{y}_j = \bar{d}_j^T\bar{y}_j + \left(1 + \frac{\bar{a}}{\bar{b}}\right)\frac{(\bar{d}^T\bar{y}_j)^2}{\bar{b}} - 2\frac{\bar{d}^T\bar{y}_j\bar{d}_j^T\bar{y}}{\bar{b}} \tag{794}$$

$$\bar{d}_j^T B_+\bar{d}_j = \bar{d}_j^T\bar{y}_j + \frac{(\bar{d}_j^T\bar{y})^2}{\bar{b}} - \frac{(\bar{d}^T\bar{y}_j)^2}{\bar{c}}, \tag{795}$$

neboť podle předpokladu platí $H\bar{y}_j = \bar{d}_j$, $j \in \hat{\mathcal{I}}$, takže

$$\begin{aligned}(H_+\bar{y}_j - \bar{d}_j)^T B_+(H_+\bar{y}_j - \bar{d}_j) &= \bar{y}_j^T H_+\bar{y}_j + \bar{d}_j^T B_+\bar{d}_j - 2\bar{d}_j^T\bar{y}_j \\ &= \frac{1}{\bar{b}}\left((\bar{d}_j^T\bar{y} - \bar{d}^T\bar{y}_j)^2 + (\bar{d}^T\bar{y}_j)^2\left(\frac{\bar{a}}{\bar{b}} - \frac{\bar{b}}{\bar{c}}\right)\right).\end{aligned}$$

(b) Nechť $\hat{D}^T\bar{y} = \hat{Y}^T\bar{d} = 0$, neboli $\bar{d}_j^T\bar{y} = \bar{d}^T\bar{y}_j = 0$, $j \in \hat{\mathcal{I}}$. Pak podle (790) platí $H_+\hat{Y} = \hat{D}$ a z (788) plyne $H_+\bar{y} = \bar{d}$, takže

$$\begin{aligned} H_+y - d &= H_+\bar{y} - H_+(\bar{y} - y) + \bar{d} + (\bar{d} - d) = H_+\hat{Y}\hat{\omega} - \hat{D}\hat{\lambda} = \hat{D}(\hat{\omega} - \hat{\lambda}), \\ B_+(H_+y - d) &= B_+\hat{D}(\hat{\omega} - \hat{\lambda}) = \hat{Y}(\hat{\omega} - \hat{\lambda}), \end{aligned}$$

což dává (792). □

Věta 189. *Nechť matice $\hat{D}^T\hat{Y}$ je regulární. Pak jediné řešení rovnic (791) má tvar*

$$\lambda^* = (\hat{Y}^T\hat{D})^{-1}\hat{Y}^Td, \qquad \omega^* = (\hat{D}^T\hat{Y})^{-1}\hat{D}^Ty. \tag{796}$$

Zvolíme-li vektory $\hat{\lambda}$, $\hat{\omega}$ v (789) libovolně, platí

$$\bar{b} = (\hat{\lambda} - \lambda^*)^T\hat{D}^T\hat{Y}(\hat{\omega} - \omega^*) + b^*, \qquad b^* = (d^*)^Ty^* = b - (\lambda^*)^T\hat{D}^T\hat{Y}\omega^*, \tag{797}$$

kde d^*, y^* jsou vektory $\hat{d}$, $\hat{y}$ pro $\hat{\lambda} = \lambda^*$, $\hat{\omega} = \omega^*$.

Důkaz Dosadíme-li vyjádření (789) do vztahů (791), dostaneme

$$\hat{Y}^T\hat{D}\hat{\lambda} = \hat{Y}^Td, \qquad \hat{D}^T\hat{Y}\hat{\omega} = \hat{D}^Ty. \tag{798}$$

Jelikož matice $\hat{D}^T\hat{Y}$ je regulární, mají tyto rovnice jediné řešení (796). Použijeme-li (798), dostaneme

$$\begin{aligned} (\hat{\lambda} - \lambda^*)^T\hat{D}^T\hat{Y}(\hat{\omega} - \omega^*) &= \hat{\lambda}^T\hat{D}^T\hat{Y}\hat{\omega} - \hat{\lambda}^T\left(\hat{D}^T\hat{Y}\omega^*\right) - \left((\lambda^*)^T\hat{D}^T\hat{Y}\right)\hat{\omega} + \left((\lambda^*)^T\hat{D}^T\hat{Y}\right)\omega^* \\ &= \left(\hat{\lambda}^T\hat{D}^T\hat{Y}\hat{\omega} - \hat{\lambda}^T\hat{D}^Ty - d^T\hat{Y}\hat{\omega}\right) + d^T\hat{Y}(\hat{D}^T\hat{Y})^{-1}\hat{D}^Ty, \end{aligned}$$

odkud plyne

$$\begin{aligned} \hat{b} &= (d - \hat{D}\hat{\lambda})^T(y - \hat{Y}\hat{\omega}) = \left(\hat{\lambda}^T\hat{D}^T\hat{Y}\hat{\omega} - \hat{\lambda}^T\hat{D}^Ty - d^T\hat{Y}\hat{\omega}\right) + b \\ &= (\hat{\lambda} - \lambda^*)^T\hat{D}^T\hat{Y}(\hat{\omega} - \omega^*) + b - d^T\bar{Y}(\hat{D}^T\hat{Y})^{-1}\hat{D}^Ty \\ &= (\hat{\lambda} - \lambda^*)^T\hat{D}^T\hat{Y}(\hat{\omega} - \omega^*) + b^*, \end{aligned}$$

neboť podle (798) platí

$$\begin{aligned} b^* = (d^*)^Ty^* &= \left(d - \hat{D}(\hat{Y}^T\hat{D})^{-1}\hat{Y}^Td\right)\left(y - \hat{Y}(\hat{D}^T\hat{Y})^{-1}\hat{D}^Ty\right) \\ &= d^Ty - d^T\bar{Y}(\hat{D}^T\hat{Y})^{-1}\hat{D}^Ty = b - (\lambda^*)^T\hat{D}^T\hat{Y}\omega^*. \end{aligned}$$

□

Metody s proměnnou metrikou vyžadují, aby matice H_+ byla symetrická. Jelikož se snažíme, aby platilo $H_+\hat{Y}_+ = \hat{D}_+$, budeme požadovat, aby matice $\hat{D}_+^T\hat{Y}_+ = \hat{Y}_+^TH_+\hat{Y}_+$ byla symetrická. Předpokládáme-li, že matice $\hat{D}^T\hat{Y}$ je symetrická, je matice

$$\hat{D}_+^T\hat{Y}_+ = \begin{bmatrix} \hat{D}^T\hat{Y} & \hat{D}^T\hat{y} \\ \hat{d}^T\hat{Y} & \hat{d}^T\hat{y} \end{bmatrix}$$

symetrická, pokud $\hat{D}^T\hat{y} = \hat{Y}^T\hat{d}$. Budeme tedy předpokládat, že $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$, kde $\mathcal{S} = \{(\bar{\lambda}, \bar{\omega}) : \hat{D}^T\bar{y} = \hat{Y}^T\bar{d}\}$. Zřejmě $(\lambda^*, \omega^*) \in \mathcal{S}$, neboť podle věty 189 platí $\hat{D}^Ty^* = \hat{Y}^Td^* = 0$. Ukážeme, že některé důležité veličiny jsou nezávislé na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$.

Věta 190. *Nechť H je symetrická pozitivně definitní matice taková, že $H\hat{Y} = \hat{D}$. Nechť matice $\hat{D}^T\hat{Y}$ je regulární a $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Pak*

(a) Matice $\hat{D}^T\hat{Y}$ je symetrická a pozitivně definitní.

(b) Rozdíly $\hat{\lambda} - \hat{\omega}$, $\bar{d} - H\bar{y}$, $B\bar{d} - \bar{y}$, $\bar{b} - \bar{a}$, $\bar{c} - \bar{b}$, jsou nezávislé na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$.

(c) Hodnoty $\bar{a}$, $\bar{b}$, $\bar{c}$ nabývají svého minima, pokud $\hat{\lambda} = \lambda^*$, $\hat{\omega} = \omega^*$, kde λ^*, ω^* jsou čísla uvedená v (796).

Důkaz (a) Matice $\hat{D}^T\hat{Y}$ je symetrická a pozitivně definitní, neboť z $H\hat{Y} = \hat{D}$ plyne $\hat{Y}^T H\hat{Y} = \hat{D}^T\hat{Y}$ a matice $\hat{Y}^T H\hat{Y}$ je symetrická a pozitivně definitní (matice H je pozitivně definitní a matice $\hat{Y}$ má lineárně nezávislé sloupce).

(b) Z předpokladu $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$ a vztahu (789) plyne $\hat{D}^T(y - \hat{Y}\hat{\omega}) = \hat{Y}^T(d - \hat{D}\hat{\lambda})$, neboli

$$\hat{D}^T\hat{Y}(\hat{\lambda} - \hat{\omega}) = \hat{Y}^T d - \hat{D}^T y, \tag{799}$$

takže rozdíl $\hat{\lambda} - \hat{\omega}$, nezávisí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Jelikož $H\hat{Y} = \hat{D}$, můžeme psát

$$\bar{d} - H\bar{y} = d - Hy - \hat{D}(\hat{\lambda} - \hat{\omega}), \qquad B\bar{d} - \bar{y} = B(\bar{d} - H\bar{y}), \tag{800}$$

takže podle (799) rozdíly $\bar{d} - H\bar{y}$, $B\bar{d} - \bar{y}$ nezávisejí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Použijeme-li vztahy (789), $H\hat{Y} = \hat{D}$ a předpoklad $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$, dostaneme

$$\begin{aligned}
\bar{b} - \bar{a} &= \bar{y}^T(\bar{d} - H\bar{y}) = y^T(\bar{d} - H\bar{y}) - \hat{\omega}^T\hat{Y}^T(\bar{d} - H\bar{y}) = y^T(\bar{d} - H\bar{y}) - \hat{\omega}^T(\hat{Y}^T\bar{d} - \hat{D}^T\bar{y}) \\
&= y^T(\bar{d} - H\bar{y}) \\
\bar{c} - \bar{b} &= \bar{d}^T(B\bar{d} - \bar{y}) = d^T(B\bar{d} - \bar{y}) - \hat{\lambda}^T\hat{D}^T(B\bar{d} - \bar{y}) = d^T(B\bar{d} - \bar{y}) - \hat{\lambda}^T(\hat{Y}^T\bar{d} - \hat{D}^T\bar{y}) \\
&= d^T(B\bar{d} - \bar{y}),
\end{aligned}$$

takže podle (800) rozdíly $\bar{b} - \bar{a}$ a $\bar{c} - \bar{b}$ nezávisejí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$.

(c) Jelikož $(\lambda^*, \omega^*) \in \mathcal{S}$, a rozdíl $\hat{\lambda} - \hat{\omega}$, nezávisí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$, platí $\hat{\lambda} - \hat{\omega} = \lambda^* - \omega^*$, neboli $\hat{\lambda} - \lambda^* = \hat{\omega} - \omega^*$, což spolu s (797) dává $\bar{b} = (\hat{\lambda} - \lambda^*)\hat{D}^T\hat{Y}(\hat{\lambda} - \lambda^*) + b^*$. Jelikož matice $\hat{D}^T\hat{Y}$ je pozitivně definitní, je tento výraz minimální, pokud $\hat{\lambda} = \lambda^*$ a tedy i $\hat{\omega} = \omega^*$. Z vyjádření $\bar{a} = \bar{b} - (\bar{b} - \bar{a})$, $\bar{c} = \bar{b} + (\bar{c} - \bar{b})$ a nezávislosti rozdílů $\bar{b} - \bar{a}$, $\bar{c} - \bar{b}$ na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$ plyne zbytek tvrzení. □

Ukážeme nyní, že modifikovaná maticová metoda s proměnnou metrikou s omezenou pamětí je v jistém smyslu optimální, aplikujeme-li ji na ryze konvexní kvadratickou funkci s pozitivně definitní Hessovou maticí G.

Lemma 85. *Nechť H je symetrická pozitivně definitní matice taková, že $H\hat{Y} = \hat{D}$, matice $\hat{D}^T\hat{Y}$ je regulární a $b^* > 0$. Pak pro libovolnou dvojici $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$ platí $\bar{b} > 0$. Nechť G je symetrická pozitivně definitní matice taková, že $G\hat{D} = \hat{Y}$ a $G(d - \hat{D}\hat{\lambda}) = y - \hat{Y}\hat{\omega}$ pro nějakou dvojici $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Pak*

(a) $G\bar{d} = \bar{y}$ pro libovolnou dvojici $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$.

(b) Nechť H_+ je matice určená podle vzorců (788), (789), kde $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Pak hodnota $\|G^{1/2}H_+G^{1/2} - I\|_F^2$ nabývá svého je minima, pokud $\hat{\lambda} = \lambda^*$ a $\hat{\omega} = \omega^*$.

Důkaz Nechť $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Jelikož jsou splněny předpoklady věty 189 a věty 190, je matice $\hat{D}^T\hat{Y}$ symetrická a pozitivně definitní a podle (797) platí $\bar{b} \geq b^* > 0$.

(a) Jelikož $G\hat{D} = \hat{Y}$, můžeme ve větě 190 nahradit matici H maticí G^{-1}, takže rozdíl $G\bar{d} - \bar{y}$ nezávisí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$.

(b) Označme $\hat{Z} = G^{1/2}\hat{D} = G^{-1/2}\hat{Y}$, $z = G^{1/2}d$, $\bar{z} = G^{1/2}\bar{d}$ a $R^{-1} = G^{1/2}HG^{1/2}$, $R_+^{-1} = G^{1/2}H_+G^{1/2}$. Pak podle (788) platí

$$R_+^{-1} = \bar{P}R^{-1}\bar{P} + \frac{\bar{z}\bar{z}^T}{\bar{z}^T\bar{z}} = I + \bar{P}(R^{-1} - I)\bar{P}, \qquad \bar{P} = I - \frac{\bar{z}\bar{z}^T}{\bar{z}^T\bar{z}}, \tag{801}$$

neboť $\bar{b} = \bar{d}^T\bar{y} = \bar{d}^T G\bar{d} = \bar{z}^T\bar{z}$ a $\bar{P}^2 = \bar{P}$. Označíme-li $M = R^{-1} - I$ a použijeme-li stejný postup jako v důkazu lemmatu 80, dostaneme

$$\|R_+^{-1} - I\|_F^2 \quad = \quad \|R^{-1} - I\|_F^2 - 2\frac{\bar{z}^T M^2 \bar{z}}{\bar{z}^T\bar{z}} + \left(\frac{\bar{z}^T M\bar{z}}{\bar{z}^T\bar{z}}\right)^2 \tag{802}$$

(vzorec (733)). Jelikož $H\hat{Y} = \hat{D}$, můžeme psát $M\hat{Z} = R^{-1}\hat{Z} - \hat{Z} = G^{1/2}(H\hat{Y} - \hat{D}) = 0$, takže platí $M\bar{z} = M(z - \hat{Z}\hat{\lambda}) = Mz$. Čísla $\bar{z}^T M^2\bar{z} = z^T M^2 z$ a $\bar{z}^T M\bar{z} = z^T Mz$ tedy nazávisejí na výběru $(\hat{\lambda}, \hat{\omega}) \in \mathcal{S}$. Dosadíme-li získané vztahy do (802), dostaneme

$$\begin{aligned}\|R_+^{-1} - I\|_F^2 &= \|R^{-1} - I\|_F^2 - \frac{2}{\bar{b}} z^T M^2 z + \frac{1}{\bar{b}^2}(z^T Mz)^2 \\ &= \|R^{-1} - I\|_F^2 - 2z^T M^2 z\xi + (z^T Mz)^2\xi^2,\end{aligned}$$

kde $\xi = 1/\bar{b}$. Jelikož

$$\frac{\mathrm{d}}{\mathrm{d}\xi}\|R_+^{-1} - I\|_F^2 = 2(z^T Mz)^2\xi - 2z^T M^2 z = 2\left(\frac{(\bar{z}^T M\bar{z})^2}{\bar{z}^T\bar{z}} - \bar{z}^T M^2\bar{z}\right) \le 0$$

(používáme Schwarzovu nerovnost aplikovanou na vektory $\bar{z}$ a $M\bar{z}$), je norma $\|(R_+)^{-1} - I\|_F$ minimální, je-li ξ maximalní a tedy je-li $\bar{b}$ minimální, což podle věty 190 nastává, pokud $\hat{\lambda} = \lambda^*$ a $\hat{\omega} = \omega^*$. □

Věta 191. *Předpokládejme, že minimalizovaná funkce má tvar (687), kde G je symetrická pozitivně definitní matice. Nechť H je symetrická pozitivně definitní matice taková, že $H\hat{Y} = \hat{D}$, a H_+ je matice určená podle (788). Pak jsou-li sloupce matice $[\hat{D}, d]$ lineárně nezávislé, je matice $\hat{D}^T\hat{Y}$ symetrická pozitivně definitní a vybíráme-li vektory $\hat{\lambda}$ a $\hat{\omega}$ tak, že $\hat{\lambda} = \lambda^*$ a $\hat{\omega} = \omega^* = \lambda^*$, platí $\bar{b} > 0$ (takže matice H_+ je pozitivně definitní), $H_+\bar{y} = \bar{d}$, $H_+ y = d$, $H_+\hat{Y} = \hat{D}$ a hodnota $\|G^{1/2}H_+G^{1/2} - I\|_F$ je minimální.*

Důkaz (a) Dokážeme indukcí, že pokud $(\hat{\lambda}_i, \hat{\omega}_i) = (\lambda_i^*, \omega_i^*)$ pro $i \in N$, platí $\hat{Y}_{i-1} = G\hat{D}_{i-1}$ a $\bar{y}_i = G\bar{d}_i$ pro $i \in N$, takže matice G splňuje předpoklady lemmatu 85. Předpokládejme, že $\hat{Y}_{i-1} = G\hat{D}_{i-1}$ (platí to pro $i = 2$, neboť $\bar{y}_1 = y_1$ a $\bar{d}_1 = d_1$, takže buď $G\hat{D}_1 = [Gd_1] = [y_1] = \hat{Y}_1$, nebo $\hat{D}_1$ a $\hat{Y}_1$ nemají žádné sloupce, takže $\bar{d}_1 = d_1$ a $\bar{y}_1 = y_1$). Pak lze psát

$$\omega_i^* = (\hat{D}_{i-1}^T\hat{Y}_{i-1})^{-1}\hat{D}_{i-1}^T y_i = (\hat{D}_{i-1}^T G\hat{D}_{i-1})^{-1}\hat{D}_{i-1}^T Gd_i = (\hat{Y}_{i-1}^T\hat{D}_{i-1})^{-1}\hat{Y}_{i-1}^T d_i = \lambda_i^*,$$

takže $y_i^* = y_i - \hat{Y}_{i-1}\omega_i^* = G(d_i - \hat{D}_{i-1}\lambda_i^*) = Gd_i^*$ a tedy $\hat{y}_i = G\bar{d}_i$, pokud $(\hat{\lambda}_i, \hat{\omega}_i) \in \mathcal{S}_i$ (lemma 85).

(b) Jelikož sloupce matice $[\hat{D}, d]$ jsou lineárně nezávislé a matice G je pozitivně definitní, je matice $[\hat{D}, d]^T G[\hat{D}, d]$ pozitivně definitní, takže podle (797) platí

$$\left[(\omega^*)^T, -1\right][\hat{D}, d]^T G[\hat{D}, d]\left[(\omega^*)^T, -1\right]^T = (\omega^*)^T\hat{D}^T\hat{Y}\omega^* - 2y^T\hat{D}\omega^* + y^T d = b - (\omega^*)^T\hat{D}^T\hat{Y}\omega^* = b^* > 0,$$

neboť $\lambda^* = \omega^*$ a vektor $[(\omega^*)^T, -1]^T$ je nenulový.

(c) Rovnosti $H_+\bar{y} = \bar{d}$, $H_+\hat{Y}_- = \hat{D}$ a $H_+ y = d$ plynou z (788), (790) a (792). Zbytek tvrzení plyne z lemmatu 85. □

Ukážeme nyní, jak se konstruuje matice H_{i+1} v obecném případě, kdy může platit $i > \bar{m}$. Budeme předpokládat, že $H_1 = \gamma_i I$ a používat vzorec (785), ve kterém $\bar{D}_i = [\bar{d}_{i-m+1}, \ldots, \bar{d}_i]$, $\bar{Y}_i = [\bar{y}_{i-m+1}, \ldots, \bar{y}_i]$, $\bar{C}_i$ obsahuje diagonálu matice $\bar{D}_i^T\bar{Y}_i$ a $\bar{R}_i$ obsahuje horní polovinu matice $\bar{D}_i^T\bar{Y}_i$. Matice $\bar{D}_i$, $\bar{Y}_i$ vzniknou z matic $\bar{D}_{i-1}$, $\bar{Y}_{i-1}$ přidáním nových sloupců $\bar{d}_i$, $\bar{y}_i$, a pokud $i > \bar{m}$, ubráním starých sloupců $\bar{d}_{i-m}$, $\bar{y}_{i-m}$. Podobně jednoduše získáme matice $\bar{D}_i^T\bar{Y}_i$, $\bar{Y}_i^T\bar{Y}_i$ z matic $\bar{D}_{i-1}^T\bar{Y}_{i-1}$, $\bar{Y}_{i-1}^T\bar{Y}_{i-1}$ a tudíž i matice $\bar{C}_i$, $\bar{R}_i$ z matic $\bar{C}_{i-1}$, $\bar{R}_{i-1}$. Tím máme k dispozici všechny matice potřebné k určení matice H_{i+1} a k výpočtu vektoru s_{i+1}. Množinu indexů $\hat{\mathcal{I}}_{i-1} \subset \{i-m+1, \ldots, i-1\}$, pomocí které se vybírají sloupce matic $\hat{D}_{i-1}$, $\hat{Y}_{i-1}$, konstruujeme rekurentně tak, že

$$\hat{\mathcal{I}}_{i-1} \setminus \{i-1\} \subset \hat{\mathcal{I}}_{i-2}, \qquad i > 1 \tag{803}$$

(kde $\hat{\mathcal{I}}_0 = \emptyset$), tedy případným ubráním některých indexů (zejména indexu $\{i-m\}$, pokud $i-m \in \hat{\mathcal{I}}_{i-2}$) a případným přidáním indexu $\{i-1\}$. Odtud plyne, že pokud pro nějaký index $k \in \{i-m+1, \dots, j-1\}$, kde $k < j < i$, platí $k \in \hat{\mathcal{I}}_{i-1}$, můžeme psát $k \in \hat{\mathcal{I}}_{j-1}$, neboť podle (803) platí

$$\hat{\mathcal{I}}_{i-1} \cap \{k, \dots, i-1\} \subset \hat{\mathcal{I}}_{i-2} \cap \{k, \dots, i-2\} \subset \cdots \subset \hat{\mathcal{I}}_{j-1} \cap \{k, \dots, j-1\}. \tag{804}$$

Výhodou, která umožňuje zobecnění studované metody s proměnnou metrikou s omezenou pamětí na případ, kdy $i > \bar{m}$, je skutečnost, že transformace (789), (796) nezávisejí na aktualizované matici H, takže potřebné vlastnosti aktualizací (ukázané v předchozích větách) zůstanou zachovány, vybíráme-li množinu $\hat{\mathcal{I}}$ tak, aby byla splněna podmínka (803).

Věta 192. *Nechť pro libovolný index $i > 1$ je $\hat{\mathcal{I}}_{i-1}$ množina konstruovaná tak, aby byla splněna podmínka (803), $\bar{b}_i > 0$ a $\hat{D}_{i-1}^T \bar{y}_i = 0$, $\hat{Y}_{i-1}^T \bar{d}_i = 0$, kde $\hat{D}_{i-1}$, $\hat{Y}_{i-1}$ jsou matice, jejichž sloupce odpovídají indexům z množiny $\hat{\mathcal{I}}_{i-1}$. Nechť $H_{i-m+1}^{i+1} = \gamma_i I$ a*

$$H_{j+1}^{i+1} = \bar{V}_j^T H_j^{i+1} \bar{V}_j + \frac{\bar{d}_j \bar{d}_j^T}{\bar{b}_j}, \qquad \bar{V}_j = I - \frac{\bar{y}_j \bar{d}_j^T}{\bar{b}_j}, \qquad i-m+1 \le j \le i \tag{805}$$

(takže $H_{i+1} = H_{i+1}^{i+1}$). Pak pro $k \in \hat{\mathcal{I}}_{i-1}$, $k < i$, platí

$$\bar{d}_k^T \bar{y}_j = \bar{y}_k^T \bar{d}_j = 0, \qquad k < j \le i, \tag{806}$$

$$H_j^{i+1} \bar{y}_k = \bar{d}_k, \qquad k < j \le i+1. \tag{807}$$

Důkaz (a) Nechť $i > 1$ a $k \in \hat{\mathcal{I}}_{i-1}$, $k < i$. Pak podle (804) pro libovolný index $k < j \le i$ platí $k \in \hat{\mathcal{I}}_{j-1}$. Jelikož vektory $\bar{d}_j$, $\bar{y}_j$ konstruujeme tak. aby platilo $\hat{D}_{j-1}^T \bar{y}_j = 0$, $\hat{Y}_{j-1}^T \bar{d}_j = 0$ a jelikož $k \in \hat{\mathcal{I}}_{j-1}$, můžeme psát $\bar{d}_k^T \bar{y}_j = \bar{y}_k^T \bar{d}_j = 0$. Tím jsme dokázali rovnosti (806).

(b) Rovnosti (807) dokážeme indukcí. Pro $j = k+1$ platí $H_j^{i+1} \bar{y}_k = \bar{d}_k$ podle (805). Předpokládejme, že tato rovnost platí pro nějaký index $k < j \le i$. Podle (805) a (806) lze psát $\bar{V}_j \bar{y}_k = \bar{y}_k$ a $\bar{V}_j^T \bar{d}_k = \bar{d}_k$, takže s použitím indukčního předpokladu a vztahů (805), (806) dostaneme $H_{j+1}^{i+1} \bar{y}_k = \bar{d}_k$, čímž je indukční krok dokončen. □

Poznámka 294. Věta 192 má velmi důležitý důsledek. Pokud v každém iteračním kroku (s indexem $i \in N$) vybíráme množinu $\hat{\mathcal{I}}_{i-1}$ tak, aby byla splněna podmínka (803) a pokládáme $\hat{\lambda}_i = \lambda_i^*$, $\hat{\omega} = \omega_i^*$ (vzorec (796)), jsou podle (806) matice $\hat{D}_{i-1}^T \hat{Y}_{i-1}$ diagonální (s prvky $\bar{b}_j$, $j \in \hat{\mathcal{I}}_{i-1}$ na hlavní diagonále). V tomto případě se většina výpočtů značně zjednoduší. Prvky vektorů λ_i^*, ω_i^* jsou čísla $d_i^T \bar{y}_j / \bar{b}_j$, $\bar{d}_j^T y_i / \bar{b}_j$, $j \in \hat{\mathcal{I}}_{i-1}$, takže vzorce (792) a (797) lze zapsat ve tvaru

$$(H_{i+1} y_i - d_i) B_{i+1} (H_{i+1} y_i - d_i) = \sum_{j \in \hat{\mathcal{I}}_{i-1}} \frac{(d_i^T \bar{y}_j - \bar{d}_j^T y_i)^2}{\bar{b}_j}, \tag{808}$$

$$b_i^* = b_i - \sum_{j \in \hat{\mathcal{I}}_{i-1}} \frac{\bar{d}_j^T y_i d_i^T \bar{y}_j}{\bar{b}_j}. \tag{809}$$

Můžeme tedy vyšetřovat vliv jednotlivých transformací odděleně a rozhodnout, kdy nevhodnou transformaci vynecháme. Vzorce (786) a (796) můžeme zapsat ve tvaru

$$\bar{d}_i = d_i - \hat{D}_{i-1} (\hat{Y}_{i-1}^T \hat{D}_{i-1})^{-1} \hat{Y}_{i-1}^T d_i = P_i d_i, \qquad \bar{y}_i = y_i - \hat{Y}_{i-1} (\hat{D}_{i-1}^T \hat{Y}_{i-1})^{-1} \hat{D}_{i-1}^T y_i = P_i^T y_i, \tag{810}$$

kde $P_i = \hat{D}_{i-1} (\hat{Y}_{i-1}^T \hat{D}_{i-1})^{-1} \hat{Y}_{i-1}^T$ je matice projekce (platí $P_i^2 = P_i$). Jsou-li splněny předpoklady věty 192, je matice $\hat{Y}_{i-1}^T \hat{D}_{i-1}$ diagonální, takže

$$P_i = I - \sum_{j \in \hat{\mathcal{I}}_{i-1}} \frac{\bar{d}_j \bar{y}_j^T}{\bar{b}_j} = \prod_{j \in \hat{\mathcal{I}}_{i-1}} \left(I - \frac{\bar{d}_j \bar{y}_j^T}{\bar{b}_j} \right) \tag{811}$$

Poznámka 295. Použijeme-li rovnosti (807) a vztahy (810), (811), můžeme psát

$$\bar{y}_i^T H_i^{i+1} \bar{y}_i = y_i^T P_i H_i^{i+1} P_i^T y_i = y_i^T H_i^{i+1} y_i - \sum_{j \in \hat{I}_{i-1}} \frac{(\bar{d}_j^T y_i)^2}{\bar{b}_j}, \tag{812}$$

$$\bar{d}_i^T B_i \bar{d}_i = d_i^T P_i^T B_i P_i d_i = d_i^T B_i d_i - \sum_{j \in \hat{I}_{i-1}} \frac{(d_i^T \bar{y}_j)^2}{\bar{b}_j} = -\alpha_i d_i^T g_i - \sum_{j \in \hat{I}_{i-1}} \frac{(d_i^T \bar{y}_j)^2}{\bar{b}_j}, \tag{813}$$

kde $\bar{y}_i^T H_i^{i+1} \bar{y}_i = \bar{a}_i$ a výraz $\bar{d}_i^T B_i \bar{d}_i = \tilde{c}_i$ aproximuje číslo $\bar{d}_i^T B_i^{i+1} \bar{d}_i = \bar{c}_i$ (číslo $\bar{c}_i$ nelze jednoduše určit, neboť $s_i = -B_i g_i \neq -B_i^{i+1} g_i$).

Věta 192 říká, že vybereme-li množinu $\hat{\mathcal{I}}_{i-1}$ tak, aby platilo $\bar{b}_i^* > 0$, a položíme-li $\hat{\lambda}_i = \lambda_i^*$, $\hat{\omega}_i = \omega_i^*$, je libovolný vektor $\bar{d}_k \in \hat{\mathcal{I}}_{i-1}$ ortogonální ke všem vektorům $\bar{y}_j$, $k < j \leq i$. Je-li minimalizovaná funkce kvadratická s pozitivně definitní Hessovou maticí G, platí podle věty 191 $b^* > 0$ pro libovolnou množinu $\hat{\mathcal{I}}_{i-1} \subset \{i-m+1, \ldots, i-1\}$, pro kterou jsou sloupce matice $[\hat{D}_i, d_i]$ lineárně nezávislé. Podle věty je tedy možné vždy volit $\hat{\mathcal{I}}_{i-1} = \{i-m+1, \ldots, i-1\}$.

Věta 193. *Uvažujme $\bar{m}$-krokovou modifikovanou maticovou metodou BFGS s omezenou pamětí, takovou že $x_{i+1} = x_i + \alpha_i s_i$, přičemž $s_1 = -g_1$ a pro $i \geq 1$ platí (785), kde sloupce matic $\bar{D}_i$, $\bar{Y}_i$ se určují podle vzorců*

$$\bar{d}_i = d_i^* = d_i - \sum_{j=i-m+1}^{i-1} \frac{\bar{y}_j^T d_i}{\bar{b}_j} \bar{d}_j, \qquad \bar{y}_i = y_i^* = y_i - \sum_{j=i-m+1}^{i-1} \frac{\bar{d}_j^T y_i}{\bar{b}_j} \bar{y}_j. \tag{814}$$

Aplikujeme-li tuto metodu na ryze konvexní kvadratickou funkci s pozitivně definitní Hessovou maticí G a není-li vektor d_i lineární kombinací vektorů $\bar{d}_j$, $j \in \{i-m+1, \ldots, i-1\}$, jsou pravdivá následující tvrzení.

(a) Vektory $\bar{d}_j$, $i-m+1 \leq j \leq i$ jsou vzájemně G-ortogonální.

(b) Matice (787) splňuje m kvazinewtonovských podmínek $H_{i+1} \bar{y}_j = \bar{d}_j$, $i-m+1 \leq j \leq i$. Navíc je splněna standardní kvazinewtonovská podmínka $H_{i+1} y_i = d_i$.

(c) Jestliže v i-tém iteračním kroku platí $\alpha_i = 1$, je vektor s_{i+1} G-ortogonální k vektorům $\bar{d}_j$, $i-m+1 \leq j \leq i-1$.

Důkaz Jelikož podle věty 191 platí $\bar{b}_i > 0$, jsou splněny předpoklady věty 192, takže podle (806) platí $\bar{d}_k^T \bar{y}_j = \bar{d}_k^T G \bar{d}_j = 0$, $i-m+1 \leq k < j \leq i$. Ze stejného důvodu je splněno m kvazinewtonovských podmínek $H_{i+1} \bar{y}_j = \bar{d}_j$, $i-m+1 \leq j \leq i$ (speciálně $H_{i+1} \hat{Y}_{i-1} = \hat{D}_{i-1}$, neboli $\hat{Y}_{i-1} = B_{i+1} \hat{D}_{i-1}$). Platnost kvazinewtonovské podmínky $H_{i+1} y_i = d_i$ plyne z (808), neboť pro kvadratickou funkci platí $d_i^T \bar{y}_j = d_i^T G \bar{d}_j = y_i^T \bar{d}_j$. Jestliže $\alpha_i = 1$, platí $s_i = d_i$, takže

$$\begin{aligned} -\hat{D}_{i-1}^T G s_{i+1} &= -\hat{Y}_{i-1}^T s_{i+1} = -\hat{D}_{i-1}^T B_{i+1} s_{i+1} = \hat{D}_{i-1}^T g_{i+1} = \hat{D}_{i-1}^T (y_i + g_i) \\ &= \hat{D}_{i-1}^T G d_i + \hat{D}_{i-1}^T g_i = \hat{Y}_{i-1}^T d_i + \hat{Y}_{i-1}^T H_i g_i = \hat{Y}_{i-1}^T (d_i - s_i) = \hat{Y}_{i-1}^T (d_i - d_i) = 0. \end{aligned}$$

□

Transformace popsané v tomto oddílu zlepšují teoretické vlastnosti metod s proměnou metrikou s omezenou pamětí. Nicméně v některých případech mohou zhoršit stabilitu iteračního procesu a také zvyšují počet potřebných numerických operací. Proto je důležité zvolit vhodnou strategii pro výběr množiny $\hat{\mathcal{I}}_{i-1}$. Množina $\hat{\mathcal{I}}_{i-1} \subset \{i-m+1, \ldots, i-1\}$ se určuje rekurentně tak, že se postupně procházejí indexy množiny $\{i-m+1, \ldots, i-1\}$ od největšího k nejmenšímu (neboť nejnovější informace jsou nejdůležitější). Předpokládáme, že pro index $j \in \{i-m+1, \ldots, i-1\}$ známe množinu $\hat{\mathcal{I}}_{i-1}^{j+1} = \hat{\mathcal{I}}_{i-1} \cap \{j+1, \ldots, i-1\}$, kde $\hat{\mathcal{I}}_{i-1}^{i} = \emptyset$, a čísla $\bar{b}_i^{j+1}$, $\bar{a}_i^{j+1}$, $\tilde{c}_i^{j+1}$, kde $\bar{b}_i^i = b_i$, $\bar{a}_i^i = y_i^T H_i^{i+1} y_i$, $\tilde{c}_i^i = d_i^T B_i d_i = -\alpha_i d_i^T g_i$ (poznámka 297 (b)). Pak $\hat{\mathcal{I}}_{i-1} = \hat{\mathcal{I}}_{i-1}^{i-m+1}$. Pravidla pro výběr množiny $\hat{\mathcal{I}}_{i-1}^j$, která mají heuristický charakter a byla získána numerickým experimentováním, jsou shrnuta v následující poznámce, kde $\Delta_i^j = (d_i^T \bar{y}_j - \bar{d}_j^T y_i)^2 / (b_i \bar{b}_j)$,

$$\bar{b}_i^j = \bar{b}_i^{j+1} - \frac{d_i^T \bar{y}_j \bar{d}_j^T y_i}{\bar{b}_j}, \qquad \bar{a}_i^j = \bar{a}_i^{j+1} - \frac{(\bar{d}_j^T y_i)^2}{\bar{b}_j}, \qquad \tilde{c}_i^j = \tilde{c}_i^{j+1} - \frac{(d_i^T \bar{y}_j)^2}{\bar{b}_j} \approx \bar{c}_i^j.$$

Poznámka 296. Index $j \in \{i-m+1,\dots,i-1\}$ nepřidáme do množiny $\hat{\mathcal{I}}_{i-1}^j$ (neboli $j \notin \hat{\mathcal{I}}_{i-1}^j$), pokud

(a) $j \notin \hat{\mathcal{I}}_{i-2} \cup \{i-1\}$ nebo $\|\bar{d}_j\| > \Delta\|d_j\|$ nebo $\|\bar{y}_j\| > \Delta\|y_j\|$,
(b) $\bar{b}_i^j < \delta_1 b_i$ nebo $\bar{a}_i^j < (1/10)\delta_1 b_i$ nebo $\bar{c}_i^j < 10\delta_1 b_i$,
(c) $\Delta_j^i > \delta_2$ nebo $\Delta_j^i > \delta_3 + (b_i^j/\bar{b}_i - 1)^4/2$,
(d) $\Delta_j^i > \delta_3$ a současně $|1 - \bar{a}_i^j/\bar{b}_i^j|(b_i/\bar{b}_i^j - 1) < 1$,
(e) $(d_i^T \bar{y}_j)^2 + (\bar{d}_j^T y_i)^2/(b_i \bar{b}_j) < \delta_4$,

kde $0 < \delta_1 < 1$, $0 < \delta_3 < \delta_2 < 1$, $0 < \delta_4 < 1$ a $\Delta > 1$ (obvykle $\delta_1 = 10^{-4}$, $\delta_2 = 10^{-2}$, $\delta_3 = 10^{-5}$, $\delta_4 = 10^{-10}$ a $\Delta = 10^3$). Pokud $j \notin \hat{\mathcal{I}}_{i-1}^j$, položíme $\bar{b}_i^j = \bar{b}_i^{j+1}$, $\bar{a}_i^j = \bar{a}_i^{j+1}$, $\bar{c}_i^j = \bar{c}_i^{j+1}$ (vrátíme se k předchozím hodnotám).

Abychom ušetřili aritmetické operace, budeme aktualizace matic a vektorů provádět tak, jak je to uvedeno v následující poznámce.

Poznámka 297. Předpokládejme, že známe matice $\bar{D}_{i-1}$, $\bar{Y}_{i-1}$, $\bar{D}_{i-1}^T\bar{Y}_{i-1}$, $\bar{D}_{i-1}^T\bar{Y}_{i-1}$,$\bar{R}_{i-1}$, $\bar{C}_{i-1}$ a vektory $\bar{D}_{i-1}^T g_i$, $\bar{Y}_{i-1}^T g_i$. Pak:

(a) Pokud $m < \bar{m}$ položíme $\tilde{D}_{i-1} = \bar{D}_{i-1}$, $\tilde{Y}_{i-1} = \bar{Y}_{i-1}$, $\tilde{D}_{i-1}^T\tilde{Y}_{i-1} = \bar{D}_{i-1}^T\bar{Y}_{i-1}$, $\tilde{Y}_{i-1}^T\tilde{Y}_{i-1} = \bar{Y}_{i-1}^T\bar{Y}_{i-1}$, $\tilde{R}_{i-1} = \bar{R}_{i-1}$, $\tilde{C}_{i-1} = \bar{C}_{i-1}$, $\tilde{D}_i^T g_i = \bar{D}_i^T g_i$, $\tilde{Y}_i^T g_i = \bar{Y}_i^T g_i$. V opačném případě určíme matice $\tilde{D}_{i-1}$, $\tilde{Y}_{i-1}$, $\tilde{D}_{i-1}^T\tilde{Y}_{i-1}$, $\tilde{Y}_{i-1}^T\tilde{Y}_{i-1}$, $\tilde{R}_{i-1}$, $\tilde{C}_{i-1}$ a vektory $\tilde{D}_{i-1}^T g_i$, $\tilde{Y}_{i-1}^T g_i$ tak že v maticích $\bar{D}_{i-1}$, $\bar{Y}_{i-1}$ vyškrtneme první sloupec (vektor s indexem $i-m$), v maticích $\bar{D}_{i-1}^T\bar{Y}_{i-1}$, $\bar{D}_{i-1}^T\bar{Y}_{i-1}$. $\bar{R}_{i-1}$, $\bar{C}_{i-1}$ vyškrtneme první řádek a první sloupec a z vektorů $\bar{D}_{i-1}^T g_i$, $\bar{Y}_{i-1}^T g_i$ odstraníme první prvek.
(b) Vypočteme vektory $\tilde{D}_{i-1}^T g_{i+1}$, $\tilde{Y}_{i-1}^T g_{i+1}$, $\tilde{D}_{i-1}^T y_i = \tilde{D}_{i-1}^T g_{i+1} - \tilde{D}_{i-1}^T g_i$, $\tilde{Y}_{i-1}^T y_i = \tilde{Y}_{i-1}^T g_{i+1} - \tilde{Y}_{i-1}^T g_i$ a čísla $d_i^T B_i d_i = -\alpha_i d_i^T g_i$,

$$\begin{aligned} y_i^T H_i^{i+1} y_i = \gamma_i y_i^T y_i \;&+\; (\tilde{D}_{i-1}^T y_i)^T (\tilde{R}_{i-1}^T)^{-1}(\tilde{C}_{i-1} + \gamma_i \tilde{Y}_{i-1}^T\tilde{Y}_{i-1})\tilde{R}_{i-1}^{-1}(\tilde{D}_{i-1}^T y_i) \\ &-\; \gamma_i (\tilde{D}_{i-1}^T y_i)^T (\tilde{R}_{i-1}^T)^{-1}(\tilde{Y}_{i-1}^T y_i) \end{aligned}$$

(využíváme toho, že matici H_i^{i+1} lze vyjádřit ve tvaru

$$H_i^{i+1} = (\tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T)^T \bar{C}_i \tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T + \gamma_i (I - \tilde{Y}_{i-1}\tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T)^T (I - \tilde{Y}_{i-1}\tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T),$$

který je analogií vztahu (787) pro matici $H_{i+1} = H_{i+1}^{i+1}$).
(c) Předpokládáme, že známe vektory $\bar{\lambda}_i$, $\bar{\omega}_i$. Vypočteme vektory $\bar{d}_i = d_i + \tilde{D}_{i-1}\bar{\lambda}_i$, $\bar{y}_i = y_i + \tilde{Y}_{i-1}\bar{\omega}_i$ a $\tilde{D}_{i-1}^T\bar{y}_i = \tilde{D}_{i-1}^T y_i + \tilde{D}_{i-1}^T\tilde{Y}_{i-1}\bar{\omega}_i$, $\tilde{Y}_{i-1}^T\bar{y}_i = \tilde{Y}_{i-1}^T y_i + \tilde{Y}_{i-1}^T\tilde{Y}_{i-1}\bar{\omega}_i$, $\tilde{Y}_{i-1}^T\bar{d}_i = \tilde{Y}_{i-1}^T d_i + \tilde{Y}_{i-1}^T\tilde{D}_{i-1}\bar{\lambda}_i$. Vektor $\tilde{Y}_{i-1}^T d_i$ vypočtene podle vzorce

$$\begin{aligned} \tilde{Y}_{i-1}^T d_i = \;&-\; \alpha_i \tilde{Y}_{i-1}^T H_i g_i = -\alpha_i \gamma_{i-1}\tilde{Y}_{i-1}^T g_i \\ &-\; \alpha_i \tilde{Y}_{i-1}^T \tilde{D}_{i-1}(\tilde{R}_{i-1}^{-1})^T \left[(\tilde{C}_{i-1} + \gamma_{i-1}\tilde{Y}_{i-1}^T\tilde{Y}_{i-1})\tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T g_i - \gamma_{i-1}\tilde{Y}_{i-1}^T g_{i+1}\right] \\ &+\; \alpha_i \gamma_{i-1}\tilde{Y}_{i-1}^T\tilde{Y}_{i-1}\tilde{R}_{i-1}^{-1}\tilde{D}_{i-1}^T g_i, \end{aligned}$$

který využívá analogii vztahu (785).
(d) Položíme $\bar{D}_i = [\tilde{D}_{i-1}, \bar{d}_i]$, $\bar{Y}_i = [\tilde{Y}_{i-1}, \bar{y}_i]$, $\bar{D}_i^T g_{i+1} = [\tilde{D}_{i-1}g_{i+1}, \bar{d}_i^T g_{i+1}]$, $\bar{Y}_i^T g_{i+1} = [\tilde{Y}_{i-1}g_{i+1}, \bar{y}_i^T g_{i+1}]$,

$$\bar{D}_i^T\bar{Y}_i = \begin{bmatrix} \bar{D}_{i-1}^T\bar{Y}_{i-1} & \bar{D}_{i-1}^T\bar{y}_i \\ \bar{d}_i^T\bar{Y}_{i-1} & \bar{d}_i^T\bar{y}_i \end{bmatrix}, \qquad \bar{Y}_i^T\bar{Y}_i = \begin{bmatrix} \bar{Y}_{i-1}^T\bar{Y}_{i-1} & \bar{Y}_{i-1}^T\bar{y}_i \\ \bar{y}_i^T\bar{Y}_{i-1} & \bar{y}_i^T\bar{y}_i \end{bmatrix}.$$

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 21. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\delta_1 = 10^{-4}$, $\delta_2 = 10^{-2}$, $\delta_3 = 10^{-5}$, $\delta_4 = 10^{-10}$, $\Delta = 1000$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \le \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$ a určíme směrový vektor s_i: Pokud $m = 0$, položíme $s_i := -g_i$, v opačném případě vypočteme směrový vektor podle vzorce (785) (kde vystupuje i místo $i+1$ a $i-1$ místo i).

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$, $y_i := g_{i+1} - g_i$. Jestliže $m = 0$, položíme $\hat{\mathcal{I}}_{i-1} = \emptyset$, $\bar{d}_i := d_i$, $\bar{d}_i := d_i$, $\bar{b}_i = \bar{d}_i^T \bar{y}_i$, $\bar{D}_i = [\bar{d}_i]$, $\bar{Y}_i = [\bar{y}_i]$, $\bar{D}_i^T \bar{Y}_i = [\bar{d}_i^T \bar{y}_i]$, $\bar{Y}_i^T \bar{Y}_i = [\bar{y}_i^T \bar{y}_i]$, $\bar{R}_i = [\bar{d}_i^T \bar{y}_i]$, $\bar{C}_i = [\bar{d}_i^T \bar{y}_i]$, vypočteme vektory $\bar{D}_i^T g_{i+1}$, $\bar{Y}_i^T g_{i+1}$ a přejdeme na krok 6.

Krok 4 Určíme množinu $\hat{\mathcal{I}}_{i-1}$ a číslo $\bar{b}_i^{m-i+1}$ podle poznámky 296. Určíme vektory $\bar{\lambda}_i$, $\bar{\omega}_i$ tak, že $e_j^T \bar{\lambda}_i = -\bar{d}_i^T \bar{y}_j / \bar{b}_j$, $e_j^T \bar{\omega}_i = -\bar{d}_j^T y_i / \bar{b}_j$ pro $j \in \hat{\mathcal{I}}_{i-1}$ a $e_j^T \bar{\lambda}_i = 0$, $e_j^T \bar{\omega}_i = 0$ pro $j \notin \hat{\mathcal{I}}_{i-1}$. Určíme vektory $\bar{d}_i$, $\bar{y}_i$ podle (786) a položíme $\bar{b}_i = \bar{d}_i^T \bar{y}_i$. Pokud $\bar{b}_i < \bar{b}_i^{m-i+1}/2$, položíme $\bar{b}_i = \bar{b}_i^{m-i+1}$.

Krok 5 Určíme matice $\bar{D}_i$, $\bar{Y}_i$, $\bar{D}_i^T \bar{Y}_i$, $\bar{D}_i^T \bar{Y}_i$, $\bar{R}_i$, $\bar{C}_i$ a vektory $\bar{D}_i^T g_{i1}$, $\bar{Y}_i^T g_{i+1}$ podle poznámky 297.

Krok 6 Zvětšíme i o 1 a přejdeme na krok 2.

Nyní dokážeme globální konvergenci algoritmu 21. Tak jako v oddílu 9.2 budeme předpokládat, že funkce $F : R^n \to R$ vyhovuje předpokladům F1, F4, F5. Pak platí (695).

Věta 194. *Uvažujme metodu s proměnnou metrikou s omezenou pamětí, realizovanou algoritmem 21, s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Nechť funkce F splňuje předpoklady F1, F4, F5. Pak směrové vektory $s_i = -H_i^i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$.*

Důkaz (a) Nechť $i \in N$ a $\hat{\mathcal{I}}_{i-1} \neq \emptyset$. Podle poznámky 296 pro $j \in \hat{\mathcal{I}}_{i-1}$ platí $\bar{b}_j \ge \delta_1 b_j$, $\|\bar{d}_j\| \le \Delta \|d_j\|$, $\|\bar{y}_j\| \le \Delta \|y_j\|$, takže s použitím (695), (810), (811) a důsledku 11 dostaneme

$$\begin{aligned} \|\bar{d}_i\| &\le \left\| \prod_{j\in\hat{\mathcal{I}}_{i-1}} \left(I - \frac{\bar{d}_j \bar{y}_j^T}{\bar{b}_j} \right) \right\| \|d_i\| \le \|d_i\| \prod_{j\in\hat{\mathcal{I}}_{i-1}} \frac{\|\bar{d}_j\| \|\bar{y}_j\|}{\bar{b}_j} \\ &\le \|d_i\| \left(\frac{\Delta^2}{\delta_1} \right)^{\bar{m}} \prod_{j\in\hat{\mathcal{I}}_{i-1}} \frac{\|d_j\| \|y_j\|}{b_j} \le \left(\frac{\Delta^2}{\delta_1} \right)^{\bar{m}} \left(\frac{\overline{G}}{\underline{G}} \right)^{\bar{m}/2} \|d_i\| \triangleq \overline{C} \|d_i\|. \end{aligned}$$

Podobným způsobem dostaneme $\|\bar{y}_i\| \le \overline{C} \|y_i\|$. Pokud $\hat{\mathcal{I}}_{i-1} = \emptyset$, platí $\bar{d}_i = d_i$ a $\bar{y}_i = y_i$, takže opět $\|\bar{d}_i\| \le \overline{C} \|d_i\|$ a $\|\bar{y}_i\| \le \overline{C} \|y_i\|$ (neboť $\overline{C} > 1$). Spojíme-li tyto nerovnosti, můžeme s použitím (695) pro $i \in N$ psát

$$\begin{aligned} \frac{\|\bar{d}_i\|^2}{\bar{b}_i} &\le \frac{\overline{C}^2}{\delta_1} \frac{\|d_i\|^2}{b_i} \le \frac{\overline{C}^2}{\delta_1 \underline{G}} \triangleq \frac{1}{K_1}, \\ \frac{\|\bar{y}_i\|^2}{\bar{b}_i} &\le \frac{\overline{C}^2}{\delta_1} \frac{\|y_i\|^2}{b_i} \le \frac{\overline{C}^2 \overline{G}}{\delta_1} \triangleq C_1. \end{aligned}$$

(b) Podle (a) a lemmatu 75 jsou splněny předpoklady lemmatu 82, takže existují čísla $0 < K < 1 < C$ taková, že $\operatorname{Tr} B_i^i \le C$ a $\det B_i^i \ge K$. Jsou tedy splněny předpoklady lemmatu 76, takže vektory $s_i = -H_i^i g_i$ jsou stejnoměrně spádové a platí $\lim_{i\to\infty} \|g_i\| = 0$. □

Poznámka 298. V důkazu globální konvergence potřebujeme, aby pro $j \in \hat{\mathcal{I}}_{i-1}^j$ platilo $\|\bar{d}_j\| \le \Delta \|d_j\|$, $\|\bar{y}_j\| \le \Delta \|y_j\|$. Tento test vyžaduje dva skalární součiny ($2n$ aritmetických operací) navíc. Proto ho raději vynecháváme a globální konvergenci zajišťujeme podle poznámky 32 (účinnost metody se tím nezhorší a výpočet se urychlí).

9.5 Metody redukovaných Hessiánů s omezenou pamětí

Nechť $m = \max(\bar{m}, i-1)$. Zatímco metody s proměnnou metrikou s omezenou pamětí generují matici H_i vždy z počáteční (obvykle jednotkové) matice pomocí m aktualizací používajících $2m$ vektorů d_j, y_j, $i-m \le j \le i-1$, používají metody redukovaných Hessiánů, studované v práci [64], jednotkovou matici pouze v prvním iteračním kroku. Směrový vektor s_i se počítá pomocí redukované matice $\tilde{H}_i$ řádu m, která se aktualizuje pomocí m vektorů z_k, $1 \le k \le m$, tvořících ortonormální bázi v podprostoru generovaném vektory d_j, $i-m \le j \le i-1$. Tento podprostor se po skončení i-tého iteračního kroku změní obvykle tak, že se nejstarší vektor z_1 odstraní, ostatní vektory se posunou a přidá se nový vektor z_m získaný ortogonalizací vektoru d_i. Abychom lépe pochopili myšlenku metod redukovaných Hessiánů, budeme nejprve předpokládat, že $\bar{m} = n$. V tomto případě jsou metody redukovaných Hessiánů ekvivalentní standardním metodám s proměnnou metrikou.

Věta 195. *Uvažujme metodu s proměnnou metrikou (270)–(272), kde $H_1 = I$. Pak*

$$H_i z \in \mathcal{G}_i, \quad \forall z \in \mathcal{G}_i, \tag{815}$$

$$H_i w = \left(\prod_{k=1}^{i-1} \gamma_k \right) w, \quad \forall w \in \mathcal{G}_i^{\perp}, \tag{816}$$

kde $\mathcal{G}_i = \mathcal{L}(g_1, \dots g_i)$ a $\mathcal{G}_i^{\perp}$ je ortogonální doplněk podprostoru $\mathcal{G}_i$.

Důkaz Důkaz provedeme indukcí. Pro $i = 1$ je $H_1 = I$, takže $H_1 z = z \in \mathcal{G}_1$, $\forall z \in \mathcal{G}_1$, a $H_1 w = w$, $\forall w \in \mathcal{G}_1^{\perp}$. Předpokládejme, že (815)–(816) platí pro nějaký index $i \in N$ a označme $\omega_i = \prod_{k=1}^{i} \gamma_k$.

(a) Nechť $H_i g_{i+1} = z + w$, kde $z \in \mathcal{G}_{i+1}$ a $w \in \mathcal{G}_{i+1}^{\perp} \subset \mathcal{G}_i^{\perp}$ (takže $w^T z = 0$). Pak podle (816) platí

$$w^T w = w^T z + w^T w = w^T H_i g_{i+1} = \omega_{i-1} w^T g_{i+1} = 0,$$

neboť $g_{i+1} \in \mathcal{G}_{i+1}$. Můžeme tedy psát $H_i g_{i+1} = z \in \mathcal{G}_{i+1}$ a jelikož podle indukčního předpokladu platí $H_i g_i \in \mathcal{G}_i \subset \mathcal{G}_{i+1}$, dostaneme $d_i = -\alpha_i H_i g_i \in \mathcal{G}_{i+1}$ a $H_i y_i = H_i(g_{i+1} - g_i) \in \mathcal{G}_{i+1}$, takže $\mathcal{L}(U_i) \subset \mathcal{G}_{i+1}$. To spolu s (271) dává $H_{i+1} g_j = \gamma_i H_i g_j + U_i M_i U_i^T g_j \in \mathcal{G}_{i+1}$, $\forall 1 \le j \le i$, a použitím (272) dostaneme

$$H_{i+1} g_{i+1} = H_{i+1} g_i + H_{i+1} y_i = H_{i+1} g_i + \rho_i d_i \in \mathcal{G}_{i+1}.$$

Tím jsme dokázali, že $H_{i+1} g_j \in \mathcal{G}_{i+1}$ $\forall 1 \le j \le i+1$, neboli $H_{i+1} z \in \mathcal{G}_{i+1}$, $\forall z \in \mathcal{G}_{i+1}$.

(b) Nechť nyní $w \in \mathcal{G}_{i+1}^{\perp} \subset \mathcal{G}_i^{\perp}$. Jelikož podle (a) platí $\mathcal{L}(U_i) \subset \mathcal{G}_{i+1}$, dostaneme $U_i^T w = 0$, což spolu s (271) dává $H_{i+1} w = \gamma_i H_i w$. Protože $w \in \mathcal{G}_i^{\perp}$, můžeme podle indukčního předpokladu psát $H_{i+1} w = \omega_i w$. □

Důsledek 26. *Nechť jsou splněny předpoklady věty 195 a nechť matice H_i, $i \in N$, jsou pozitivně definitní. Pak $\mathcal{S}_i = \mathcal{G}_i$, $i \in N$, kde $\mathcal{S}_i = \mathcal{L}(s_1, \dots, s_i)$.*

Důkaz Jelikož $s_1 = -g_1$, můžeme psát $\mathcal{S}_1 = \mathcal{L}(s_1) = \mathcal{L}(g_1) = \mathcal{G}_1$. Předpokládejme, že tvrzení platí pro nějaký index $i \in N$. Podle věty 195 platí $s_{i+1} = -H_{i+1} g_{i+1} \in \mathcal{G}_{i+1}$, takže $\mathcal{S}_{i+1} \subset \mathcal{G}_{i+1}$. Nechť $g_{i+1} \notin \mathcal{G}_i$ (v opačném případě platí $\mathcal{G}_{i+1} = \mathcal{G}_i$ a není co dokazovat). Ukážeme, že také $H_i g_{i+1} \notin \mathcal{G}_i$. Pokud totiž $H_i g_{i+1} \in \mathcal{G}_i$, musí platit $w^T H_i g_{i+1} = \omega_{i-1} w^T g_{i+1} = 0$, $\forall w \in \mathcal{G}_i^{\perp}$, takže nutně $g_{i+1} \in \mathcal{G}_i$, což je spor s předpokladem že $g_{i+1} \notin \mathcal{G}_i$. Jelikož $H_i g_{i+1} \notin \mathcal{G}_i$ a $H_i g_i \in \mathcal{G}_i$, platí $H_i y_i = H_i(g_{i+1} - g_i) \notin \mathcal{G}_i$. Jelikož vektor s_{i+1} je podle (452) lineární kombinací vektorů d_i a $H_i y_i$, kde $d_i = -\alpha_i H_i g_i \in \mathcal{G}_i$ a koeficient u $H_i y_i$ je nenulový (neboť matice H_{i+1} je podle předpokladu pozitivně definitní), musí platit $s_{i+1} \notin \mathcal{G}_i$, takže $\mathcal{S}_{i+1} = \mathcal{G}_{i+1}$. □

Věta 196. *Nechť jsou splněny předpoklady věty 195 a nechť Z_i je matice, jejíž sloupce tvoří ortonormální bázi v $\mathcal{G}_i$. Pak platí*

$$H_i = Z_i \tilde{H}_i Z_i^T + \left(\prod_{k=1}^{i-1} \gamma_k \right) (I - Z_i Z_i^T), \qquad \tilde{H}_i = Z_i^T H_i Z_i. \tag{817}$$

Důkaz Podle věty 195 platí $H_i Z_i = Z_i \tilde{H}_i$, kde $\tilde{H}_i$ je nějaká čtvercová regulární matice. Vynásobíme-li tuto rovnost maticí Z_i^T, dostaneme $\tilde{H}_i = Z_i^T H_i Z_i$ (neboť $Z_i^T Z_i = I$). Lze tedy psát

$$\left(Z_i \tilde{H}_i Z_i^T + \omega_{i-1}(I - Z_i Z_i^T)\right) Z_i = Z_i \tilde{H}_i = H_i Z_i.$$

Nechť W_i je matice, jejíž sloupce tvoří ortonormální bázi v $\mathcal{G}_i^\perp$, takže $Z_i^T W_i = 0$. Pak podle věty 195 platí

$$\left(Z_i \tilde{H}_i Z_i^T + \omega_{i-1}(I - Z_i Z_i^T)\right) W_i = \omega_{i-1} W_i = H_i W_i.$$

Jelikož čtvercová matice $[Z_i, W_i]$ je ortogonální (a tedy regulární), platí (817). □

Poznámka 299. Jelikož $g_i \in \mathcal{G}_i$, můžeme podle (817) psát $H_i g_i = Z_i \tilde{H}_i Z_i^T g_i$. Směrový vektor $s_i = -H_i g_i$ se tedy vypočte podle vzorců

$$s_i = Z_i \tilde{s}_i, \qquad \tilde{s}_i = -\tilde{H}_i \tilde{g}_i, \qquad \tilde{g}_i = Z_i^T g_i.$$

Poznamenejme, že v těchto vzorcích se používá pouze redukovaná matice $\tilde{H}_i$ a matice Z_i jejíž sloupce tvoří ortonormální bázi v $\mathcal{G}_i$.

Poznámka 300. Předpokládejme, že vektor g_{i+1} neleží v $\mathcal{G}_i$ a označme $P_i = I - Z_i Z_i^T$ (takže $P_i g_{i+1} \neq 0$). Pak vektor $z_{i+1} = P_i g_{i+1} / \|P_i g_{i+1}\|$ leží v $\mathcal{G}_{i+1} \cap \mathcal{G}_i^\perp$ a sloupce matice $Z_{i+1} = [Z_i, z_{i+1}]$ tvoří ortonormální bázi v $\mathcal{G}_{i+1}$. V dalších úvahách budeme používat označení

$$\hat{H}_i = \begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \prod_{k=1}^{i-1} \gamma_k \end{bmatrix}.$$

Věta 197. *Nechť jsou splněny předpoklady věty 195 a vektor g_{i+1} neleží v $\mathcal{G}_i$. Pak platí $\hat{H}_i = Z_{i+1}^T H_i Z_{i+1}$ a*

$$H_i = Z_{i+1} \hat{H}_i Z_{i+1}^T + \left(\prod_{k=1}^{i-1} \gamma_k\right)(I - Z_{i+1} Z_{i+1}^T), \tag{818}$$

takže

$$s_i = Z_{i+1} \hat{s}_i, \qquad \hat{s}_i = -\hat{H}_i \hat{g}_i, \qquad \hat{g}_i = Z_{i+1}^T g_i.$$

Nechť

$$\tilde{H}_{i+1} = \gamma_i(\hat{H}_i + \hat{U}_i M_i \hat{U}_i^T), \qquad \hat{U}_i = [\hat{d}_i, \hat{H}_i \hat{y}_i], \tag{819}$$

kde $\hat{d}_i = \alpha_i \hat{s}_i$ a $\hat{y}_i = Z_{i+1}^T y_i$. Pak platí

$$H_{i+1} = \gamma_i(H_i + U_i M_i U_i^T), \qquad U_i = [d_i, H_i y_i]$$

(vztah (271)). Přitom $a_i = y_i^T H_i y_i = \hat{y}_i^T \hat{H}_i \hat{y}_i$ a $b_i = y_i^T d_i = \hat{y}_i^T \hat{d}_i$.

Důkaz (a) Jelikož $z_{i+1} \in \mathcal{G}_i^\perp$, můžeme podle (817) psát $Z_i^T H_i z_{i+1} = \tilde{H}_i Z_i^T z_{i+1} = 0$ a $z_{i+1}^T H_i z_{i+1} = \omega_{i-1}$, kde $\omega_{i-1} = \prod_{k=1}^{i-1} \gamma_k$ (neboť $z_{i+1}^T z_{i+1} = 1$). Platí tedy $\hat{H}_i = Z_{i+1}^T H_i Z_{i+1}$. Dále podle věty 196 dostaneme

$$\begin{aligned} Z_{i+1} \hat{H}_i Z_{i+1}^T + \omega_{i-1}(I - Z_{i+1} Z_{i+1}^T) &= [Z_i, z_{i+1}] \begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \omega_{i-1} \end{bmatrix} \begin{bmatrix} Z_i^T \\ z_{i+1}^T \end{bmatrix} + \omega_{i-1}(I - Z_i Z_i^T - z_{i+1} z_{i+1}^T) \\ &= Z_i \tilde{H}_i Z_i^T + \omega_{i-1} z_{i+1} z_{i+1}^T + \omega_{i-1}(I - Z_i Z_i^T - z_{i+1} z_{i+1}^T) \\ &= Z_i \tilde{H}_i Z_i^T + \omega_{i-1}(I - Z_i Z_i^T) = H_i \end{aligned}$$

a jelikož $g_i \in \mathcal{G}_i \subset \mathcal{G}_{i+1}$, platí $s_i = -H_i g_i = -Z_{i+1} \hat{H}_i Z_{i+1} g_i$, neboli $\hat{g}_i = Z_{i+1}^T g_i$, $\hat{s}_i = -\hat{H}_i \hat{g}_i$, $s_i = Z_{i+1} \hat{s}_i$. Poznamenejme, že z $g_i \in \mathcal{G}_i$ plyne $z_{i+1}^T g_i = 0$, takže

$$\hat{g}_i = Z_{i+1}^T g_i = \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix}, \quad \hat{s}_i = -\hat{H}_i \hat{g}_i = -\begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \omega_{i-1} \end{bmatrix} \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{s}_i \\ 0 \end{bmatrix}, \quad \hat{d}_i = \alpha_i \hat{s}_i = \begin{bmatrix} \tilde{d}_i \\ 0 \end{bmatrix}. \tag{820}$$

Jelikož $z_{i+1}^T g_i = 0$, můžeme psát

$$z_{i+1}^T y_i = z_{i+1}^T g_{i+1} = \frac{g_{i+1}^T P_i g_{i+1}}{\|P_i g_{i+1}\|} = \|P_i g_{i+1}\|$$

(neboť matice $P_i = I - Z_i Z_i^T$ je idempotemtní), takže

$$\hat{y} = Z_{i+1}^T y_i = \begin{bmatrix} Z_i^T y_i \\ z_{i+1}^T y_i \end{bmatrix} = \begin{bmatrix} \tilde{y}_i \\ \|P_i g_{i+1}\| \end{bmatrix}. \tag{821}$$

Číslo $\|P_i g_{i+1}\|$ známe, neboť ho potřebujeme k určení vektoru z_{i+1} (poznámka 300).

(b) Zřejmě $d_i = Z_{i+1}\hat{d}_i$ a $b_i = y_i^T d_i = y_i^T Z_{i+1}\hat{d}_i = \hat{y}_i^T \hat{d}_i$. Jelikož $y_i = g_{i+1} - g_i \in \mathcal{G}_i$, platí podle (818) $H_i y_i = Z_{i+1}\hat{H}_i\hat{y}_i$ a $a_i = y_i^T H_i y_i = \hat{y}_i^T \hat{H}_i \hat{y}_i$. Můžeme tedy psát $U_i = Z_{i+1}\hat{U}_i$. Použijeme-li (817) a (819), dostaneme

$$\begin{aligned} H_{i+1} &= Z_{i+1}\tilde{H}_{i+1}Z_{i+1}^T + \omega_i(I - Z_{i+1}Z_{i+1}^T) \\ &= \gamma_i\left(Z_{i+1}\hat{H}_i Z_{i+1}^T + Z_{i+1}\hat{U}_i M_i \hat{U}_i^T Z_{i+1}^T + \omega_{i-1}(I - Z_{i+1}Z_{i+1}^T)\right) \\ &= \gamma_i(H_i + U_i M_i U_i^T), \end{aligned}$$

což je právě vztah (271). □

Poznámka 301. Leží-li vektor g_{i+1} v $\mathcal{G}_i$, můžeme položit $Z_{i+1} = Z_i$, $\hat{H}_i = \tilde{H}_i$ a v aktualizaci (819) použít vektory $\hat{d}_i = \tilde{d}_i = \alpha_i \tilde{s}_i$ a $\hat{y}_i = \tilde{y}_i = Z_i^T y_i$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 22. Data $\underline{\delta} > 0$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $Z_1 := [\, g_1/\|g_1\| \,]$, $\tilde{H}_1 := [1]$ a $i := 1$.

Krok 2 Pokud $\|g_i\| \leq \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $\tilde{g}_i := Z_i^T g_i$, $\tilde{s}_i := -\tilde{H}_i \tilde{g}_i$ a $s_i := Z_i \tilde{s}_i$.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $\tilde{d}_i := \alpha_i \tilde{s}_i$, $\tilde{y}_i := Z_i^T(g_{i+1} - g_i)$.

Krok 4 Pokud $\|P_i g_{i+1}\| \leq \underline{\delta}\|g_{i+1}\|$, položíme $Z_{i+1} := Z_i$, $\hat{H}_i := \tilde{H}_i$, $\hat{d}_i := \tilde{d}_i$, $\hat{y}_i := \tilde{y}_i$. Pokud $\|P_i g_{i+1}\| > \underline{\delta}\|g_{i+1}\|$, určíme matice Z_{i+1}, $\hat{H}_i$ podle poznámky 300, vektor $\hat{d}_i$ podle (820) a vektor $\hat{y}_i$ podle (821).

Krok 5 Zvolíme parametry ρ_i, γ_i a η_i (tak jako v Algoritmu 9) a určíme matici $\tilde{H}_{i+1}$ podle (819).

Krok 6 Zvětšíme i o 1 a přejdeme na krok 2.

Algoritmus 22 je teoreticky shodný s algoritmem 9 a měl by být správě uveden v oddílu 4.8 jako modifikace algoritmu 9. Výhoda algoritmu 22 spočívá v tom, že se pro $i < n$ pracuje s menšími maticemi, což vede v tomto případě k úspoře času. Jelikož operace s ortonormálními bázemi vyžadují aritmetické operace navíc, je výhodné přejít pro $i \geq n$ na algoritmus 9. Hlavní přínos metod redukovaných Hessiánů spočívá v tom, že lze jejich myšlenku použít pro rozsáhlé úlohy, omezíme-li řád redukovaných matic na $\bar{m} \ll n$. Dostaneme tak metody redukovaných Hessiáru s omezenou pamětí, studované v práci [65].

Nechť $0 < \bar{m} < n$, $i \in N$ a $m = \min(\bar{m}, i)$. Označme $\mathcal{G}_i = \mathcal{L}(g_{i-m+1}, \ldots, g_i)$,

$$\mathcal{S}_i = \mathcal{L}(s_{i-m+1}, \ldots, s_{i-1}, s_i), \qquad S_i = [s_{i-m+1}, \ldots, s_{i-1}, s_i],$$

$$\mathcal{S}'_i = \mathcal{L}(s_{i-m+1}, \ldots, s_{i-1}, g_i), \qquad S'_i = [s_{i-m+1}, \ldots, s_{i-1}, g_i]$$

a

$$H_i = Z_i \tilde{H}_i Z_i^T + \hat{\gamma}_i (I - Z_i Z_i^T), \qquad \tilde{H}_i = Z_i^T H_i Z_i, \tag{822}$$

kde Z_i je matice, jejíž sloupce tvoří bázi v $\mathcal{S}'_i$ ($\hat{\gamma}_i$ je hodnota specifikovaná v poznámce 305). Metody redukovaných Hessiánů s omezenou pamětí používají směrový vektor $s_i = -H_i g_i$, kde H_i je matice určená vztahem (822). Jelikož $g_i \in \mathcal{S}'_i$, platí $(I - Z_i Z_i^T) g_i = 0$ a směrový vektor lze určit podle vzorců uvedených v poznámce 299. Stačí tedy uchovávat pouze redukovanou matici $\tilde{H}_i$ a matici Z_i, jejíž sloupce tvoří ortonormální bázi v $\mathcal{S}'_i$.

Poznámka 302. Ke konstrukci matice $\tilde{H}_i$ bychom mohli použít podprostor $\mathcal{G}_i$ místo $\mathcal{S}'_i$. Metody používající podprostor $\mathcal{S}'_i$ jsou však účinnější. Je to způsobeno tím, že po snížení dimenze (vyškrtnutí prvního sloupce v matici $[S_i, g_{i+1}]$) již neplatí $\mathcal{S}'_i = \mathcal{G}_i$ a použití podprostoru $\mathcal{S}'_i$ je pro metody redukovaných Hessiánů s omezenou pamětí výhodnější.

Algoritmus metod redukovaných Hessiánů s omezenou pamětí se od algoritmu 22 liší zejména tím, že matice Z_i a $\tilde{H}_i$ mohou mít nanejvýš $\bar{m}$ sloupců. Kromě toho se používají horní trojúhelníkové matice R_i, R'_i takové, že $S_i = Z_i R_i$, $S'_i = Z_i R'_i$. V každém iteračním kroku vycházíme z rozkladu $S'_i = Z_i R'_i$. Po nalezení směrového vektoru určíme rozklad $S_i = Z_i R_i$ podle lemmatu 86 a po vypočtení nového gradientu určíme rozklad $[S_i, g_{i+1}] = Z^+_{i+1} R^+_{i+1}$ podle lemmatu 87. Jestliže $i < \bar{m}$, platí $S_{i+1} = [S_i, g_{i+1}]$, takže $Z_{i+1} = Z^+_{i+1}$ a $R_{i+1} = R^+_{i+1}$. Jestliže $i \geq \bar{m}$, vyškrtneme v $[S_i, g_{i+1}]$ první sloupec a určíme rozklad $S'_{i+1} = Z_{i+1} R_{i+1}$ podle lemmatu 88. Poslední procedura vnáší do algoritmu metod redukovaných Hessiánů s omezenou pamětí nové problémy technického charakteru, které se řeší pomocí Givensových matic elementárních rotací.

Lemma 86. *Nechť $S'_i = Z_i R'_i$. Pak $R'_i = [R_i^-, \tilde{g}_i]$ a položíme-li $R_i = [R_i^-, \tilde{s}_i]$, platí $S_i = Z_i R_i$.*

Důkaz Nechť $S'_i = Z_i R'_i$ a $R'_i = [R_i^-, r'_i]$. Pak $g_i = Z_i r'_i$ a $\tilde{g}_i = Z_i^T g_i = Z_i^T Z_i r'_i = r'_i$ (neboť $Z_i^T Z_i = I$), takže $S'_i = Z_i [R_i^-, \tilde{g}_i]$. Podobně $S_i = Z_i [R_i^-, r_i] = Z_i [R_i^-, \tilde{s}_i]$, neboť $s_i = Z_i \tilde{s}_i$. □

Lemma 87. *Nechť $S_i = Z_i R_i$, kde Z_i, R_i jsou matice vystupující v lemmatu 86. Předpokládejme, že $P_i g_{i+1} = (I - Z_i Z_i^T) g_{i+1} \neq 0$ a položme*

$$Z^+_{i+1} = \left[Z_i, \frac{P_i g_{i+1}}{\|P_i g_{i+1}\|} \right], \qquad R^+_{i+1} = \begin{bmatrix} R_i, & Z_i^T g_{i+1} \\ 0, & \|P_i g_{i+1}\| \end{bmatrix}.$$

Pak platí $[S_i, g_{i+1}] = Z^+_{i+1} R^+_{i+1}$.

Důkaz Roznásobíme-li součin $Z^+_{i+1} R^+_{i+1}$, dostaneme $Z_i R_i = S_i$ a $Z_i Z_i^T g_{i+1} + P_i g_{i+1} = Z_i Z_i^T g_{i+1} + (I - Z_i Z_i^T) g_{i+1} = g_{i+1}$. □

Lemma 88. *Nechť matice Z^+_{i+1}, $R^+_{i+1} = [t_{i+1}, T_{i+1}]$, vystupující v lemmatu 87, mají $\bar{m}+1$ sloupců (takže t_{i+1} je vektor a T_{i+1} je horní Hessenbergova matice, která má $\bar{m}+1$ řádků a $\bar{m}$ sloupců). Nechť Q_{i+1} je ortogonální matice řádu $\bar{m}+1$ taková, že*

$$Q_{i+1}^T T_{i+1} = \begin{bmatrix} R'_{i+1} \\ 0 \end{bmatrix}, \tag{823}$$

kde R'_{i+1} je horní trojúhelníková matice řádu $\bar{m}$. Pak $S'_{i+1} = Z_{i+1} R'_{i+1}$, kde $[S_i, g_{i+1}] = [s_{i-m+1}, S'_{i+1}]$ a Z_{i+1} je matice obsahující prvních $\bar{m}$ sloupců matice $Z^+_{i+1} Q_{i+1}$. Sloupce matice Z_{i+1} tvoří ortonormální bázi v $\mathcal{S}'_{i+1}$

Důkaz Podle lemmatu 87 platí $[S_i, g_{i+1}] = [s_{i-m+1}, S'_{i+1}] = Z^+_{i+1} R^+_{i+1} = Z^+_{i+1} [t_{i+1}, T_{i+1}]$, neboli $S'_{i+1} = Z^+_{i+1} T_{i+1}$, kde T_{i+1} je horní Hessenbergova matice, která má $\bar{m}+1$ řádků a $\bar{m}$ sloupců. Nechť Q_{i+1} je ortogonální matice řádu $\bar{m}+1$, pro kterou platí (823). Pak lze psát

$$S'_{i+1} = Z^+_{i+1} T_{i+1} = Z^+_{i+1} Q_{i+1} Q_{i+1}^T T_{i+1} = Z^+_{i+1} Q_{i+1} \begin{bmatrix} R'_{i+1} \\ 0 \end{bmatrix}.$$

Jelikož $(Z^+_{i+1} Q_{i+1})^T Z^+_{i+1} Q_{i+1} = Q_{i+1}^T Q_{i+1} = I$, tvoří sloupce matice Z_{i+1} ortonormální bázi v $\mathcal{S}'_{i+1}$. □

Poznámka 303. Ortogonální matice Q_{i+1} je obvykle součinem Givensových matic elementárních rotací (poznámka 272). Pak

$$Q_{i+1} = Q_{12}Q_{23}\dots Q_{\bar{m},\bar{m}+1},$$

kde $Q_{j,j+1}$, $1 \le j \le \bar{m}$, jsou matice definované v poznámce 270.

Poznámka 304. Nechť $\tilde{H}_{i+1}^+ = (Z_{i+1}^+)^T H_{i+1} Z_{i+1}^+$ je symetrická matice řádu $\bar{m}+1$ a Q_{i+1} je ortogonální matice řádu $\bar{m}+1$ použitá v lemmatu 88. Pak vyškrtneme-li v matici

$$Q_{i+1}^T \tilde{H}_{i+1}^+ Q_{i+1} = (Z_{i+1}^+ Q_{i+1})^T H_{i+1} Z_{i+1}^+ Q_{i+1}$$

poslední řádek a poslední sloupec, dostaneme matici $\tilde{H}_{i+1} = Z_{i+1}^T H_{i+1} Z_{i+1}$. Z toho plyne, že aplikujeme-li elementární rotace na řádky Hessenbergovy matice T_{i+1}, musíme je též aplikovat na řádky a sloupce redukované matice $\tilde{H}_{i+1}^+$.

Poznámka 305. Snížením dimenze redukované matice (poznámka 304) ztrácíme část informací z předchozích iteračních kroků. Proto není logické používat v matici $\hat{H}_i$ součin škálovacích koeficientů ze všech předchozích iterací. Tak jako u metod s proměnnou metrikou s omezenou pamětí použijeme pouze poslední škálovací koeficient. Položíme

$$\hat{H}_i = \begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \hat{\gamma}_i \end{bmatrix}, \tag{824}$$

kde buď $\hat{\gamma}_i = 1$ nebo $\hat{\gamma}_i = \gamma_i$. Pokud $\hat{\gamma}_i = 1$, použijeme v aktualizaci

$$\tilde{H}_{i+1}^+ = \gamma_i(\hat{H}_i + \hat{U}_i M_i \hat{U}_i^T), \qquad \hat{U}_i = [\hat{d}_i, \hat{H}_i \hat{y}_i], \tag{825}$$

škálovací koeficient γ_i. Pokud $\hat{\gamma}_i = \gamma_i$, položíme v (825) $\gamma_i = 1$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 23. Data $\bar{m} < n$, $\underline{\delta} > 0$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $Z_1 := [\, g_1/\|g_1\| \,]$, $R_1' := [\|g_1\|]$, $\tilde{H}_1 := [1]$ a $i := 1$.

Krok 2 Pokud $\|g_i\| \le \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $\tilde{g}_i := Z_i^T g_i$, $\tilde{s}_i := -\tilde{H}_i \tilde{g}_i$, $s_i := Z_i \tilde{s}_i$ a určíme matici R_i tak jako v lemmatu 86.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $\tilde{d}_i := \alpha_i \tilde{s}_i$, $\tilde{y}_i := Z_i^T(g_{i+1} - g_i)$.

Krok 4 Pokud $\|P_i g_{i+1}\| \le \underline{\delta}\|g_{i+1}\|$, položíme $Z_{i+1}^+ := Z_i$, $R_{i+1}^+ := R_i$, $\hat{H}_i := \tilde{H}_i$, $\hat{d}_i := \tilde{d}_i$, $\hat{y}_i := \tilde{y}_i$. Pokud $\|P_i g_{i+1}\| > \underline{\delta}\|g_{i+1}\|$, určíme matice Z_{i+1}^+, R_{i+1}^+ tak jako v lemmatu 87, matici $\hat{H}_i$ podle (824) a vektory $\hat{d}_i$, $\hat{y}_i$ podle (820), (821).

Krok 5 Zvolíme parametry ρ_i, γ_i a η_i a určíme matici $\tilde{H}_{i+1}^+$ podle (825).

Krok 6 Mají-li matice Z_{i+1}^+, R_{i+1}^+ nejvýše $\bar{m}$ sloupců, položíme $Z_{i+1} := Z_{i+1}^+$, $R_{i+1}' := R_{i+1}^+$ a $\tilde{H}_{i+1} := \tilde{H}_{i+1}^+$. V opačném případě určíme matice Z_{i+1}, R_{i+1}' tak jako v lemmatu 88 a matici $\tilde{H}_{i+1}$ tak jako v poznámce 304.

Krok 7 Zvětšíme i o 1 a přejdeme na krok 2.

Abychom se v algoritmu 23 mohli odvolávat na příslušné vzorce, používáme matice Z, R, H s různými indexy. Ve skutečnosti jsou všechny matice označené stejným písmenem uloženy na stejném místě v paměti počítače, takže příkazy $Z_{i+1}^+ := Z_i$, $R_{i+1}^+ := R_i$, $\hat{H}_i := \tilde{H}_i$ v kroku 4 a příkazy $Z_{i+1} := Z_{i+1}^+$, $R_{i+1}' := R_{i+1}^+$, $\tilde{H}_{i+1} := \tilde{H}_{i+1}^+$ v kroku 6 jsou fiktivní (nic se nepřesouvá).

Algoritmus 23 představuje pouze jeden způsob, jak lze realizovat metody redukovaných Hessiánů s omezenou pamětí. Často se místo matic $\tilde{H}_i$ používají matice $\tilde{B}_i = \tilde{H}_i^{-1}$.

Lemma 89. *Nechť H_i je matice určená vztahem (822) a*

$$B_i = Z_i \tilde{B}_i Z_i^T + \frac{1}{\hat{\gamma}_i}(I - Z_i Z_i^T), \qquad \tilde{B}_i = Z_i^T B_i Z_i. \tag{826}$$

Pak $H_i B_i = I$ právě tehdy, jestliže $\tilde{H}_i \tilde{B}_i = I$

Důkaz Jelikož $Z_i^T Z_i = I$, můžeme psát $H_i B_i = Z_i \tilde{H}_i \tilde{B}_i Z_i^T + I - Z_i Z_i^T$. Jestliže $\tilde{H}_i \tilde{B}_i = I$, dostaneme $H_i B_i = Z_i Z_i^T + I - Z_i Z_i^T = I$. Jestliže $H_i B_i = I$, můžeme psát $I = Z_i(\tilde{H}_i \tilde{B}_i - I) Z_i^T + I$, což po vynásobení Z_i^T zleva a Z_i zprava dává $\tilde{H}_i \tilde{B}_i - I = 0$. □

Poznámka 306. Ukázali jsme, že pokud $H_i = B_i^{-1}$, platí $\tilde{H}_i = \tilde{B}_i^{-1}$ a (822) lze zapsat ve tvaru

$$H_i = Z_i \tilde{B}_i^{-1} Z_i^T + \hat{\gamma}_i(I - Z_i Z_i^T), \qquad \tilde{B}_i = Z_i^T B_i Z_i.$$

V tomto případě se směrový vektor určuje podle vzorců

$$s_i = Z_i \tilde{s}_i, \qquad \tilde{B}_i \tilde{s}_i = -\tilde{g}_i, \qquad \tilde{g}_i = Z_i^T g_i.$$

Místo matice $\tilde{B}_i$ se používá Choleského rozklad $B_i = L_i L_i^T$ kde L_i je dolní trojúhelníková matice, která se aktualizuje metodami popsanými v oddílu 4.7.

Poznámka 307. Používáme-li Choleského rozklad matice $\tilde{B}_i$, zkomplikuje se krok 6 algoritmu 23. Abychom získali rozklad $\tilde{B}_{i+1} = L_{i+1} L_{i+1}$ z rozkladu $\tilde{B}_{i+1}^+ = L_{i+1}^+ L_{i+1}^+$, je třeba matici L_{i+1}^+ vynásobit zleva maticí Q_{i+1}. Tím se ale poruší její tvar (není již dolní trojúhelníková, ale dolní Hessenbergova). Například v matici $Q_{12} L_{i+1}^+$ vznikne nový nenulový prvek v prvním řádku a druhém sloupci. Abychom tento prvek opět vynulovali, musíme matici $Q_{12} L_{i+1}^+$ vynásobit zprava vhodnou Givensovou maticí $\tilde{Q}_{12}$. Pokračujeme-li takto dále, dostaneme dolní trojúhelníkovou matici

$$Q_{i+1}^T L_{i+1}^+ \tilde{Q}_{i+1} = (Q_{12} Q_{23} \dots Q_{\bar{m},\bar{m}+1})^T L_{i+1}^+ \tilde{Q}_{12} \tilde{Q}_{23} \dots \tilde{Q}_{\bar{m},\bar{m}+1},$$

takže $Q_{i+1}^T \tilde{B}_{i+1}^+ Q_{i+1} = Q_{i+1}^T L_{i+1}^+ \tilde{Q}_{i+1} (Q_{i+1}^T L_{i+1}^+ \tilde{Q}_{i+1})^T$ (neboť $\tilde{Q}_{i+1} \tilde{Q}_{i+1}^T = I$). Matici L_{i+1} dostaneme tak, že v matici $Q_{i+1}^T L_{i+1}^+ \tilde{Q}_{i+1}$ vyškrtneme poslední řádek a poslední sloupec.

Algoritmus 23 lze modifikovat tak, abychom ušetřili operace potřebné k úpravám matice Z_{i+1}^+ (jde o ortogonální transformace použité v lemmatu 88). Z tohoto důvodu používáme místo matice Z_i některou z matic $S_i'(R_i')^{-1}$, $S_i(R_i)^{-1}$. V maticích S_i', S_i se pouze vyměňují sloupce, takže jejich úpravy jsou nenáročné, a matice R_i', R_i jsou malé a horní trojúhelníkové. Modifikovaný algoritmus 23 vypadá takto.

(a) Směrový vektor v kroku 2 počítáme postupně podle vzorců $\tilde{u}_i := (S_i')^T g_i$, $(R_i')^T \tilde{g}_i = \tilde{u}_i$, $\tilde{s}_i := -\tilde{H}_i \tilde{g}_i$, $R_i' \tilde{v}_i = \tilde{s}_i$, $s_i := S_i' \tilde{v}_i$.

(b) V kroku 3 vypočteme vektor $\tilde{g}_{i+1} = Z_i^T g_{i+1}$ řešením soustavy rovnic $R_i^T \tilde{g}_{i+1} = S_i^T g_{i+1}$, položíme $\tilde{y}_i = \tilde{g}_{i+1} - \tilde{g}_i$ a spočteme číslo $\|P_i g_{i+1}\| = \sqrt{g_{i+1}^T g_{i+1} - \tilde{g}_{i+1}^T \tilde{g}_{i+1}}$.

Matice Z_i se v modifikovaném algoritmu nepoužívá.

9.6 Posunuté metody s proměnnou metrikou s omezenou pamětí

Směrový vektor získaný metodou redukovaných gradientů lze zapsat ve tvaru

$$s_i = -Z_i \tilde{H}_i Z_i^T g_i = -S_i S_i^T g_i, \qquad S_i = Z_i \tilde{H}_i^{1/2}$$

(matice $S_i \in R^{n \times m}$, kde $m = \max(\bar{m}, i-1)$, má zde stejný význam jako v oddílu 4.2 a je různá od matice S_i použité v oddílu 9.5). Podstatné je, že matice Z_i je zkonstruovaná tak, že $g_i \in \mathcal{L}(S_i) = \mathcal{L}(Z_i)$, takže $S_i^T g_i \neq 0$ (a tudíž $s_i \neq 0$), pokud $g_i \neq 0$. Požadavek, aby platilo $S_i^T g_i \neq 0$, pokud $g_i \neq 0$, ztěžuje použití

obecnějších metod s proměnnou metrikou s omezenou pamětí takových, že $s_i = -S_i S_i^T g_i$, kde matice S_i, která má nanejvýš $\bar{m}$ sloupců, se aktualizuje pomocí metod popsaných v oddílu 4.2. Pro praktické použití je výhodnější předpokládat, že $s_i = -H_i g_i$, kde

$$H_i = \zeta_i I + \bar{H}_i = \zeta_i I + \bar{U}_i \bar{M}_i \bar{U}_i^T,$$

přičemž $\zeta_i > 0$ a matice $\bar{U}_i$ má omezený počet sloupců. V tomto tvaru lze zapsat metody s proměnnou metrikou s omezenou pamětí (vzorec (760)), kdy $\zeta_i = \gamma_{i-1}$ a matice $\bar{U}_i$ má nejvýše $2\bar{m}$ slouců, i metody redukovaných Hessiánů s omezenou pamětí (vzorec (822)), kdy $\zeta_i = \hat{\gamma}_i$, $\bar{U}_i = Z_i$ a $\bar{M}_i = \tilde{H}_i - \hat{\gamma}_i I$. Poznamenejme, že matice $\bar{U}_i \bar{M}_i \bar{U}_i^T$ nemusí být pozitivně semidefinitní.

V tomto oddílu se budeme zabývat metodami popsanými v práci [158], které používají směrový vektor $s_i = -H_i g_i$, kde

$$H_i = \zeta_i I + \bar{H}_i = \zeta_i I + \bar{S}_i \bar{S}_i^T, \tag{827}$$

přičemž číslo $\zeta_i > 0$ a matice $\bar{H}_i = \bar{S}_i \bar{S}_i^T$ se aktualizují tak, aby byla splněna kvazinewtonovská podmínka $H_{i+1} y_i = \rho_i d_i$, $\rho_i > 0$, neboli

$$\bar{H}_{i+1} y_i = \rho_i \bar{d}_i, \qquad \bar{d}_i = d_i - \zeta_{i+1} y_i \tag{828}$$

(v prvním iteračním kroku pokládáme $\zeta_1 = 1$ a $\bar{H}_1 = 0$). Poznamenejme, že matice $\bar{H}_i = \bar{S}_i \bar{S}_i^T$, $\bar{S} \in R^{m \times n}$, je vždy pozitivně semidefinitní.

Abychom lépe porozuměli posunutým metodám s proměnnou metrikou, budeme nejprve předpokládat, že $\bar{m} = n$, takže matice $\bar{S}$ může mít n sloupců a matice $\bar{H}$ může být regulární. Podobným způsobem jako v oddílu 4.1 můžeme odvodit obecnou aktualizaci

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T - \frac{1}{\bar{a}} \bar{H} y (\bar{H} y)^T + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H} y \right) \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H} y \right)^T, \tag{829}$$

kde $\bar{a} = y^T \bar{H} y$ a $\bar{b} = y^T \bar{d}$, která vyhovuje kvazinewtonovské podmínce (828). Jelikož v prvních n iteračních krocích je matice $\bar{H}$ singulární, může nastat případ, že $\bar{H} y = 0$. V tomto případě vynecháme v (829) všechny členy obsahující $\bar{H} y$, takže

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T. \tag{830}$$

Věta 198. *Nechť matice $\bar{H}$ je pozitivně semidefinitní, $\rho > 0$, $\eta \geq 0$ a $0 < \zeta_+ < b/y^T y$. Pak matice $\bar{H}_+$ určená vztahem (829) je pozitivně semidefinitní a matice $H_+ = \zeta_+ + \bar{H}_+$ je pozitivně definitní.*

Důkaz Jelikož $\eta \geq 0$, můžeme použít pseudosoučinový vzorec (291) s $\rho > 0$ a $\gamma = 1$, kam dosadíme hodnoty s pruhem. Podle tohoto vzorce je matice $\bar{H}_+$ pozitivně semidefinitní pokud $\bar{b} > 0$ (hodnotu $\bar{b} = 0$ vylučujeme, neboť $\bar{b}$ se vyskytuje ve jmenovateli pseudosoučinového vzorce). Stačí tedy ověřit podmínku $\bar{b} = b - \zeta_+ y^T y > 0$, což dává $\zeta_+ < b/y^T y$. Jelikož $\zeta_+ > 0$ a matice $\bar{H}_+$ je pozitivně semidefinitní, platí $v^T H_+ v = \zeta_+ v^T v + v^T \bar{H}_+ v > 0$ pro libovolný vektor $v \neq 0$.

□

U posunutých metod s proměnnou metrikou velmi záleží na hodnotě parametru ζ_{i+1}. Abychom vyhověli předpokladům věty 198 položíme $\zeta_{i+1} = \sigma_i b_i / y_i^T y_i$, kde $0 < \sigma_i < 1$. Je-li hodnota σ_i příliš malá, má matice H_{i+1} malé nejmenší vlastní číslo, takže může platit $\|s_i\| \approx 0$ i když $\|g_i\| > 0$. Je-li hodnota σ_i příliš velká, norma natice $\bar{S}_{i+1}$ obvykle exponenciálně narůstá a její sloupce se stávají lineárně závislými. Volíme-li hodnotu σ_i konstantní, je vhodné, aby ležela v intervalu $0.20 \leq \sigma_i \leq 0.25$ (například $\sigma_i = 0.22$). Teoreticky podloženější hodnotu parametru σ_i lze získat pomocí následující věty.

Věta 199. *Nechť $\bar{V} = I - y\,\bar{d}^T / y^T \bar{d}$, kde $\bar{d} = d - \sigma (y^T d / y^T y) y$, a nechť v je libovolný vektor takový, že $y^T v = y^T d$. Pak platí*

$$\left| \frac{\|\bar{V}^T v\|}{\|v\|} - \omega \right| \leq (1 + \omega) \frac{\|v - d\|}{\|v\|},$$

kde

$$\omega = \frac{\sigma}{1 - \sigma} \sqrt{1 - \tau}, \qquad \tau = \frac{(y^T d)^2}{y^T y d^T d}.$$

Důkaz Podle předpokladu platí $y^T\bar{d} = y^T(d - \sigma(y^Td/y^Ty)y) = (1-\sigma)y^Td$ a

$$\bar{V}^Tv = v - \frac{y^Tv}{y^T\bar{d}}\bar{d} = v - \frac{y^Td}{y^T\bar{d}}\bar{d} = v - \frac{1}{1-\sigma}\bar{d} = v - d - \frac{\sigma}{1-\sigma}\left(d - \frac{y^Td}{y^Ty}y\right)$$

Protože

$$\left\|d - \frac{y^Td}{y^Ty}y\right\| = \sqrt{\left(d - \frac{y^Td}{y^Ty}y\right)^T\left(d - \frac{y^Td}{y^Ty}y\right)} = \|d\|\sqrt{1-\tau},$$

můžeme psát $|\,\|\bar{V}^Tv\| - \omega\|d\|\,| \le \|v-d\|$, takže

$$\left|\frac{\|\bar{V}^Tv\|}{\|v\|} - \omega\right| \le \left|\frac{\|\bar{V}^Tv\|}{\|v\|} - \omega\frac{\|d\|}{\|v\|}\right| + \omega\frac{\|v-d\|}{\|v\|} \le (1+\omega)\frac{\|v-d\|}{\|v\|}.$$

□

Jak již bylo konstatováno, je-li hodnota σ příliš velká, norma natice $\bar{S}$ obvykle exponenciálně narůstá a její sloupce se stávají lineárně závislými (dokládají to numerické experimenty). V tomto případě je například první sloupec matice S, který označíme symbolem v, téměř rovnoběžný s vektorem d a vhodnou normalizací lze docílit toho, že $y^Tv = y^Td$ a číslo $\|v-d\|/\|v\|$ je malé. Pak podle věty 199 platí $\|\bar{V}^Tv\|/\|v\| \approx \omega$. Jelikož vektor $\bar{V}^Tv$ je podle (839) prvním sloupcem matice $\bar{S}_+$, je vhodné volit číslo σ tak, aby platilo $\omega \le 1$, což dává $\sigma \le 1/(1+\sqrt{1-\tau})$. Většinou je nutné tuto hodnotu ještě zmenšit. Ukázalo se, že je vhodné vynásobit ji číslem $\sqrt{1-\bar{a}/a}$, takže

$$\sigma = \frac{\sqrt{1-\bar{a}/a}}{1+\sqrt{1-b^2/(y^Tyd^Td)}}. \tag{831}$$

Volba (831) vyhovuje předpokladům věty 203 a odůvodňuje ji také následující věta.

Věta 200. *Nechť $\bar{H}_+$ je matice určená podle vzorce (829), kde $\bar{H} = 0$. Pak matice $H_+ = \zeta_+I + \bar{H}_+$, kde $\zeta_+ = \sigma\bar{b}/y^Ty$ a číslo σ je určeno podle (831), je optimálně podmíněná.*

Důkaz Pokud $\bar{H} = 0$, platí $\bar{H}y = 0$ a $\bar{a} = 0$ a použitím (830) dostaneme $H_+ = \zeta_+I + (\rho/\bar{b})\bar{d}\bar{d}^T$, takže matice $(1/\zeta_+)H_+$ má podle lemmatu 31 $n-1$ jednotkových vlastních čísel a zbylé vlastní číslo, které se rovná číslu podmíněnosti, je dáno vztahem $\kappa_+ = 1 + (1/\zeta_+)(\rho/\bar{b})\bar{d}^T\bar{d}$. Použijeme-li vztahy $\zeta_+ = \sigma b/y^Ty$, $\bar{b} = (1-\sigma)b$ a $\bar{d} = d - \sigma by/y^Ty$, dostaneme pro číslo podmíněnosti matice $(1/\zeta_+)H_+$ (a tedy i H_+) rovnost

$$\begin{aligned}\frac{1}{\rho}(\kappa_+ - 1) = \frac{1}{\zeta_+\bar{b}}\bar{d}^T\bar{d} &= \frac{y^Ty}{\sigma(1-\sigma)b^2}\left(d^Td - 2\sigma\frac{b^2}{y^Ty} + \sigma^2\frac{b^2}{y^Ty}\right)\\ &= \frac{1}{\sigma(1-\sigma)}\left(\frac{y^Tyd^Td}{b^2} - 2\sigma + \sigma^2\right) = \frac{1}{1-\sigma}\left(\frac{1}{\sigma\tau} - 1\right) - 1,\end{aligned}$$

takže

$$\frac{1}{\rho}\kappa'_+ = \frac{1}{(1-\sigma)^2}\left(\frac{1}{\sigma\tau} - 1\right) - \frac{1}{1-\sigma}\left(\frac{1}{\sigma^2\tau}\right) = -\frac{\sigma^2\tau - 2\sigma + 1}{\sigma^2(1-\sigma)^2\tau}.$$

Optimální hodnota parametru σ tedy vyhovuje kvadratické rovnici $\sigma^2\tau - 2\sigma + 1 = 0$, která má řešení $\sigma = (1-\sqrt{1-\tau})/\tau = 1/(1+\sqrt{1-\tau})$ (bereme kořen, pro který platí $0 < \sigma < 1$), a protože $\bar{a} = 0$, dostaneme (831). □

V kvazinewtonovské podmínce (828) používáme parametr $\rho > 0$. Tento parametr má poněkud jiný význam, než v případě standardních metod s proměnnou metrikou. Jeho význam ukazuje následující věta.

Věta 201. *Nechť $H_+ = \zeta_+I + \bar{H}_+$, kde $\zeta_+ = \sigma y^Td/y^Ty$ a $\bar{H}_+$ je matice získaná aktualizací (829) (kde $\bar{d} = d - \zeta_+y$). Pak jestliže $\rho = \sigma/(1-\sigma)$, platí*

$$\frac{y^T\bar{H}_+y}{y^Ty} = \zeta_+.$$

Důkaz Z $\bar{H}_+ y = \rho\bar{d}$ plyne $H_+ y = (1-\rho)\zeta_+ y + \rho d$, takže $y^T H_+ y = (1-\rho)\zeta_+ y^T y + \rho b = \sigma b + (1-\sigma)\rho b$. Položíme-li $\rho = \sigma/(1-\sigma)$, platí $y^T H_+ y = 2\sigma b = 2\zeta_+ y^T y$. Ze vztahu

$$\frac{y^T H_+ y}{y^T y} = \zeta_+ + \frac{y^T \bar{H}_+ y}{y^T y} = 2\zeta_+$$

pak plyne tvrzení věty. □

Poznámka 308. Z věty 201 plyne, že pokud $\rho = \sigma/(\sigma-1)$, jsou oba členy v (827) v jistém smyslu souměřitelné. Jiná vhodná hodnota parametru ρ je $\rho = \zeta/(\zeta+\zeta_+)$.

Nyní se budeme zabývat globální konvergencí posunutých metod s proměnnou metrikou. Tak jako v oddílu 4.5 budeme předpokládat že funkce $F : R^n \to R$ vyhovuje předpokladům F1, F4, F5. Pak (podobně jako v důkazu lemmatu 48) dostaneme

$$\underline{G} \le \frac{y^T d}{d^T d} \le \overline{G}, \qquad \underline{G} \le \frac{y^T y}{y^T d} \le \overline{G}, \qquad \frac{\underline{\sigma}}{\overline{G}} \le \frac{\sigma}{\overline{G}} \le \zeta_+ \le \frac{\sigma}{\underline{G}} \le \frac{\overline{\sigma}}{\underline{G}} \tag{832}$$

(pokud $0 < \underline{\sigma} \le \sigma \le \overline{\sigma} < 1$), neboť $\zeta_+ = \sigma y^T d/y^T y$. Dále budeme předpokládat, že posunutá metoda s proměnnou metrikou je realizována jako metoda spádových směrů (s výběrem délky kroku splňujícím slabou Wolfeho podmínku). Nejprve dokážeme globální konvergenci metody DFP.

Lemma 90. *Uvažujme posunutou metodu s proměnnou metrikou s aktualizací (829), kde $0 < \underline{\rho} \le \rho \le \overline{\rho}$ a $0 < \underline{\sigma} \le \sigma \le \overline{\sigma} < 1$, s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Nechť funkce F splňuje předpoklady F1, F4, F5. Pak, existuje-li konstanta $C > 0$ taková, že*

$$\mathrm{Tr}\bar{H}_{i+1} \le \mathrm{Tr}\bar{H}_i + C \qquad \forall i \in N, \tag{833}$$

platí

$$\liminf_{i\to\infty} \|g_i\| = 0. \tag{834}$$

Důkaz Platí-li (833), můžeme s použitím (832) psát

$$\|H_{i+1}\| \le \zeta_{i+1} + \|\bar{H}_{i+1}\| \le \frac{\overline{\sigma}}{\underline{G}} + \mathrm{Tr}\bar{H}_i + C \le \frac{\overline{\sigma}}{\underline{G}} + \mathrm{Tr}\bar{H}_1 + Ci \le \overline{C}(i+1),$$

kde $\overline{C} = \max(C, \overline{\sigma}/\underline{G} + \mathrm{Tr}\bar{H}_1)$. Dostaneme tedy

$$\cos^2\theta_i = \frac{(s_i^T g_i)^2}{g_i^T g_i s_i^T s_i} = \frac{g_i^T(\zeta_i I + \bar{H}_i) g_i}{g_i^T g_i} \frac{s_i^T H_i^{-1} s_i}{s_i^T s_i} \ge \frac{\zeta_i}{\|H_i\|} \ge \frac{\underline{\sigma}}{\overline{C}\,\overline{G} i},$$

takže platí $\sum_{i=1}^{\infty} \cos^2\theta_i = \infty$, z čehož podle věty 11 plyne (834). □

Věta 202. *Nechť jsou splněny předpoklady lemmatu 90 a $\eta = 0$. Pak platí (834).*

Důkaz Jelikož pro libovolnou symetrickou matici $\bar{H}$ a pro libovolné dva vektory u, v platí

$$\mathrm{Tr}(\bar{H} + uu^T - vv^T) = \mathrm{Tr}\bar{H} + u^T u - v^T v \le \mathrm{Tr}\bar{H} + u^T u,$$

dostaneme použitím (287) nerovnost

$$\begin{aligned} \mathrm{Tr}\bar{H}_+ &\le \mathrm{Tr}\bar{H} + \rho\frac{\bar{d}^T\bar{d}}{y^T\bar{d}} = \mathrm{Tr}\bar{H} + \rho\frac{d^T d + \sigma(\sigma-2)(y^T d)^2/y^T y}{(1-\sigma)y^T d} \\ &\le \mathrm{Tr}\bar{H} + \frac{\overline{\rho}}{1-\overline{\sigma}}\left(\frac{d^T d}{y^T d} + \frac{y^T d}{y^T y}\right) \le \mathrm{Tr}\bar{H} + \frac{\overline{\rho}}{1-\overline{\sigma}}\frac{2}{\underline{G}}, \end{aligned}$$

takže podle lemmatu 90 platí (834). □

Poznámka 309. Nerovnost (833) obecně neplatí. Její platnost zaručuje omezenost výrazu $(\bar{a}/\bar{b})\eta/(1-\eta)$. Jelikož podíl $\bar{a}/\bar{b}$ může teoreticky růst nade všechny meze, nemusí býpodmínka (833) splněna pro žádnou kladnou hodnotu parametru η.

Nyní vyšetříme obecný případ, kdy $0 \leq \eta \leq 1$.

Lemma 91. *Uvažujme posunutou metodu s proměnnou metrikou s aktualizací (829), kde* $0 \leq \eta \leq 1$. *Pak platí*

$$\frac{\det H_+}{\det H} \leq \frac{\bar{d}^T B \bar{d}}{\bar{b}} \left(\rho + \zeta \frac{y^T y}{\bar{b}}\right) \left(1 + \frac{\zeta_+}{\zeta}\right)^n. \tag{835}$$

Důkaz (a) Z vyjádření (293) a z důsledku 8 (c) plyne, že nerovnost (835) stačí dokázat pro $\eta = 1$, tedy pro metodu BFGS. Aktualizaci metody BFGS lze zapsat ve tvaru (296) (kde vystupují veličiny s pruhem a kde $\gamma = 1$), takže položíme-li $\omega = \rho + \bar{a}/\bar{b}$, můžeme psát

$$H^{-1/2}(\zeta I + \bar{H}_+)H^{-1/2} = I + \frac{B^{1/2}(\omega\bar{d} - \bar{H}y)(\omega\bar{d} - \bar{H}y)^T B^{1/2} - B^{1/2}\bar{H}y(\bar{H}y)^T B^{1/2}}{\omega\bar{b}}.$$

Označíme-li $U = [u - v, v]$ a $M = \mathrm{diag}(1, -1)$ a použijeme-li důsledek 9 (d), dostaneme

$$\det(I + UMU^T) = \det(I + MU^T U) = (1 + \|u-v\|^2)(1 - \|v\|^2) + ((u-v)^T v)^2 = u^T u + (1 - u^T v)^2 - u^T u v^T v,$$

což pro $u = \omega B^{1/2}\bar{d}/\sqrt{\omega\bar{b}}$ a $v = B^{1/2}\bar{H}y/\sqrt{\omega\bar{b}}$ dává

$$\frac{\det(\zeta I + \bar{H}_+)}{\det H} = \omega\frac{\bar{d}^T B\bar{d}}{\bar{b}} + \left(1 - \frac{\bar{d}^T B\bar{H}y}{\bar{b}}\right)^2 - \frac{\bar{d}^T B\,\bar{d}y^T\bar{H}B\bar{H}y}{\bar{b}^2}.$$

Jelikož podle (827) platí $\bar{H}y = Hy - \zeta y$, můžeme psát $\bar{d}^T B\bar{H}y = \bar{b} - \zeta\bar{d}^T By$ a $y^T\bar{H}B\bar{H}y = \bar{a} - \zeta y^T y + \zeta^2 y^T By$. Dosadíme-li tyto hodnoty do předchozí rovnosti a použijeme-li Schwarzovu nerovnost, dostaneme

$$\begin{aligned}\frac{\det(\zeta I + \bar{H}_+)}{\det H} &= \left(\rho + \frac{\bar{a}}{\bar{b}}\right)\frac{\bar{d}^T B\bar{d}}{\bar{b}} + \zeta^2\frac{(\bar{d}By)^2}{\bar{b}^2} - \frac{\bar{d}^T B\bar{d}\,y^T\bar{H}B\bar{H}y}{\bar{b}^2} \\ &= \left(\rho + \zeta\frac{y^T y}{\bar{b}}\right)\frac{\bar{d}^T B\bar{d}}{\bar{b}} + \zeta^2\frac{(\bar{d}^T By)^2 - \bar{d}^T B\bar{d}\,y^T By}{\bar{b}^2} \leq \left(\rho + \zeta\frac{y^T y}{\bar{b}}\right)\frac{\bar{d}^T B\bar{d}}{\bar{b}}.\end{aligned}$$

(b) Označme λ_i $1 \leq i \leq n$, vlastní čísla matice $\zeta I + \bar{H}_+$, takže $\lambda_i \geq \zeta$, $1 \leq i \leq n$ (matice $\bar{H}_+$ je pozitivně semidefinitní). Jelikož $H_+ = \zeta_+ + \bar{H}_+$, má matice H_+ vlastní čísla $\lambda_i + \zeta_+ - \zeta$, $1 \leq i \leq n$, takže

$$\frac{\det H_+}{\det(\zeta I + \bar{H}_+)} = \prod_{i=1}^n \left(1 + \frac{\zeta_+ - \zeta}{\lambda_i}\right) \leq \prod_{i=1}^n \left(1 + \frac{\zeta_+}{\lambda_i}\right) \leq \left(1 + \frac{\zeta_+}{\zeta}\right)^n.$$

Spojíme-li tuto nerovnost s nerovností odvozenou v (a), dostaneme tvrzení lemmatu. □

Důkaz globální konvergence lze snadno dokončit, je-li splněna podmínka $\sigma \leq \sqrt{1 - \bar{a}/a}$. Jelikož zároveň požadujeme, aby platilo $\sigma \geq \underline{\sigma}$, a výraz $1 - \bar{a}/a$ může být libovolně malý, je třeba volit nějaký kompromis. Budeme tedy předpokládat, že $\underline{\sigma} \leq \sigma \leq \sqrt{1 - \bar{a}/a}$, pokud $\underline{\sigma} \leq \sqrt{1 - \bar{a}/a}$, a $\sigma = \underline{\sigma}$ v opačném případě.

Lemma 92. *Uvažujme posunutou metodu s proměnnou metrikou s výběrem délky kroku splňujícím slabou Wolfeho podmínku a s aktualizací (829), kde* $0 \leq \eta \leq 1$, $0 < \underline{\rho} \leq \rho \leq \overline{\rho}$ *a buď* $\underline{\sigma} \leq \sigma \leq \sqrt{1 - \bar{a}/a}$ *nebo* $\sqrt{1 - \bar{a}/a} < \underline{\sigma} = \sigma$. *Nechť funkce F splňuje předpoklady F1, F4, F5. Pak existují konstanty* C_1, C_2 *takové, že*

$$\frac{\det H_+}{\det H} \leq C_1\frac{c}{b} + C_2\underline{\sigma},$$

přičemž konstanta C_2 *nezávisí na* $\underline{\sigma}$.

Důkaz (a) Předpokládejme nejprve, že $\underline{\sigma} \le \sigma \le \sqrt{1-\bar{a}/a}$. Pak, jelikož platí $ac - b^2 \ge 0$ (Schvarzova nerovnost), dostaneme $\sigma^2 \le 1 - \bar{a}/a = \zeta y^T y/a \le \zeta y^T yc/b^2$, takže $\zeta_+^2 = \sigma^2 b^2/(y^T y)^2 \le \zeta c/y^T y$. Protože nejmenší vlastní číslo matice H je zdola omezeno číslem ζ, je největší vlastní číslo matice $B = H^{-1}$ shora omezeno číslem $1/\zeta$, takže $\zeta_+^2 y^T By \le \zeta c y^T By/y^T y \le c$. Jelikož pro libovolné dva vektory $u \in R^n$ a $v \in R^n$ platí $(u+v)^T(u+v) \le (\|u\| + \|v\|)^2$ a pro libovolná dvě čísla $a > 0$, $b > 0$ platí $(a+b)^2 \le 2(a^2+b^2)$, můžeme psát

$$\bar{d}^T B\bar{d} = (d - \zeta_+ y)^T B(d - \zeta_+ y) \le (\sqrt{d^T Bd} + \zeta_+\sqrt{y^T By})^2 \le 2(d^T Bd + \zeta_+^2 y^T By), \tag{836}$$

což spolu s $d^T Bd = c$ a $\zeta_+^2 y^T By \le c$ dává $\bar{d}^T B\bar{d} \le 4c$. Zbylé činitele ve výrazu (835) již odhadneme snadno. Podle (832) platí

$$\rho + \zeta\frac{y^T y}{\bar{b}} \le \overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}} \tag{837}$$

(neboť $\bar{b} = (1-\sigma)b \ge (1-\overline{\sigma})b$) a z (828) plyne $1 + \zeta_+/\zeta \le 1 + \overline{\sigma}\,\overline{G}/(\underline{\sigma}\,\underline{G})$. Po dosazení dostaneme

$$\frac{\det H_+}{\det H} \le \frac{4}{1-\overline{\sigma}}\left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}}\right)\left(1 + \frac{\overline{\sigma}\,\overline{G}}{\underline{\sigma}\,\underline{G}}\right)^n \frac{c}{\bar{b}}.$$

(b) Nechť nyní $\sigma = \underline{\sigma} > \sqrt{1-\bar{a}/a}$. Pak podle (832) platí

$$\frac{\zeta_+}{\zeta} \le \frac{\sigma}{\underline{G}}\frac{\overline{G}}{\underline{\sigma}} = \frac{\overline{G}}{\underline{G}}, \tag{838}$$

což s použitím (836) dává

$$\bar{d}^T B\bar{d} \le 2(d^T Bd + \zeta_+^2 y^T By) \le 2(c + \frac{\zeta_+^2}{\zeta} y^T y) = 2c + 2\sigma b\frac{\overline{G}}{\underline{G}} = 2c + 2\underline{\sigma} b\frac{\overline{G}}{\underline{G}}$$

(neboť $y^T By/y^T y \le \|B\| \le 1/\zeta$) a použijeme-li odhady (837), (838), můžeme psát

$$\frac{\det H_+}{\det H} \le \frac{2}{1-\overline{\sigma}}\left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}}\right)\left(1 + \frac{\overline{G}}{\underline{G}}\right)^n \left(\frac{c}{\bar{b}} + \underline{\sigma}\frac{\overline{G}}{\underline{G}}\right).$$

(c) Položíme-li

$$C_1 = \frac{4}{1-\overline{\sigma}}\left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}}\right)\left(1 + \frac{\overline{\sigma}\,\overline{G}}{\underline{\sigma}\,\underline{G}}\right)^n, \qquad C_2 = \frac{2}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}}\left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}}\frac{\overline{G}}{\underline{G}}\right)\left(1 + \frac{\overline{G}}{\underline{G}}\right)^n,$$

dostaneme tvrzení lemmatu. □

Věta 203. *Nechť jsou splněny předpoklady lemmatu 92. Pak je-li číslo $\underline{\sigma} > 0$ dostatečně malé (je-li $\underline{\sigma} < 1/C^2$), platí (834).*

Důkaz (a) Abychom nemuseli vyšetřovat první iterační krok zvlášť, začneme od druhého kroku. Protože v každém iteračním kroku platí $\det H \ge \zeta^n \ge (\underline{\sigma}/\overline{G})^n$, můžeme pro $k \in N$ psát

$$C \triangleq \frac{\underline{\sigma}^n}{\overline{G}^n \det H_2} \le \frac{\det H_{k+2}}{\det H_2} = \prod_{i=2}^{k+1}\frac{\det H_{i+1}}{\det H_i} \le \left(\frac{1}{k}\sum_{i=2}^{k+1}\frac{\det H_{i+1}}{\det H_i}\right)^k$$

(používáme nerovnost (19)). Podle lemmatu 92 tedy platí

$$kC^{1/k} \le \sum_{i=2}^{k+1}\frac{\det H_{i+1}}{\det H_i} \le C_1\sum_{i=2}^{k+1}\frac{c_i}{b_i} + kC_2\underline{\sigma}.$$

Předpokládejme, že $\underline{\sigma} < 1/C_2$. Jelikož $C^{1/k} \to 1$ pro $k \to \infty$, existuje index $\underline{k} \in N$ takový, že $C^{1/k} \geq (1 + C_2\underline{\sigma})/2$, $\forall k \geq \underline{k}$, takže podle předchozí nerovnosti platí

$$\sum_{i=2}^{k+1} \frac{c_i}{b_i} \geq \frac{k}{C_1}(C^{1/k} - C_2\underline{\sigma}) \geq \frac{k}{2C_1}(1 - C_2\underline{\sigma})$$

$\forall k \geq \underline{k}$, takže $\sum_{i=2}^{k+1} c_i/b_i \to \infty$, pro $k \to \infty$.

(b) Použijeme-li (23) a (832), můžeme psát

$$\sum_{i=2}^{k+1} \cos^2\theta_i = \sum_{i=2}^{k+1} \frac{(s_i^T g_i)^2}{g_i^T g_i s_i^T s_i} = \sum_{i=2}^{k+1} \frac{g_i^T H_i g_i}{g_i^T g_i} \frac{d_i^T B_i d_i}{d_i^T d_i} = \sum_{i=2}^{k+1} \frac{g_i^T H_i g_i}{g_i^T g_i} \frac{d_i^T B_i d_i}{y_i^T d_i} \frac{y_i^T d_i}{d_i^T d_i} \geq \underline{\sigma}\frac{\underline{G}}{\overline{G}} \sum_{i=2}^{k+1} \frac{c_i}{b_i}.$$

Jelikož pravá strana konverguje podle (a) k nekonečnu, musí i levá strana konvergovat k nekonečnu, takže podle věty 11 platí (834). □

Posunuté metody s proměnnou metrikou s $\bar{m} = n$ by měly být správě uvedeny v oddílu 4.8 jako modifikace klasických metod s proměnnou metrikou. Posunuté metody dávají lepší výsledky než neškálované klasické metody. Jelikož však není známo, jak posunuté metody vhodně škálovat, jsou řízeně škálované klasické metody mnohem účinnější. Hlavní přínos posunutých metod s proměnnou metrikou spočívá v tom, že lze jejich myšlenku použít pro rozsáhlé úlohy, omezíme-li hodnost matic na $\bar{m} \ll n$.

Nechť nyní $m = \min(\bar{m}, i-1)$, kde $i \in N$ a $\bar{m} < n$. V tomto případě mají matice $\bar{H}_i = \bar{S}_i\bar{S}_i^T$, $\bar{S}_i \in R^{n\times m}$, $i \in N$, omezenou hodnost a odpovídající posunuté metody s proměnnou metrikou jsou metodami s omezenou pamětí. Iterační proces posunutých metod s proměnnou metrikou s omezenou pamětí má dvě fáze. V počáteční fázi pokládáme $\zeta_1 = 1$ a $\bar{H}_1 = 0$ (takže matice $\bar{S}_1$ v (827) neobsahuje žádný sloupec) a pro $1 \leq i \leq \bar{m}$ používáme aktualizaci BFGS

$$\bar{H}_{i+1} = \bar{V}_i^T \bar{H}_i \bar{V}_i + \frac{\rho_i}{\bar{b}_i} \bar{d}_i \bar{d}_i^T, \qquad \bar{V}_i = I - \frac{1}{\bar{b}_i} y_i \bar{d}_i^T,$$

neboli

$$\bar{S}_{i+1} = \left[\bar{V}_i^T \bar{S}_i, \sqrt{\frac{\rho_i}{\bar{b}_i}}\bar{d}_i\right]. \tag{839}$$

V tomto případě se hodnost matice $\bar{H}_i$ i počet sloupců matice $\bar{S}_i$ zvětšují o jednotku až do hodnoty $\bar{m}$. Poznamenejme, že použití obecnějšího pseudosoučinového vzorce (291) (ve kterém vystupují veličiny s pruhem) je znesnadněno tím, že může platit $\bar{H}_i y_i = 0$ a $\bar{a}_i = 0$. V tomto případě vždy používáme metodu BFGS (vyplývá to ze vztahu (292), kde vynechání členu s Hy má stejný důsledek jako volba $\eta = 1$). Ve druhé fázi se hodnost matice $\bar{H}_i$ ani počet sloupců matice $\bar{S}_i$ nemění a matice $\bar{S}_{i+1}$ se získává variačně odvozenými aktualizacemi, které se podobají aktualizacím popsaným v oddílu 4.3.

Ve druhé fázi výpočtu předpokládáme, že $H = \zeta I + \bar{H} = \zeta I + \bar{S}\bar{S}^T$ a hledáme matici $H_+ = \zeta_+ I + \bar{S}_+\bar{S}_+^T$ tak, aby byla splněna kvazinewtonovská podmínka

$$\bar{S}_+^T y = \tilde{z}, \qquad \bar{S}_+\tilde{z} = \rho\bar{d}, \qquad \tilde{z}^T\tilde{z} = \rho\bar{b}, \tag{840}$$

kde $\bar{d} = d - \zeta_+ y$, $\bar{b} = y^T\bar{d}$, a aby Frobeniova norma $\|T^{-1/2}(\bar{S}_+ - \bar{S})\|_F$ byla minimální. Aplikujeme-li na tuto úlohu větu 92, dostaneme

$$\bar{S}_+ = \bar{S} - \frac{Ty}{y^T T y}\tilde{y}^T + \left(\rho\bar{d} - \bar{z} + \frac{y^T\bar{z}}{y^T T y} T y\right)\frac{\tilde{z}^T}{\tilde{z}^T\tilde{z}}, \tag{841}$$

kde $\tilde{y} = \bar{S}^T y$ a $\bar{z} = \bar{S}\tilde{z}$ (viz (364)). Zvolíme-li matici T tak, aby platilo $Ty = \rho\bar{d} - \bar{z}$, výraz (841) se velmi zjednoduší. Po dosazení a úpravě dostaneme

$$\bar{S}_+ = \bar{S} - \frac{\rho\bar{d} - \bar{z}}{\rho - y^T\bar{z}}(\tilde{y} - \tilde{z})^T. \tag{842}$$

Použití vzorců (841) a (842) není tak jednoduché, jako v případě standardních metod s proměnnou metrikou vyšetřovaných v oddílu 4.3. Předně neplatí $\bar{d} \in \mathcal{L}(\bar{S})$, takže neexistuje vektor $\tilde{d}$ takový, že $\bar{d} = \bar{S}\tilde{d}$, což je nutné k tomu abychom dostali aktualizaci (829) (poznámka 313). Nemůžeme ani položit $\tilde{d} = \bar{S}^T\bar{B}\bar{d}$, kde $\bar{B} = \hat{H}^\dagger$ (což by odpovídalo volbě (330)), neboť vektor $\bar{B}\bar{d}$ neumíme jednoduše spočítat. Místo toho budeme předpokládat, že platí $\tilde{d} = \bar{S}^T B d = -\alpha \bar{S}^T g$.

Poznámka 310. Položíme-li

$$\tilde{z} = \vartheta\tilde{d} = \vartheta\bar{S}^T B d, \qquad \vartheta = \pm\sqrt{\rho\bar{b}/\bar{c}}, \qquad \bar{c} = d^T B\bar{H}Bd$$

do (842), dostaneme

$$\bar{S}_+ = \bar{S} - \frac{\rho\bar{d} - \vartheta\bar{H}Bd}{\rho\bar{b} - \vartheta y^T\bar{H}Bd}\left(y - \vartheta Bd\right)^T\bar{S}. \tag{843}$$

Znaménko koeficientu ϑ je vhodné volit tak, aby jmenovatel v (843) byl co největší, tedy tak, aby platilo $\vartheta y^T\bar{H}Bd \leq 0$.

Poznámka 311. Podle věty 93 dostaneme standardní metodu BFGS, dosadíme-li $Ty = d$ a $\tilde{z} = \vartheta S^T B d$ do 364 (ϑ se volí tak, aby byla splněna poslední rovnost v (363)). Analogicky můžeme dosadit $Ty = \bar{d}$ a $\tilde{z} = \vartheta\bar{S}^T B d$ do (841). V tomto případě platí

$$\bar{S}_+ = \bar{S} - \frac{1}{\bar{b}}\bar{d}y^T\bar{S} + \frac{1}{\bar{c}}\left[\left(\frac{\rho}{\vartheta} + \frac{y^T\bar{H}Bd}{\bar{b}}\right)\bar{d} - \bar{H}Bd\right]d^TB\bar{S} \tag{844}$$

(neboť $\tilde{z}^T\tilde{z} = \vartheta^2\bar{c}$), kde čísla ϑ a $\bar{c}$ mají stejný význam jako v předchozí poznámce. Znaménko koeficientu ϑ volíme opět tak, aby platilo $\vartheta y^T\bar{H}Bd \leq 0$.

Poznámka 312. Vztahy (843) a (844) definují jednoduché metody, které jsou v jistém smyslu zobecněním metody BFGS. Obecnější metody dostaneme, volíme-li

$$Ty = \frac{\sqrt{\eta}}{\bar{b}}\bar{d} + \frac{1-\sqrt{\eta}}{\bar{a}}\bar{H}y \qquad \tilde{z} = \vartheta\bar{S}^T\left(\frac{\sqrt{\eta}}{\bar{b}}Bd + \frac{1-\sqrt{\eta}}{\bar{a}}y\right), \tag{845}$$

kde parametr ϑ se vybírá tak, aby platilo $\tilde{z}^T\tilde{z} = \rho\bar{b}$. Položíme-li

$$\hat{a} = y^T\bar{H}y, \qquad \hat{b} = y^T\bar{H}Bd, \qquad \hat{c} = d^TB\bar{H}Bd$$

(takže $\hat{a} = \bar{a}$), dostaneme

$$\begin{aligned}\rho\bar{b} &= \tilde{z}^T\tilde{z} = \vartheta^2\left(\frac{\eta}{\bar{b}^2}\hat{c} + 2\frac{\sqrt{\eta}(1-\sqrt{\eta})}{\bar{a}\bar{b}}\hat{b} + \frac{(1-\sqrt{\eta})^2}{\bar{a}^2}\hat{a}\right)\\ &= \frac{\vartheta}{\bar{a}\bar{b}^2}\left(\eta\hat{a}\hat{c} + 2\bar{b}\hat{b}\sqrt{\eta}(1-\sqrt{\eta}) + \hat{b}^2(1-\sqrt{\eta})^2\right) = \frac{\vartheta}{\bar{a}\bar{b}^2}\left(\eta\left(\hat{a}\hat{c} - \hat{b}^2\right) + \left(\bar{b} + (\hat{b}-\bar{b})\sqrt{\eta}\right)^2\right),\end{aligned}$$

takže

$$\vartheta^2 = \frac{\rho\bar{a}\bar{b}^3}{\eta\left(\hat{a}\hat{c} - \hat{b}^2\right) + \left(\bar{b} + (\hat{b}-\bar{b})\sqrt{\eta}\right)^2}. \tag{846}$$

Poznamenejme, že $\eta \geq 0$ podle předpokladu a $\hat{a}\hat{c} - \hat{b}^2 \geq 0$ podle Schwarzovy nerovnosti, takže výraz na pravé straně je kladný, pokud je definován. Vzorce (845)–(846) dosazujeme do obecného výrazu (841) (platí $y^TTy = 1$ a $\tilde{z}^T\tilde{z} = \rho\bar{b}$).

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 24. Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$ a vypočteme hodnoty $F_1 := F(x_1)$, $g_1 := g(x_1)$. Položíme $i := 1$.

Krok 2 Pokud $\|g_i\| \leq \underline{\varepsilon}$, ukončíme výpočet. V opačném případě položíme $m := \min(\bar{m}, i-1)$ a určíme směrový vektor s_i. Pokud $m = 0$, položíme $s_i := -g_i$, v opačném případě položíme $s_i := -\zeta_i g_i - \bar{S}_i \bar{S}_i^T g_i$.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1, položíme $x_{i+1} := x_i + \alpha_i s_i$ a vypočteme hodnoty $F_{i+1} := F(x_{i+1})$, $g_{i+1} := g(x_{i+1})$. Položíme $d_i := x_{i+1} - x_i$ a $y_i := g_{i+1} - g_i$.

Krok 4 Vypočteme číslo σ_i podle (831). Pokud $\sigma_i < 0.2$, položíme $\sigma_i := 0.2$. Pokud $\sigma_i > 0.8$, položíme $\sigma_i := 0.8$. Položíme $\zeta_{i+1} := \sigma_i y_i^T d_i / y_i^T y_i$.

Krok 5 Pokud $i \leq \bar{m}$, určíme matici $\bar{S}_{i+1}$ podle vzorce (839). Pokud $i > \bar{m}$, určíme matici $\bar{S}_{i+1}$ podle vzorce (844 nebo podle vzorce (841) s volbou (845)–(846). Zvětšíme i o 1 a přejdeme na krok 2.

Nyní vyšetříme globální konvergenci metody s proměnnou metrikou s omezenou pamětí realizované algoritmem 24.

Lemma 93. *Uvažujme posunutou metodu s proměnnou metrikou s omezenou pamětí s aktualizací (841) a volbou (845)–(846). Pak platí*

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T - \frac{1}{\bar{a}} \bar{H}y(\bar{H}y)^T + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right) \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right)^T - uu^T, \tag{847}$$

kde $u = (I - Tyy^T / y^T T y) \bar{S} \tilde{z} / \|\tilde{z}\|$.

Důkaz Roznásobením se snadno ukáže, že vzorec (841) lze zapsat ve tvaru

$$\bar{S}_+ = \left(I - \frac{Tyy^T}{y^T T y} \right) \bar{S} \left(I - \frac{\tilde{z}\tilde{z}^T}{\tilde{z}^T \tilde{z}} \right) + \rho \frac{\bar{d}\tilde{z}^T}{\tilde{z}^T \tilde{z}}.$$

Využijeme-li toho, že matice $I - \tilde{z}\tilde{z}^T / \tilde{z}^T \tilde{z}$ je idempotentní a platí $(I - \tilde{z}\tilde{z}^T / \tilde{z}^T \tilde{z})\tilde{z} = 0$, dostaneme po dosazení

$$\begin{aligned} \bar{H}_+ = \bar{S}_+ \bar{S}_+^T &= \left(I - \frac{Tyy^T}{y^T T y} \right) \bar{S} \left(I - \frac{\tilde{z}\tilde{z}^T}{\tilde{z}^T \tilde{z}} \right) \bar{S}^T \left(I - \frac{Tyy^T}{y^T T y} \right)^T + \rho \frac{\bar{d}\bar{d}^T}{\bar{b}} \\ &= \left(I - \frac{Tyy^T}{y^T T y} \right) \bar{H} \left(I - \frac{Tyy^T}{y^T T y} \right)^T + \rho \frac{\bar{d}\bar{d}^T}{\bar{b}} - uu^T, \end{aligned}$$

kde $u = (I - Tyy^T / y^T T y) \bar{S} \tilde{z} / \|\tilde{z}\|$. Použijeme-li první rovnost v (845), můžeme tento vzorec zapsat ve tvaru

$$\begin{aligned} \bar{H}_+ &= \left(I - \frac{Tyy^T}{y^T T y} \right) \bar{H} \left(I - \frac{Tyy^T}{y^T T y} \right)^T + \rho \frac{\bar{d}\bar{d}^T}{\bar{b}} - uu^T \\ &= \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T - \frac{1}{\bar{a}} \bar{H}y(\bar{H}y)^T + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right) \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right)^T - uu^T \end{aligned}$$

(lze se o tom přesvědčit prostým dosazením a roznásobením závorek). □

Poznámka 313. Z vyjádření (847) vyplývá, že aktualizace (841) s volbou (845)–(846) je ekvivalentní aktualizaci (829) právě tehdy, když $u = 0$ v (847), neboli když $S\tilde{z} \parallel Ty$ v (845). To lze dosáhnout pouze tehdy, když $\bar{d} = \bar{H}Bd$, neboli když $\bar{d} = \bar{S}\tilde{d}$, kde $\tilde{d} = \bar{S}^T Bd$. To nelze obecně zajistit, neboť nemusí platit $\bar{d} \in \mathcal{L}(\bar{S})$.

Lemma 93 použijeme k důkazu globální konvergence.

Důsledek 27. *Nechť jsou splněny předpoklady věty 203, kde aktualizace (829) je nahražena aktualizací (841) s volbou (845)–(846). Pak platí (834).*

Důkaz Označíme-li $\hat{H}_+$ matici určenou vztahem (829) a použitou v lemmatu 91, můžeme psát $\bar{H}_+ = \hat{H}_+ - uu^T$ a použijeme-li důsledek 12, dostaneme

$$\begin{aligned} \det(\zeta I + \bar{H}_+) &= \det(\zeta I + \hat{H}_+ - uu^T) = \det(\zeta I + \hat{H}_+)\, \det(1 - u^T((\zeta I + \hat{H}_+)^{-1}u) \\ &\leq \det((\zeta I + \hat{H}_+) \leq \left(\rho + \zeta \frac{y^T y}{\bar{b}}\right) \frac{\bar{d}^T B \bar{d}}{\bar{b}} \det H. \end{aligned}$$

Můžeme tedy použít lemma 91, lemma 92 a větu 203. □

9.7 Vektorové diferenční verze Newtonovy metody

Vektorové diferenční verze Newtonovy metody se obvykle realizují jako nepřesné metody spádových směrů (algoritmus 6) nebo nepřesné metody s lokálně omezeným krokem (algoritmus 11). V iteračním kroku metody sdružených gradientů se nepoužívá matice $B = G(x)$. Místo toho se násobení $q = Bp = G(x)p$ nahradí numerickým derivováním

$$G(x)p \approx \frac{g(x + \delta p) - g(x)}{\delta}, \tag{848}$$

kde $\delta = \varepsilon/\|p\|$ je vhodná diference (obvykle $\varepsilon = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost). Jinak se iterační krok metody sdružených gradientů nemění. Jestliže výpočet gradientu vyžaduje $O(n)$ operací, je tento způsob úspornější než násobení matice vektorem (obecně $O(n^2)$ operací). Navíc není třeba počítat druhé derivace.

Symbol $\approx$ v (848) znamená, že limita diferenčního podílu na pravé straně (pro $\delta \to 0$) se rovná výrazu na levé straně. Stejný význam bude mít symbol $\approx$ všude v tomto oddílu. Vliv diference $\delta = \varepsilon/\|p\|$ na přesnost Newtonovy metody udává tato věta.

Věta 204. *Nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňuje předpoklad (F6). Nechť $q = G(x)p$ a*

$$\tilde{q} = \frac{g(x + \delta p) - g(x)}{\delta}, \qquad \delta = \frac{\varepsilon}{\|p\|},$$

kde $x \in \mathcal{D}$ a $x + \delta p \in \mathcal{D}$. Pak platí

$$\|\tilde{q} - q\| \leq \frac{1}{2}\varepsilon \overline{L} \|p\|.$$

Důkaz Podle věty o střední hodnotě (tvrzení 3) platí

$$g(x + \delta p) = g(x) + \int_0^1 G(x + \tau\delta p)\delta p \mathrm{d}\tau,$$

takže

$$\begin{aligned} \|\tilde{q} - q\| &= \frac{1}{\delta}\left\| \int_0^1 (G(x + \tau\delta p) - G(x))\delta p \mathrm{d}\tau \right\| \leq \frac{1}{\delta}\int_0^1 \|G(x + \tau\delta p) - G(x))\|\|\delta p\| \mathrm{d}\tau \\ &\leq \frac{1}{\delta}\int_0^1 \overline{L}\|\delta p\|^2 \tau \mathrm{d}\tau = \frac{1}{2}\overline{L}\delta\|p\|^2 = \frac{1}{2}\varepsilon\overline{L}\|p\| \end{aligned}$$

□

V dalším výkladu budeme předpokládat, že směrový vektor s se určuje předpodmíněnou metodou sdružených gradientů popsanou v oddílu 3.8. Aby nedocházelo k nedorozumění, budeme pro vnější iterace (kroky Newtonovy metody) používat index i a pro vnitřní iterace (kroky metody sdružených gradientů) index j. Index i budeme často vynechávat.

Věta 205. *Uvažujme předpodmíněnou metodu sdružených gradientů (definice 35) aplikovanou na soustavu lineárních rovnic $G(x)s + g = 0$, kde vektory $q_j = G(x)p_j$ jsou nahraženy vektory $\tilde{q}_j = (g(x + \delta_j p_j) - g(x))/\delta_j$, $\delta_j = \varepsilon/\|p_j\|$. Předpokládejme, že jsou splněny předpoklady důsledku 4 a věty 204 a označme*

$$s_{m+1} = s_1 + \sum_{j=1}^{m} \alpha_j p_j, \qquad g_{m+1} = g_1 + \sum_{j=1}^{m} \alpha_j q_j, \qquad \tilde{g}_{m+1} = g_1 + \sum_{j=1}^{m} \alpha_j \tilde{q}_j$$

(takže $g_{m+1} = g + G(x)s_{m+1}$, počítáme-li přesně). Pak platí

$$\|\tilde{g}_{m+1} - g_{m+1}\| \le \overline{\vartheta}\|s_{m+1}\|, \qquad \overline{\vartheta} = \frac{1}{2} m\varepsilon\overline{L}\sqrt{\kappa(C)}, \tag{849}$$

kde $\kappa(C)$ je spektrální číslo podmíněnosti matice C.

Důkaz Použijeme-li nerovnost uvedenou ve větě 204, dostaneme

$$\|\tilde{g}_{m+1} - g_{m+1}\| = \left\| \sum_{j=1}^{m} \alpha_j (\tilde{q}_j - q_j) \right\| \le \sum_{j=1}^{m} \alpha_j \|\tilde{q}_j - q_j\| \le \frac{1}{2}\varepsilon\overline{L} \sum_{j=1}^{m} \alpha_j \|p_j\|.$$

Aplikujeme-li postup uvedený v části (c) důkazu věty 69 na předpodmíněnou metodu sdružených gradientů a použijeme-li poznámku 95, můžeme pro $1 \le j \le m$ psát $\alpha_j\|p_j\|_C \le \|s_{j+1}\|_C \le \|s_{m+1}\|_C$, neboli $\alpha_j\|p_j\| \le \sqrt{\kappa(C)}\|s_{m+1}\|$, takže

$$\sum_{i=j}^{m} \alpha_j \|p_j\| \le m\sqrt{\kappa(C)}\|s_{m+1}\|,$$

což spolu s předchozí nerovností dokazuje tvrzení věty. □

Poznámka 314. Předpokládejme, že v m-tém kroku metody sdružených gradientů platí $\|\tilde{g}_{m+1}\| \le \overline{\omega}\|g\|$, $0 \le \overline{\omega} < 1$. Pak, položíme-li $s = s_{m+1}$ a $\tilde{g} = \tilde{g}_{m+1}$, můžeme podle předpokladu a podle věty 205 psát

$$\frac{\|\tilde{G}s + g\|}{\|g\|} \le \overline{\omega}, \qquad \frac{\|(\tilde{G} - G)s\|}{\|s\|} \le \overline{\vartheta},$$

kde $\tilde{G}$ je nějaká symetrická matice, pro kterou platí $\tilde{G}s + g = \tilde{g}$, a kde $\overline{\vartheta} = m\varepsilon\overline{L}\sqrt{\kappa(C)}/2$. Zvolíme-li číslo ε tak, aby platilo $\varepsilon < (1 - \overline{\omega})\underline{G}/(m\overline{L}\sqrt{\kappa(C)})$, dostaneme $\overline{\vartheta} < (1 - \overline{\omega})\,\underline{G}/2$ a můžeme použít větu 22, podle které diferenční verze Newtonovy metody konverguje lineárně s asymptotickou rychlostí alespoň $(\overline{\omega}\,\underline{G} + \overline{\vartheta})/(\underline{G} - \overline{\vartheta})$. Poznamenejme, že nerovnosti použité ve větě 22 a ve větě 205 jsou značně nadhodnocené, takže diferenční verze nepřesné Newtonovy metody obvykle fungují velmi dobře i pro standardní volbu $\varepsilon = \sqrt{\varepsilon_M}$.

Nevýhodou vektorových diferenčních verzí Newtonovy metody je skutečnost, že počet vnitřních iterací metody sdružených gradientů, tedy i počet vyčíslení gradientů minimalizované funkce, může být značně velký, je-li matice $G = G(x)$ špatně podmíněná. Proto je účelné metodu sdružených gradientů vhodně předpodmínit. Potíž je v tom, že neznáme matici G, takže není možné použít standardní postupy. V tomto oddílu popíšeme čtyři základní způsoby předpodmiňování.

(1) Použití metod s proměnnou metrikou s omezenou pamětí [71], [120].
(2) Použití pásových matic určených standardní metodou BFGS odvozenou z předpodmíněné metody sdružených gradientů [124], [108].
(3) Použití pásových matic určených směrovým derivováním (automatickým nebo numerickým) [108], [116], [140].
(4) Použití pásových matic určených analyticky podle vzorců definujících prvky Hessovy matice.
(5) Použití tridiagonálních matic určených Lanczosovou metodou odvozenou z nepředpodmíněné metody sdružených gradientů [50], [51], [124].

Předpodmiňování pomocí metod s proměnnou metrikou s omezenou pamětí je velmi jednoduché a přímočaré. V i-tém kroku Newtonovy metody se používá předpodmiňovač $C = (H_i^i)^{-1}$, kde H_i^i je matice uvedená v definici 57. Vektory $C^{-1}g_j = H_i^i\, g_j$, sloužící k výpočtu vektorů p_j (definice 35), určujeme buď pomocí Strangových rekurencí (důsledek 25), kde místo vektoru g_i používáme postupně vektory g_j, $1 \le j \le m$, nebo pomocí maticových reprezentací (věta 184, věta 186). Protože se násobení maticí H_i^i používá v každém kroku metody sdružených gradientů, je třeba aby počet použitých aktualizací byl co nejmenší. Vhodným kompromisem je volba $\bar{m} = 3$ (maximálně tři aktualizace).

Další způsob předpodmínění využívá toho, že metody s proměnnou metrikou, s vhodnou počáteční maticí a s přesným výběrem délky kroku, aplikované na ryze konvexní kvadratickou funkci (237) generují stejnou posloupnost vektorů p_j, $1 \le j \le m$, jako předpodmíněná metoda sdružených gradientů uvedená v definici 35 (věta 75 a důsledek 7). Pro metodu BFGS platí $B_j p_j + g_j = 0$, $1 \le j \le m$, kde $B_1 = C$ a

$$B_{j+1} = B_j + \frac{y_j y_j^T}{d_j^T y_j} - \frac{B_j d_j (B_j d_j)^T}{d_j^T B_j d_j} = B_j + \frac{Gp_j (Gp_j)^T}{p_j^T G p_j} + \frac{g_j g_j^T}{p_j^T g_j},$$

přičemž $d_j = s_{j+1} - s_j = \alpha_j p_j$ a $y_j = g_{j+1} - g_j = G d_j$. Použijeme-li místo vektorů $q_j = Gp_j$ a g_j vektory $\tilde{q_j}$ (určené numerickým derivováním) a $\tilde{g_j}$, můžeme psát $B_1 = C$ a

$$B_{j+1} = B_j + \frac{\tilde{q}_j \tilde{q}_j^T}{p_j^T \tilde{q}_j} + \frac{\tilde{g}_j \tilde{g}_j^T}{p_j^T \tilde{g}_j}, \qquad 1 \le j \le m. \tag{850}$$

Z tohoto vyjádření je patrné, že k aktualizaci matic B_j, $1 \le j \le m$, se používají pouze vektory generované předpodmíněnou metodou sdružených gradientů (s maticovým násobením nahraženým numerickým derivováním). Matice B_j, $1 \le j \le m$, se v korekčních členech nevyskytují, takže můžeme ukládat pouze jejich části. Jsou-li vektory $\tilde{q_j}$ a $\tilde{g_j}$ dobrou aproximací vektorů q_j a g_j, jsou matice B_j, $1 \le j \le m$, pozitivně definitní a je-li počet kroků metody sdružených gradientů dostatečně velký, je matice B_{m+1} dobrou aproximací matice G a můžeme ji (nebo její část) použít jako předpodmiňovač v dalším iteračním kroku Newtonovy metody.

I když matice $B = B_{m+1}$ je dobrou aproximací matice G, je volba $C = B$ nevhodná, neboť tato matice je obvykle hustá. Proto je výhodnější použít jako předpodmiňovač pásovou matici, která vznikne z B vynulováním prvků neležících v použitém pásu (v (850) se tedy aktualizují pouze prvky ležící v použitém pásu). V dalším výkladu se omezíme na diagonální, tridiagonální a pentadiagonální předpodmiňovače. Ukazuje se, že je důležité, aby tyto předpodmiňovače byly pozitivně definitní.

V případě, že $C = D$, kde D je diagonální matice obsahující diagonální prvky matice B, nenastávají žádné potíže, neboť pozitivně definitní matice B má kladné prvky na hlavní diagonále. Použití matice $C = D$ zdůvodňuje toto tvrzení uvedené v [82].

Tvrzení 7. *Nechť $\mathcal{D}_n$ je množina všech diagonálních matic řádu n a D je diagonální matice obsahující diagonální prvky matice G. Pak platí*

$$\kappa(GD^{-1}) \le l \min_{M \in \mathcal{D}_n} \kappa(GM^{-1}),$$

kde κ je spektrální číslo podmíněnosti a l je maximální počet nenulových prvků v řádcích matice G (pro pentadiagonální matici G je $l = 5$).

Nechť nyní $C = T$, kde T je symetrická tridiagonální matice obsahující prvky tří hlavních diagonál matice B. V tomto případě nemusí být matice C pozitivně definitní (i když B je pozitivně definitní). Jako příklad uvažujme matice

$$B = \begin{bmatrix} 2 & -2 & 2 \\ -2 & 3 & -3 \\ 2 & -3 & 4 \end{bmatrix}, \qquad T = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 3 & -3 \\ 0 & -3 & 4 \end{bmatrix}.$$

Obě tyto matice mají kladné prvky na hlavní diagonále a kladné hlavní subdeterminanty druhého řádu. Platí ale $\det B = 2$ a $\det T = -10$, takže T není pozitivně definitní, i když B je pozitivně definitní.

Abychom tento nedostatek odstranili, je třeba matici T upravit. To lze zařídit během provádění Gillova-Murrayova rozkladu matice T. Výhodnější je však upravit matici T předem tak, aby byla pozitivně definitní. Jednou z možností je implicitní použití modifikovaného Choleského rozkladu matice T.

Lemma 94. *Uvažujme tridiagonální matici*

$$T = \begin{bmatrix} \alpha_1, & \beta_1, & \ldots, & 0, & 0, \\ \beta_1, & \alpha_2, & \ldots, & 0, & 0, \\ -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \alpha_{n-1}, & \beta_{n-1} \\ 0, & 0, & \ldots, & \beta_{n-1}, & \alpha_n \end{bmatrix}. \tag{851}$$

a označme Δ_i hlavní subdeterminant i-tého řádu matice T (obsahující řádky a sloupce s indexy $1, 2, \ldots, i$). Pak platí $\Delta_1 = \alpha_1$ a

$$\Delta_{i+1} = \alpha_{i+1}\Delta_i - \beta_i^2\Delta_{i-1}, \qquad 1 \le i \le n-1, \tag{852}$$

kde pokládáme $\Delta_0 = 1$.

Důkaz Pro $i = 1$ je tvrzení lemmatu zřejmé. Nechť $i > 1$. Rozvedeme-li subdeterminant Δ_{i+1} podle $i+1$-tého řádku, dostaneme $\Delta_{i+1} = \alpha_{i+1}\Delta_i - \beta_i^2\Delta_{i-1}$, neboť $i+1$-tý řádek obsahuje pouze dva prvky. □

Věta 206. *Tridiagonální matice (851) je pozitivně definitní právě tehdy, když $\tau_i > 0$ pro $1 \le i \le n$, kde $\tau_1 = \alpha_1$ a*

$$\tau_{i+1} = \alpha_{i+1} - \frac{\beta_i^2}{\tau_i}, \qquad 1 \le i \le n-1. \tag{853}$$

Důkaz Dokážeme indukcí, že $\Delta_i = \tau_i\Delta_{i-1}$ pro $1 \le i \le n$, kde opět $\Delta_0 = 1$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme, že pro nějaký index $i \ge 1$ platí $\Delta_i = \tau_i\Delta_{i-1}$. Použijeme-li (852) a (853), dostaneme

$$\begin{aligned} \Delta_{i+1} &= \alpha_{i+1}\Delta_i - \beta_i^2\Delta_{i-1} = \alpha_{i+1}\Delta_i + \tau_i(\tau_{i+1} - \alpha_{i+1})\Delta_{i-1} \\ &= (\Delta_i - \tau_i\Delta_{i-1})\alpha_{i+1} + \tau_i\tau_{i+1}\Delta_{i-1} = \tau_{i+1}\Delta_i, \end{aligned}$$

čímž je indukční krok dokončen. Jelikož $\Delta_0 = 1$ a $\Delta_i = \tau_i\Delta_{i-1}$ pro $1 \le i \le n$, platí $\Delta_i > 0$ právě tehdy, když $\tau_i > 0$ (pro $1 \le i \le n$). □

Poznámka 315. Větu 206 můžeme použít tak, že počítáme čísla τ_{i+1}, $1 \le i \le n-1$, podle (853) a platí-li pro nějaký index $\tau_{i+1} \le 0$, zmenšíme mimodiagonální prvek β_i tak, aby platilo $\beta_i^2 < \tau_i\alpha_{i+1}$ (například položíme $\beta_i^2 = \lambda_i\tau_i\alpha_{i+1}$, kde $0 < \lambda_i < 1$). Pak je nové číslo τ_{i+1} kladné. Čísla τ_i, $1 \le i \le n$, jsou prvky diagonální matice D vystupující v Choleského rozkladu $T = LDL^T$. Zmenšování mimodiagonálních prvků matice T lze tedy považovat za korekci Choleského rozkladu.

Postup uvedený v poznámce 315 není příliš efektivní (je obtížné nalézt vhodné koeficienty λ_i, $1 \le i \le n$). Pro praktické účely je výhodnější použít následující větu a její důsledek.

Věta 207. *Uvažujme tridiagonální matici (851) s kladnými prvky na hlavní diagonále. Pak jsou-li matice*

$$\begin{bmatrix} \alpha_i, & 2\beta_i \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix}, \qquad 1 \le i < n-1, \tag{854}$$

pozitivně semidefinitní je matice T pozitivně definitní.

Důkaz Pro libovolný nenulový vektor $v \in R^n$ platí

$$\begin{aligned} v^T T v &= \sum_{i=1}^{n} \alpha_i v_i^2 + 2 \sum_{i=1}^{n-1} \beta_i v_i v_{i+1} \\ &= \frac{1}{2}\alpha_1 v_1^2 + \frac{1}{2} \sum_{i=1}^{n-1} \left(\alpha_i v_i^2 + \alpha_{i+1} v_{i+1}^2 + 4\beta_i v_i v_{i+1} \right) + \frac{1}{2}\alpha_n v_n^2 \\ &= \frac{1}{2}\alpha_1 v_1^2 + \frac{1}{2} \sum_{i=1}^{n-1} [v_i, v_{i+1}] \begin{bmatrix} \alpha_i, & 2\beta_i, \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} + \frac{1}{2}\alpha_n v_n^2 \end{aligned} \tag{855}$$

Jelikož matice vystupující v této rovnosti jsou podle předpokladu pozitivně semidefinitní, platí $v^T T v \geq 0$. Předpokládejme, že $v^T T v = 0$. Dokážeme indukcí, že $v = 0$. Jelikož $\alpha_1 > 0$, musí být $v_1 = 0$. Předpokládejme, že $v_j = 0$ pro $1 \leq j \leq i$, kde $i < n$. Pak jelikož

$$[v_i, v_{i+1}] \begin{bmatrix} \alpha_i, & 2\beta_i \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} = \alpha_{i+1} v_{i+1}^2$$

a $\alpha_{i+1} > 0$, musí být $v_{i+1} = 0$. Dokázali jsme, že pro libovolný vektor $v \in R^n$ platí $v^T T v \geq 0$ a že z $v^T T v = 0$ plyne $v = 0$, takže matice T je pozitivně definitní. □

Poznámka 316. Věta 207 udává podmínky postačující, nikoliv však nutné. Jelikož vlastní čísla pozitivně definitní matice jsou kladná a spojitě závislá na prvcích této matice, lze prvky matice změnit tak, že předpoklady věty 207 neplatí, ale vlastní čísla zůstanou kladná. Tato výhoda se projeví, počítáme-li prvky matice T nepřesně pomocí numerického derivování (věta 213).

Poznámka 317. Podíváme-li se na poslední výraz ve vzorci (855), vidíme, že člen $\alpha_1 v_1^2$ lze přidat k prvnímu členu v součtu a člen $\alpha_n v_n^2$ k poslednímu. Matice T je tedy pozitivně definitní, jsou-li matice

$$\begin{bmatrix} 2\alpha_1, & 2\beta_1 \\ 2\beta_1, & \alpha_2 \end{bmatrix}, \qquad \begin{bmatrix} \alpha_i, & 2\beta_i \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix}, \qquad \begin{bmatrix} \alpha_{n-1}, & 2\beta_{n-1} \\ 2\beta_{n-1}, & 2\alpha_n \end{bmatrix},$$

kde $2 \leq i < n-2$, pozitivně semidefinitní a alespoň jedna z nich je pozitivně definitní. Tyto podmínky jsou velmi užitečné, neboť prvky α_1 a α_n jsou často menší, než ostatní diagonální prvky (věta 213).

Důsledek 28. *Nechť tridiagonální matice T obsahuje hlavní diagonálu a poloviny vedlejších diagonál symetrické pozitivně definitní matice B (takže $\alpha_i = b_{i,i}$, $1 \leq i \leq n-1$ a $\beta_i = b_{i,i+1}/2$, $1 \leq i \leq n-1$). Pak T je pozitivně definitní.*

Důkaz Dosadíme-li $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$ a $\beta_i = b_{i,i+1}/2$, dostaneme

$$\begin{bmatrix} \alpha_i, & 2\beta_i \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix} = \begin{bmatrix} b_{i,i}, & b_{i,i+1} \\ b_{i,i+1}, & b_{i+1,i+1} \end{bmatrix}, \qquad 1 \leq i \leq n-1.$$

Tyto matice jsou pozitivně definitní, neboť matice B je pozitivně definitní. □

Poznámka 318. Větu 207 a důsledek 28 můžeme použít třemi různými způsoby.

(1) Prvky matice T určíme podle důsledku 28 tak, že pro $1 \leq i \leq n-1$ položíme $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$ a $\beta_i = b_{i,i+1}/2$.

(2) Pro $1 \leq i \leq n-1$ položíme $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$, $\beta_i = b_{i,i+1}$ a vypočteme determinant $\alpha_i \alpha_{i+1} - 4\beta_i^2$ matice (854). Pokud $\alpha_i \alpha_{i+1} - 4\beta_i^2 \geq 0$, ponecháme β_i beze změny, v opačném případě zmenšíme β_i na polovinu.

(3) Pro $1 \leq i \leq n-1$ položíme $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$, $\beta_i = b_{i,i+1}$ a vypočteme determinant $\alpha_i \alpha_{i+1} - 4\beta_i^2$ matice (854). Pokud $\alpha_i \alpha_{i+1} - 4\beta_i^2 \geq 0$, ponecháme β_i beze změny, v opačném případě položíme $\beta_i = (1/2)\sqrt{\alpha_i \alpha_{i+1}}$.

Ve všech těchto případech je výsledná matice positivně definitní.

Větu 207 a její důsledek můžeme zobecnit tak, aby platila i pro další symetrické pásové matice. Ukážeme, jak to vypadá v případě pentadiagonální matice

$$P = \begin{bmatrix} \alpha_1, & \beta_1, & \gamma_{,1} & \ldots, & 0, & 0, & 0 \\ \beta_1, & \alpha_2, & \beta_2, & \ldots, & 0, & 0, & 0 \\ \gamma_1, & \beta_2, & \alpha_3, & \ldots, & 0, & 0, & 0 \\ -- & -- & -- & -- & -- & -- & -- \\ 0, & 0, & 0, & \ldots, & \alpha_{n-2}, & \beta_{n-2}, & \gamma_{n-2} \\ 0, & 0, & 0, & \ldots, & \beta_{n-2}, & \alpha_{n-1}, & \beta_{n-1} \\ 0, & 0, & 0, & \ldots, & \gamma_{n-2}, & \beta_{n-1}, & \alpha_n \end{bmatrix}. \tag{856}$$

Věta 208. *Uvažujme pentadiagonální matici (856) s kladnými prvky na hlavní diagonále. Pak, jsou-li matice*

$$\begin{bmatrix} \alpha_i, & (3/2)\beta_i, & 3\gamma_i \\ (3/2)\beta_i, & \alpha_{i+1}, & (3/2)\beta_{i+1} \\ 3\gamma_i, & (3/2)\beta_{i+1}, & \alpha_{i+2} \end{bmatrix}, \qquad 1 \le i < n-2, \tag{857}$$

pozitivně semidefinitní, je matice P pozitivně definitní.

Důkaz Pro libovolný nenulový vektor $v \in R^n$ platí

$$\begin{aligned} v^T P v &= \sum_{i=1}^{n} \alpha_i v_i^2 + 2\sum_{i=1}^{n-1} \beta_i v_i v_{i+1} + 2\sum_{i=1}^{n-2} \gamma_i v_i v_{i+2} \\ &= \frac{1}{3}\alpha_1 v_1^2 + \frac{1}{3}(\alpha_1 v_1^2 + \alpha_2 v_2^2) + \beta_1 v_1 v_2 \\ &+ \frac{1}{3}\sum_{i=1}^{n-2} \left(\alpha_i v_i^2 + \alpha_{i+1} v_{i+1}^2 + \alpha_{i+2} v_{i+2}^2 + 3\beta_i v_i v_{i+1} + 3\beta_{i+1} v_{i+1} v_{i+2} + 6\gamma_i v_i v_{i+2}\right) \\ &+ \frac{1}{3}(\alpha_{n-1} v_{n-1}^2 + \alpha_n v_n^2) + \beta_{n-1} v_{n-1} v_n + \frac{1}{3}\alpha_n v_n^2 \\ &= \frac{1}{3}\alpha_1 v_1^2 + \frac{1}{3}\,[v_1, v_2] \begin{bmatrix} \alpha_1, & (3/2)\beta_1 \\ (3/2)\beta_1, & \alpha_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &+ \frac{1}{3}\sum_{i=1}^{n-2} [v_i, v_{i+1}, v_{i+2}] \begin{bmatrix} \alpha_i, & (3/2)\beta_i, & 3\gamma_i \\ (3/2)\beta_i, & \alpha_{i+1}, & (3/2)\beta_{i+1} \\ 3\gamma_i, & (3/2)\beta_{i+1}, & \alpha_{i+2} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \\ v_{i+2} \end{bmatrix} \\ &+ \frac{1}{3}\,[v_{n-1}, v_n] \begin{bmatrix} \alpha_{n-1}, & (3/2)\beta_{n-1} \\ (3/2)\beta_{n-1}, & \alpha_n \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} + \frac{1}{3}\alpha_n v_n^2 \end{aligned}$$

Jelikož matice vystupující v této rovnosti jsou podle předpokladu pozitivně semidefinitní, platí $v^T P v \ge 0$. Předpokládejme, že $v^T P v = 0$. Dokážeme indukcí, že $v = 0$. Tak jako v důkazu věty 207, musí být $v_1 = 0$ a $v_2 = 0$. Předpokládejme, že $v_j = 0$ pro $1 \le j \le i+1$, kde $i \le n-2$. Pak jelikož

$$[v_i, v_{i+1}, v_{i+1}] \begin{bmatrix} \alpha_i, & (3/2)\beta_i, & 3\gamma_i \\ (3/2)\beta_i, & \alpha_{i+1}, & (3/2)\beta_{i+1} \\ 3\gamma_i, & (3/2)\beta_{i+1}, & \alpha_{i+2} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \\ v_{i+2} \end{bmatrix} = \alpha_{i+2} v_{i+2}^2$$

a $\alpha_{i+2} > 0$, musí být $v_{i+2} = 0$. Dokázali jsme, že pro libovolný vektor $v \in R^n$ platí $v^T P v \ge 0$ a že z $v^T P v = 0$ plyne $v = 0$, takže matice P je pozitivně definitní. □

Důsledek 29. *Nechť pentadiagonální matice P obsahuje hlavní diagonálu, dvě třetiny prvních vedlejších diagonál a třetiny druhých vedlejších diagonál pozitivně definitní matice B (takže $\alpha_i = b_{i,i}$, $1 \le i \le n$, $\beta_i = 2b_{i,i+1}/3$, $1 \le i \le n-1$, a $\gamma_i = b_{i,i+2}/3$, $1 \le i \le n-2$). Pak P je pozitivně definitní.*

Důkaz Dosadíme-li $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$, $\alpha_{i+2} = b_{i+2,i+2}$, $\beta_i = 2b_{i,i+1}/3$, $\beta_{i+1} = 2b_{i+1,i+2}/3$ a $\gamma_i = b_{i,i+2}/3$, dostaneme

$$\begin{bmatrix} \alpha_i, & (3/2)\beta_i, & 3\gamma_i \\ (3/2)\beta_i, & \alpha_{i+1}, & (3/2)\beta_{i+1} \\ 3\gamma_i, & (3/2)\beta_{i+1}, & \alpha_{i+2} \end{bmatrix} = \begin{bmatrix} b_{i,i}, & b_{i,i+1}, & b_{i,i+2} \\ b_{i,i+1}, & b_{i+1,i+1}, & b_{i+1,i+2} \\ b_{i,i+2}, & b_{i+1,i+2}, & b_{i+2,i+2} \end{bmatrix}, \qquad 1 \le i \le n-2.$$

Tyto matice jsou pozitivně definitní, neboť matice B je pozitivně definitní. □

Věta 209. *Nechť jsou splněny přepoklady věty 208. Pak determinanty Δ_i matic (857) spočteme podle vzorce*

$$\Delta_i = \alpha_{i+1}\left(\alpha_i\alpha_{i+2} - 9\gamma_i^2\right) - \frac{9}{4}\left(\alpha_i\beta_{i+1}^2 + \alpha_{i+2}\beta_i^2 - 6\beta_i\beta_{i+1}\gamma_i\right). \tag{858}$$

Determinant Δ_i je nezáporný právě tehdy, když $\underline{\gamma}_i \le \gamma_i \le \overline{\gamma}_i$, kde

$$\begin{aligned} \underline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} - \sqrt{D_i}\right), \\ \overline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} + \sqrt{D_i}\right) \end{aligned}$$

jsou kořeny kvadratické rovnice $\Delta_i = 0$ a

$$D_i = \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right)$$

je diskriminant této rovnice (vydělený číslem 36), který je nezáporný, pokud $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2 \ge 0$ a $\alpha_{i+1}\alpha_{i+2} - (9/4)\beta_{i+1}^2 \ge 0$.

Důkaz Vztah (858) dostaneme snadno vyčíslením příslušného determinantu. Jelikož kvadratický člen v (858) má záporné znaménko, je determinant Δ_i nezáporný právě tehdy, když $\underline{\gamma}_i \le \gamma_i \le \overline{\gamma}_i$, kde $\underline{\gamma}_i$, $\overline{\gamma}_i$ jsou kořeny kvadratické rovnice $\Delta_i = 0$. Podle (858) se diskriminant této rovnice (vydělený číslem 36) rovná

$$\begin{aligned} D_i &= \frac{81}{16}\beta_i^2\beta_{i+1}^2 - \frac{9}{4}\alpha_i\alpha_{i+1}\beta_{i+1}^2 - \frac{9}{4}\alpha_{i+1}\alpha_{i+2}\beta_i^2 + \alpha_i\alpha_{i+1}^2\alpha_{i+2} \\ &= \frac{9}{4}\beta_{i+1}^2\left(\frac{9}{4}\beta_i^2 - \alpha_i\alpha_{i+1}\right) - \alpha_{i+1}\alpha_{i+2}\left(\frac{9}{4}\beta_i^2 - \alpha_i\alpha_{i+1}\right) \\ &= \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right). \end{aligned}$$

□

Věta 209 nabízí dvě možnosti, jak volit nový prvek γ_i v případě, že $\Delta_i < 0$. V prvním případě pokládáme $\gamma_i := \underline{\gamma}_i$, pokud $\gamma_i < \underline{\gamma}_i$, nebo $\gamma_i := \overline{\gamma}_i$, pokud $\gamma_i > \overline{\gamma}_i$. Tento způsob je náročnější na výpočet a dává horší praktické výsledky. Výhodnější je pokládat

$$\gamma_i = \frac{1}{2}(\underline{\gamma}_i + \overline{\gamma}_i) = \frac{3}{4}\frac{\beta_i\beta_{i+1}}{\alpha_{i+1}}. \tag{859}$$

Poznámka 319. Větu 208 a důsledek 29 můžeme použít třemi různými způsoby.

(1) Prvky matice P určíme podle důsledku 28. Pro $1 \le i \le n$ položíme $\alpha_i = b_{i,i}$. Pro $1 \le i \le n-1$ položíme $\beta_i = 2b_{i,i+1}/3$. Pro $1 \le i \le n-2$ položíme $\gamma_i = b_{i,i+2}/3$.

(2) Pro $1 \le i \le n-2$ položíme $\alpha_i = b_{i,i}$, $\alpha_{i+1} = b_{i+1,i+1}$, $\alpha_{i+2} = b_{i+2,i+2}$, $\beta_i = b_{i,i+1}$, $\beta_{i+1} = b_{i+1,i+2}$ a $\gamma_i = b_{i,i+2}$. Je-li matice (857) pozitivně definitní, ponecháme β_i, β_{i+1} a γ_i beze změny, v opačném případě položíme $\beta_i = 2b_{i,i+1}/3$, $\beta_{i+1} = 2b_{i+1,i+2}/3$ a $\gamma_i = b_{i,i+2}/3$.

(3) Pro $1 \leq i \leq n$ položíme $\alpha_i = b_{i,i}$. Pro $1 \leq i \leq n-1$ položíme $\beta_i = b_{i,i+1}$ a vypočteme hlavní subdeterminant $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2$ matice (857). Pokud $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2 \geq 0$, ponecháme β_i beze změny, v opačném případě položíme $\beta_i = (2/3)\sqrt{\alpha_i\alpha_{i+1}}$. Pro $1 \leq i \leq n-2$ položíme $\gamma_i = b_{i,i+2}$ a vypočteme determinant Δ_i podle (858). Pokud $\Delta_i \geq 0$, ponecháme γ_i beze změny, v opačném případě vypočteme γ_i podle (859).

Ve všech těchto případech je výsledná matice pozitivně definitní.

Poznámka 320. Zatím jsme předpokládali, že předpodmiňovač má nanejvýš pět diagonál, ale důsledek 29 lze zobecnit pro jakoukoliv šířku pásu. Nechť C je symetrická pásová matice s pásem šířky l, takže má hlavní diagonálu a $k-1 = (l-1)/2$ párů vedlejších diagonál, které jsou shodné s odpovídajícími si diagonálami pozitivně definitní matice B. Pak vynásobíme-li i-tý pár vedlejších diagonál číslem $(k-i)/k$ (pro $1 \leq i \leq k-1$), je výsledná matice pozitivně definitní. Důkaz tohoto tvrzení je podobný důkazu důsledku 29 (používá se analogie věty 208).

Další možností, jak konstruovat pásové předpodmiňovače, je předpokládat, že Hessova matice má pásovou strukturu a určovat její prvky pomocí směrového derivování. K určení všech prvků pásové matice, která má $k-1$ párů vedlejších diagonál (takže $k = (l+1)/2$, kde l je šířka pásu), stačí použít k vektorů druhých směrových derivací. Vyšetříme opět tři speciální případy.

Předpokládejme, že Hessova matice minimalizované funkce je diagonální, neboli $G(x) = D(x)$, kde $D(x) = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$. Pak lze všechny její prvky určit pomocí jednoho vektoru druhých směrových derivací

$$G(x)v = g'(x,v) \stackrel{\Delta}{=} \lim_{\varepsilon \to 0} \frac{g(x+\varepsilon v) - g(x)}{\varepsilon}, \tag{860}$$

kde $v = [\delta_1, \ldots, \delta_n]^T$. Jelikož $\alpha_i\delta_i = e_i^T D(x)v$, kde e_i, $1 \leq i \leq n$, jsou sloupce jednotkové matice, platí

$$\alpha_i = \frac{e_i^T g'(x,v)}{\delta_i}, \qquad 1 \leq i \leq n. \tag{861}$$

Poznámka 321. Vektor (860) můžeme určit přesně pomocí automatického derivování (oddíl 14). V tomto případě volíme vektor v tak, že $\delta_i = 1$, $1 \leq i \leq n$. Vektor (860) lze také spočítat přibližně pomocí numerického derivování. Pak

$$G(x)v = g'(x,v_1) \approx D_\varepsilon(x)v \stackrel{\Delta}{=} \frac{g(x+\varepsilon v) - g(x)}{\varepsilon}, \tag{862}$$

kde $\varepsilon > 0$ je vhodně zvolená hodnota (obvykle $\varepsilon \approx \sqrt{\varepsilon_M}$). Čísla $\delta_i > 0$, $1 \leq i \leq n$, lze volit dvěma různými způsoby. Buď

$$\delta_i = \delta > 0, \quad 1 \leq i \leq n, \quad (\text{obvykle } \delta \approx \sqrt{1/n}), \tag{863}$$

nebo

$$\delta_i = \max(|x_i|, 1), \quad 1 \leq i \leq n, \tag{864}$$

Vzorce (861) a (862) použijeme ke konstrukci diagoálních předpodmiňovačů $D(x)$ a $D_\varepsilon(x)$ a to i v případě, že Hessova matice $G(x)$ není diagonální, takže $G(x) \neq D(x)$. Nicméně pro vektor $v = [\delta_1, \ldots, \delta_n]^T$ i v tomto případě platí $G(x)v = D(x)v$. Pokud $\delta_i = \delta$, $1 \leq i \leq n$ (takže $v = \delta e$, kde $e = [1, \ldots, 1]$), lze prvky matice $D(x)$ vyjádřit ve tvaru

$$\alpha_i = e_i^T D(x)e = e_i^T G(x)e = \sum_{j=1}^{n} G_{ij}, \quad 1 \leq i \leq n, \tag{865}$$

kde e_i, $1 \leq i \leq n$ jsou sloupce jednotkové matice. V dalších úvahách budeme předpokládat, že $D(x)$ je matice, která má prvky (861), a $D_\varepsilon(x)$ je matice určená podle vzorce (862).

Nevýhodou předpodmiňovačů založených na numerickém derivování je skutečnost, že nemusejí být pozitivně definitní, i když Hessova matice je pozitivné definitní.

Příklad 5. *Uvažujme ryze konvexní kvadratickou funkci $F : R^2 \to R$, kde*

$$F(x) = \frac{1}{2} x^T \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x, \qquad G(x) = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}.$$

Pak platí

$$G(x)e = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

takže

$$D(x)e = \begin{bmatrix} \alpha_1, & 0 \\ 0, & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

což dává $\alpha_1 = -1$, $\alpha_2 = 4$, takže matice D není pozitivně definitní.

Je-li matice $D(x)$ pozitivně definitní, je pro dostatečně malou hodnotu $\varepsilon > 0$ i matice $D_\varepsilon(x)$ pozitivně definitní (plyne to ze spojité závislosti vlastních čísel matice na jejích prvcích). Nutné a postačující podmínky pro to, aby matice $D(x)$, definovaná vztahy (865) (které platí, jsou-li všechny diference stejné), byla pozitivně definitní udává tato věta.

Věta 210. *Diagonální matice $D(x)$, definovaná vztahy (865), je pozitivně definitní právě tehdy, má-li Hessova matice $G(x)$ kladné řádkové součty.*

Důkaz Tvrzení plyne bezprostředně z vyjádření (865) (i-tý diagonální prvek matice $D(x)$ je i-tým řádkovým součtem matice $G(x)$). □

Důsledek 30. *Je-li Hessova matice $G(x)$ pozitivně definitní a diagonálně dominantní, je diagonální matice $D(x)$, definovaná vztahy (865), pozitivně definitní.*

Důkaz Z vyjádření (865) a z toho, že Hessova matice je diagonálně dominantní, plynou nerovnosti

$$e_i^T D(x) e = \sum_{j=1}^{n} G_{ij} \geq G_{ii} - \sum_{j \neq i} |G_{ij}| > 0, \quad 1 \leq i \leq n.$$

□

Důsledek 31. *Má-li Hessova matice $G(x)$ kladné prvky a žádný její řádek není nulový, je diagonální matice $D(x)$, definovaná vztahy (865), pozitivně definitní.*

Důkaz Tvrzení plyne bezprostředně z věty 210 (řádkové součty jsou kladné). □

Jsou-li splněny předpoklady důsledku 31, vyhovuje matice $D(x)$ předpokladům následujícího tvrzení, dokázaného v [82], které jiným způsobem zdůvodňuje použití diagonálního předpodmiňovače $D(x)$.

Tvrzení 8. *Nechť $\mathcal{D}_n$ je množina všech diagonálních matic řádu n a $D = \operatorname{diag}(\alpha_1, \ldots, \alpha_n)$ je diagonální matice taková, že*

$$\alpha_i = \sum_{j=1}^{n} |G_{ij}|, \qquad 1 \leq j \leq n,$$

kde G_{ij}, $1 \leq j \leq n$, jsou prvky i-tého řádku matice G. Pak platí

$$\kappa_1(GD^{-1}) = \min_{M \in \mathcal{D}_n} \kappa_1(GM^{-1}),$$

kde κ_1 je l_1 číslo podmíněnosti (součin l_1 norem matice a její inverze).

Má-li matice G pouze kladné prvky a je-li $D(x)$ matice definovaná vztahy (865), platí

$$\alpha_i = e_i^T D(x) e = \sum_{j=1}^{n} G_{ij} = \sum_{j=1}^{n} |G_{ij}|, \quad 1 \le i \le n$$

a matice $C = D(x)$ je podle tvrzení 8 vhodným diagonálním předpodmiňovačem (optimálním v l_1 normě) pro soustavu rovnic $G(x)s + g = 0$. Nemá-li matice $G(x)$ pouze kladné prvky, můžeme místo (861) položit

$$\alpha_i = \frac{|e_i^T g'(x, v)|}{\delta_i}, \qquad 1 \le i \le n.$$

Pak platí

$$\alpha_i = \left| \sum_{j=1}^{n} G_{ij} \right| \le \sum_{j=1}^{n} |G_{ij}|,$$

takže prvky takto vytvořené diagonální matice jsou dolním odhadem prvků optimálního diagonálního předpodmiňovače. Tato úvaha je podkladem pro jednoduchou korekci diagonálního předpodmiňovače získaného numerickým derivováním. Je-li prvek α_i (určený podle (861)) záporný, změníme jeho znaménko. Tytéž úvahy platí pro matici $D_\varepsilon(x) \approx D(x)$.

Nyní ukážeme, jak lze směrové derivování použít ke konstrukci tridiagonálního předpodmiňovače. Abychom určili prvky tridiagonální Hessovy matice, stačí použít dva vektory druhých směrových derivací

$$G(x)v_1 = g'(x, v_1) \triangleq \lim_{\varepsilon \to 0} \frac{g(x + \varepsilon v_1) - g(x)}{\varepsilon}, \tag{866}$$

$$G(x)v_2 = g'(x, v_2) \triangleq \lim_{\varepsilon \to 0} \frac{g(x + \varepsilon v_2) - g(x)}{\varepsilon}, \tag{867}$$

kde $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \dots]^T$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \dots]^T$.

Věta 211. *Předpokládejme, že Hessova matice $G(x) = T(x)$ funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ je tridiagonální maticí tvaru (851). Položme $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \dots]^T$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \dots]^T$, kde $\delta_i > 0$, $1 \le i \le n$. Pak pro $2 \le i \le n-1$ platí*

$$\begin{aligned}
\alpha_1 &= \frac{e_1^T g'(x, v_1)}{\delta_1}, & \beta_1 &= \frac{e_1^T g'(x, v_2)}{\delta_2}, \\
\alpha_i &= \frac{e_i^T g'(x, v_1)}{\delta_i}, & \beta_i &= \frac{e_i^T g'(x, v_2)}{\delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, & &\mathrm{mod}(i,2) = 1, \\
\alpha_i &= \frac{e_i^T g'(x, v_2)}{\delta_i}, & \beta_i &= \frac{e_i^T g'(x, v_1)}{\delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, & &\mathrm{mod}(i,2) = 0, \\
\alpha_n &= \frac{e_n^T g'(x, v_1)}{\delta_n}, & &\mathrm{mod}(n,2) = 1, \\
\alpha_n &= \frac{e_n^T g'(x, v_2)}{\delta_n}, & &\mathrm{mod}(n,2) = 0.
\end{aligned}$$

Důkaz Použijeme-li vzorce (866), (867) a (851), můžeme psát

$$\begin{aligned}
\alpha_1\delta_1 &= e_1^T T(x)v_1 = e_1^T g'(x, v_1), & \beta_1\delta_2 &= e_1^T T(x)v_2 = e_1^T g'(x, v_2), & & \\
\alpha_i\delta_i &= e_i^T T(x)v_1 = e_i^T g'(x, v_1), & \beta_{i-1}\delta_{i-1} + \beta_i\delta_{i+1} &= e_i^T T(x)v_1 = e_i^T g'(x, v_2), & &\mathrm{mod}(i,2) = 1, \\
\alpha_i\delta_i &= e_i^T T(x)v_2 = e_i^T g'(x, v_2), & \beta_{i-1}\delta_{i-1} + \beta_i\delta_{i+1} &= e_i^T T(x)v_1 = e_i^T g'(x, v_1), & &\mathrm{mod}(i,2) = 0, \\
\alpha_n\delta_n &= e_n^T T(x)v_1 = e_n^T g'(x, v_1), & \mathrm{mod}(n,2) &= 1, & & \\
\alpha_n\delta_n &= e_n^T T(x)v_2 = e_n^T g'(x, v_2), & \mathrm{mod}(n,2) &= 0. & &
\end{aligned}$$

Z těchto vztahů plynou vzorce uvedené ve větě 211. □

Vektory (866) a (867) mohou být určeny přesně pomocí automatického derivování (oddíl 14). V tomto případě volíme vektory v_1 a v_2 tak, že $\delta_i = 1$, $1 \le i \le n$. Vektory (866) a (867) lze také určit přibližně pomocí numerického derivování. Pak pro aproximaci $T_\varepsilon(x)$ matice $T(x)$ platí

$$G(x)v_1 = g'(x, v_1) \approx T_\varepsilon(x)v_1 = \frac{g(x + \varepsilon v_1) - g(x)}{\varepsilon}, \tag{868}$$

$$G(x)v_2 = g'(x, v_2) \approx T_\varepsilon(x)v_2 = \frac{g(x + \varepsilon v_2) - g(x)}{\varepsilon}, \tag{869}$$

kde $\varepsilon > 0$ je vhodně zvolená hodnota (obvykle $\varepsilon \approx \sqrt{\varepsilon_M}$). Čísla $\delta_i > 0$, $1 \le i \le n$, se volí podle (863) nebo (864) (kde $\delta = \sqrt{2/n}$).

Poznámka 322. Pro prvky tridiagonální matice $T_\varepsilon(x) \approx T(x)$, vystupující ve vzorcích (868) a (869), platí

$$\begin{aligned}
\alpha_1 &= \frac{g_1(x + \varepsilon v_1) - g_1(x)}{\varepsilon\delta_1}, & \beta_1 &= \frac{g_1(x + \varepsilon v_2) - g_1(x)}{\varepsilon\delta_2}, \\
\alpha_i &= \frac{g_i(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_i} = \alpha_i, & \beta_i &= \frac{g_i(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 1, \\
\alpha_i &= \frac{g_i(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_i} = \alpha_i, & \beta_i &= \frac{g_i(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_{i+1}} - \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 0, \\
\alpha_n &= \frac{g_n(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_n}, & & \mathrm{mod}(n,2) = 1, \\
\alpha_n &= \frac{g_n(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_n}, & & \mathrm{mod}(n,2) = 0,
\end{aligned}$$

Poznamenejme, že tyto vzorce vyžadují dva další gradienty $g(x + \varepsilon v_1)$ a $g(x + \varepsilon v_2)$ v každém kroku Newtonovy metody.

Vzorce uvedené ve větě 211 a v poznámce 322 použijeme ke konstrukci tridiagonálních předpodmiňovačů $T(x)$ a $T_\varepsilon(x)$ a to i v případě, že Hessova matice $G(x)$ není tridiagonální, takže $G(x) \neq T(x)$. Nicméně pro vektory $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \dots]^T$ a $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \dots]^T$ i v tomto případě platí $G(x)v_1 = T(x)v_1$ a $G(x)v_2 = T(x)v_2$. Pokud $\delta_i = \delta$, $1 \le i \le n$ (takže všechny diference jsou stejné), lze prvky matice $T(x)$ vyjádřit ve tvaru

$$\alpha_i = \sum_{\mathrm{mod}(j,2)=1} G_{ij}, \qquad \beta_i + \beta_{i-1} = \sum_{\mathrm{mod}(j,2)=0} G_{ij}, \qquad \mathrm{mod}(i,2) = 1, \tag{870}$$

$$\alpha_i = \sum_{\mathrm{mod}(j,2)=0} G_{ij}, \qquad \beta_i + \beta_{i-1} = \sum_{\mathrm{mod}(j,2)=1} G_{ij}, \qquad \mathrm{mod}(i,2) = 0, \tag{871}$$

kde $\beta_0 = \beta_n = 0$. V dalších úvahách budeme předpokládat, že $T(x)$ je matice získaná podle věty 211 a $T_\varepsilon(x)$ je matice určená podle poznámky 322.

Tridiagonální matice $T(x)$ a $T_\varepsilon(x)$ nemusí být pozitivně definitní, i když Hessova matice $G(x)$ je pozitivně definitní a diagonálně dominantní.

Příklad 6. *Uvažujme ryze konvexní kvadratickou funkci $F : R^4 \to R$, kde*

$$F(x) = \frac{1}{2}x^T \begin{bmatrix} 7 & 0 & -2 & 4 \\ 0 & 7 & 0 & -2 \\ -2 & 0 & 7 & 0 \\ 4 & -2 & 0 & 7 \end{bmatrix} x, \qquad G(x) = \begin{bmatrix} 7 & 0 & -2 & 4 \\ 0 & 7 & 0 & -2 \\ -2 & 0 & 7 & 0 \\ 4 & -2 & 0 & 7 \end{bmatrix}.$$

Pak

$$G(x)v_1 = \begin{bmatrix} 7 & 0 & -2 & 4 \\ 0 & 7 & 0 & -2 \\ -2 & 0 & 7 & 0 \\ 4 & -2 & 0 & 7 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 5 \\ 0 \\ 5 \\ 4 \end{bmatrix},$$

$$G(x)v_2 = \begin{bmatrix} 7 & 0 & -2 & 4 \\ 0 & 7 & 0 & -2 \\ -2 & 0 & 7 & 0 \\ 4 & -2 & 0 & 7 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 4 \\ 5 \\ 0 \\ 5 \end{bmatrix},$$

kde $v_1 = [1, 0, 1, 0]^T$, $v_2 = [0, 1, 0, 1]^T$, *takže podle věty 211 platí*

$$T(x) = \begin{bmatrix} 5 & 4 & 0 & 0 \\ 4 & 5 & -4 & 0 \\ 0 & -4 & 5 & 4 \\ 0 & 0 & 4 & 5 \end{bmatrix}.$$

Tato matice není pozitivně definitní, neboť determinant

$$\det \begin{bmatrix} 5 & 4 & 0 \\ 4 & 5 & -4 \\ 0 & -4 & 5 \end{bmatrix} = 5\,(25 - 32) = -35$$

její hlavní submatice je záporný.

Uvedený příklad naznačuje, že neplatí analogie důsledku 30. Ukážeme však, že je-li Hessova matice pentadiagonální, analogie tohoto důsledku již platí.

Věta 212. *Nechť Hessova matice $G(x)$ je pentadiagonální, pozitivně definitní a diagonálně dominantní. Nechť vektory v_1 a v_2 jsou vybrány tak, že $\delta_i = \delta$, $1 \le i \le n$. Pak matice $T(x)$ je pozitivně definitní a diagonálně dominantní. Je-li číslo ε dostatečně malé, je matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní.*

Důkaz Uvažujme pentadiagonální Hessovu matici tvaru

$$G(x) = \begin{bmatrix} \tilde\alpha_1, & \tilde\beta_1, & \tilde\gamma_1, & \dots, & 0, & 0, & 0 \\ \tilde\beta_1, & \tilde\alpha_2, & \tilde\beta_2, & \dots, & 0, & 0, & 0 \\ \tilde\gamma_1, & \tilde\beta_2, & \tilde\alpha_3, & \dots, & 0, & 0, & 0 \\ --, & --, & --, & --, & --, & --, & \dots \\ 0, & 0, & 0, & \dots, & \tilde\alpha_{n-2}, & \tilde\beta_{n-2}, & \tilde\gamma_{n-2} \\ 0, & 0, & 0, & \dots, & \tilde\beta_{n-2}, & \tilde\alpha_{n-1}, & \tilde\beta_{n-1} \\ 0, & 0, & 0, & \dots, & \tilde\gamma_{n-2}, & \tilde\beta_{n-1}, & \tilde\alpha_n \end{bmatrix} \tag{872}$$

a pro zjednodušení zápisu položme $\tilde\gamma_{-1} = \tilde\gamma_0 = \tilde\beta_0 = \beta_0 = 0$, $\tilde\gamma_{n-1} = \tilde\gamma_n = \tilde\beta_n = \beta_n = 0$. Pak podle předpokladu diagonální dominance platí

$$\tilde\alpha_i > |\tilde\gamma_{i-2}| + |\tilde\beta_{i-1}| + |\tilde\beta_i| + |\tilde\gamma_i|$$

pro $1 \le i \le n$. Nechť $\delta_i = \delta$ pro $1 \le i \le n$. Z rovností $G(x)v_1 = T(x)v_1$ a $G(x)v_2 = T(x)v_2$, kde $G(x)$ je pentadiagonální matice tvaru (872) a $T(x)$ je tridiagonální matice tvaru (851), plyne, že pro i liché platí

$$\begin{aligned} \alpha_i = \frac{1}{\delta} e_i^T T(x) v_1 &= \frac{1}{\delta} e_i^T G(x) v_1 = \tilde\gamma_{i-2} + \tilde\alpha_i + \tilde\gamma_i, \\ \beta_{i-1} + \beta_i = \frac{1}{\delta} e_i^T T(x) v_2 &= \frac{1}{\delta} e_i^T G(x) v_2 = \tilde\beta_{i-1} + \tilde\beta_i, \end{aligned}$$

neboli

$$\alpha_i = \tilde{\gamma}_{i-2} + \tilde{\alpha}_i + \tilde{\gamma}_i, \qquad \beta_{i-1} + \beta_i = \tilde{\beta}_{i-1} + \tilde{\beta}_i. \tag{873}$$

Vyměníme-li vektory v_1 a v_2, dostaneme stejné rovnice pro i sudé. Platí tedy $\beta_i = \tilde{\beta}_i$, což spolu s (873) dává

$$\alpha_i - |\beta_{i-1}| - |\beta_i| = \tilde{\alpha}_i + \tilde{\gamma}_{i-2} + \tilde{\gamma}_i - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| \geq \tilde{\alpha}_i - |\tilde{\gamma}_{i-2}| - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| - |\tilde{\gamma}_i| > 0$$

pro $1 \leq i \leq n$. Z toho plyne, že symetrická tridiagonální matice $T(x)$ má kladné prvky na hlavní diagonále a je diagonálně dominantní. Podle Geršgorinovy věty (Geršgorinovy kruhy leží striktně napravo od imaginární osy) je tedy pozitivně definitní a diagonálně dominantní. Pro dostatečně malou hodnotu $\varepsilon > 0$ je i matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní (plyne to ze spojité závislosti vlastních čísel matice na jejích prvcích). □

Podmínka, aby pentadiagonální Hessova matice byla diagonálně dominantní, je velmi silná. Ukazuje se však, že pro mnohé praktické úlohy jsou matice $T(x)$ a $T_\varepsilon(x)$ pozitivně definitní, i když pentadiagonální Hessova matice není diagonálně dominantní.

Věta 213. *Nechť Hessova matice $G(x)$ je pentadiagonální tvaru (872), přičemž*

$$\begin{aligned} \tilde{\alpha}_1 \geq \psi_1^2 + 1, \quad \tilde{\alpha}_n \geq \psi_n^2 + 1, \quad \tilde{\alpha}_i \geq \psi_i^2 + 2, \qquad & 2 \leq i \leq n-1, \\ |\tilde{\beta}_i| \leq |\psi_i + \psi_{i+1}|, \qquad & 1 \leq i \leq n-1, \\ \tilde{\gamma}_i \geq 1, \qquad & 1 \leq i \leq n-2, \end{aligned} \tag{874}$$

kde ψ_i, $1 \leq i \leq n$, jsou libovolná reálná čísla taková, že alespoň jeden z výrazů $\psi_1\psi_2 - 2$, $\psi_i\psi_{i+1} - 4$, $2 \leq i \leq n-1$, $\psi_{n-1}\psi_n - 2$ je nenulový. Nechť vektory v_1 a v_2 jsou zvoleny tak, že $\delta_i = \delta$, $1 \leq i \leq n$. Pak matice $T(x)$ je pozitivně definitní. Je-li číslo $\varepsilon > 0$ dostatečně malé, je i matice $T_\varepsilon(x)$ pozitivně definitní.

Důkaz Je-li Hessova matice $G(x)$ pentadiagonální tvaru (872), můžeme prvky tridiagonální matice $T(x)$ vyjádřit podle vzorců (873) (kde $\tilde{\gamma}_{-1} = \tilde{\gamma}_0 = \tilde{\beta}_0 = \beta_0 = 0$ a $\tilde{\gamma}_{n-1} = \tilde{\gamma}_n = \tilde{\beta}_n = \beta_n = 0$). Dosadíme-li nerovnosti (874) do vzorců (873), můžeme psát

$$\begin{aligned} \alpha_1 = \tilde{\gamma}_{-1} + \tilde{\alpha}_1 + \tilde{\gamma}_1 &\geq \psi_1^2 + 2, \\ \alpha_2 = \tilde{\gamma}_0 + \tilde{\alpha}_2 + \tilde{\gamma}_2 &\geq \psi_2^2 + 3, \\ \alpha_i = \tilde{\gamma}_{i-2} + \tilde{\alpha}_i + \tilde{\gamma}_i &\geq \psi_i^2 + 4, \qquad 3 \leq i \leq n-2, \\ \alpha_{n-1} = \tilde{\gamma}_{n-3} + \tilde{\alpha}_{n-1} + \tilde{\gamma}_{n-1} &\geq \psi_{n-1}^2 + 3, \\ \alpha_n = \tilde{\gamma}_{n-2} + \tilde{\alpha}_n + \tilde{\gamma}_n &\geq \psi_n^2 + 2, \end{aligned}$$

a $\beta_i = \tilde{\beta}_i$, $|\tilde{\beta}_i| \leq |\psi_i + \psi_{i+1}|$, $1 \leq i \leq n-1$. Nyní využijeme toho, že výraz $2v^T T(x)v$ lze podobně jako v

(855) zapsat ve tvaru

$$\begin{aligned}
2v^T T(x)v &= [v_1,v_2]\begin{bmatrix} 2\alpha_1, & 2\beta_1 \\ 2\beta_1, & \alpha_2-1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&+ [v_2,v_3]\begin{bmatrix} \alpha_2+1, & 2\beta_2 \\ 2\beta_2, & \alpha_3 \end{bmatrix}\begin{bmatrix} v_2 \\ v_3 \end{bmatrix} \\
&+ \sum_{i=3}^{n-3}[v_i,v_{i+1}]\begin{bmatrix} \alpha_i, & 2\beta_i \\ 2\beta_i, & \alpha_{i+1} \end{bmatrix}\begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\
&+ [v_{n-2},v_{n-1}]\begin{bmatrix} \alpha_{n-2}, & 2\beta_{n-2} \\ 2\beta_{n-2}, & \alpha_{n-1}+1 \end{bmatrix}\begin{bmatrix} v_{n-2} \\ v_{n-1} \end{bmatrix} \\
&+ [v_{n-1},v_n]\begin{bmatrix} \alpha_{n-1}-1, & 2\beta_{n-1} \\ 2\beta_{n-1}, & 2\alpha_n \end{bmatrix}\begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} \\
&\geq [v_1,v_2]\begin{bmatrix} 2(\psi_1^2+2), & 2\tilde{\beta}_1 \\ 2\tilde{\beta}_1, & \psi_2^2+2 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&+ \sum_{i=2}^{n-2}[v_i,v_{i+1}]\begin{bmatrix} \psi_i^2+4, & 2\tilde{\beta}_i \\ 2\tilde{\beta}_i, & \psi_{i+1}^2+4 \end{bmatrix}\begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\
&+ [v_{n-1},v_n]\begin{bmatrix} \psi_{n-1}^2+2, & 2\tilde{\beta}_{n-1} \\ 2\tilde{\beta}_{n-1}, & 2(\psi_n^2+2) \end{bmatrix}\begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix}. \qquad (875)
\end{aligned}$$

Jelikož

$$\begin{aligned}
2(\psi_i^2+2)(\psi_{i+1}^2+2)-4\beta_i^2 &\geq 2(\psi_i^2+2)(\psi_{i+1}^2+2)-4(\psi_i+\psi_{i+1})^2 = 2\psi_i^2\psi_{i+1}^2+8-8\psi_i\psi_{i+1} \\
&= 2(\psi_i\psi_{i+1}-2)^2 \geq 0, \qquad i\in\{1,n-1\}, \\
(\psi_i^2+4)(\psi_{i+1}^2+4)-4\beta_i^2 &\geq (\psi_i^2+4)(\psi_{i+1}^2+4)-4(\psi_i+\psi_{i+1})^2 = \psi_i^2\psi_{i+1}^2+16-8\psi_i\psi_{i+1} \\
&= (\psi_i\psi_{i+1}-4)^2 \geq 0, \qquad 2\leq i\leq n-2,
\end{aligned}$$

mají všechny matice na pravé straně součtu (875) kladné diagonální prvky a jsou pozitivně semidefinitní. Je-li alespoň jeden z výrazů $\psi_1\psi_2-2$, $\psi_i\psi_{i+1}-4$, $2\leq i\leq n-1$, $\psi_{n-1}\psi_n-2$ nenulový, je matice $T(x)$ pozitivně definitní podle poznámky 317. Jelikož vlastní čísla symetrické matice závisí spojitě na jejích prvcích, je pro dostatečne malé číslo $\varepsilon>0$ i matice $T_\varepsilon(x)$ pozitivně definitní. □

Příklad 7. Uvažujme okrajovou úlohu pro obyčejnou diferenciální rovnici druhého řádu

$$y''(t)=\varphi(y(t)), \quad 0\leq t\leq 1, \quad y(0)=\bar{y}_0, \quad y(1)=\bar{y}_1,$$

kde funkce $\varphi: R\to R$ je dvakrát spojitě diferencovatelná na R. Rozdělíme-li interval $[0,1]$ na $n+1$ částí pomocí uzlových bodů $t_i=ih$, $0\leq i\leq n+1$, kde $h=1/(n+1)$ je krok sítě, a nahradíme-li druhé derivace v uzlových bodech diferencemi

$$y''(t_i)=\frac{y(t_{i-1})-2y(t_i)+y(t_{i+1})}{h^2},$$

kde $1\leq 1\leq n$, dostaneme soustavu n nelineárních rovnic

$$f_i(x)\triangleq h^2\varphi(x_i)+2x_i-x_{i-1}-x_{i+1}=0, \qquad (876)$$

kde $x_i=y(t_i)$, $0\leq 1\leq n+1$, takže $x_0=\bar{y}_0$ a $x_{n+1}=\bar{y}_1$. Řešíme-li tuto soustavu metodou nejmenších čtverců (kapitola 8), má minimalizovaná funkce tvar

$$F(x)=\frac{1}{2}f^T(x)f(x)=\frac{1}{2}\sum_{i=1}^{n}f_i^2(x)=\frac{1}{2}\sum_{i=1}^{n}\left(h^2\varphi(x_i)+2x_i-x_{i-1}-x_{i+1}\right)^2, \qquad (877)$$

kde $x = [x_1, \ldots, x_n]^T$ a $f = [f_1, \ldots, f_n]^T$. Derivujeme-li pouze podle prvků x_{i-1}, x_i, x_{i+1}, dostaneme

$$\nabla f_i(x) = \begin{bmatrix} -1 \\ \psi(x_i) \\ -1 \end{bmatrix}, \qquad \nabla^2 f_i(x) = \begin{bmatrix} 0, & 0, & 0 \\ 0, & \psi'(x_i), & 0 \\ 0, & 0, & 0 \end{bmatrix},$$

kde $\psi(x_i) = 2 + h^2\varphi'(x_i)$ a $\psi'(x_i) = h^2\varphi''(x_i)$. Pro součet čtverců lze Hessovu matici vyjádřit ve tvaru $G(x) = J^T(x)J(x) + C(x)$ (viz (651)). Omezíme-li se pro jednoduchost na submatice řádu pět, můžeme psát

$$J(x) = \begin{bmatrix} \psi_1, & -1, & 0, & 0, & 0 \\ -1, & \psi_2, & -1, & 0, & 0 \\ 0, & -1, & \psi_3, & -1, & 0 \\ 0, & 0, & -1, & \psi_4, & -1 \\ 0, & 0, & 0, & -1, & \psi_5 \end{bmatrix}, \quad C(x) = \begin{bmatrix} f_1\psi_1', & 0, & 0, & 0, & 0 \\ 0, & f_2\psi_2', & 0, & 0, & 0 \\ 0, & 0, & f_3\psi_3', & 0, & 0 \\ 0, & 0, & 0, & f_4\psi_4', & 0 \\ 0, & 0, & 0, & 0, & f_5\psi_5' \end{bmatrix},$$

$$J^T(x)J(x) = \begin{bmatrix} \psi_1^2 + 1, & -(\psi_1 + \psi_2), & 1, & 0, & 0 \\ -(\psi_1 + \psi_2), & \psi_2^2 + 2, & -(\psi_2 + \psi_3), & 1, & 0 \\ 1, & -(\psi_2 + \psi_3), & \psi_3^2 + 2, & -(\psi_3 + \psi_4), & 1 \\ 0, & 1, & -(\psi_3 + \psi_4), & \psi_4^2 + 2, & -(\psi_4 + \psi_5) \\ 0, & 0, & 1, & -(\psi_4 + \psi_5), & \psi_5^2 + 1 \end{bmatrix},$$

kde $\psi_i = \psi(x_i)$, $\psi_i' = \psi'(x_i)$, $1 \le i \le n$, odkud vidíme, že Hessova matice $G(x) = J^T(x)J(x) + C(x)$ je pentadiagonální. Je-li funkce $\varphi : R \to R$ lineární (takže $\varphi'(x_i) = \varphi'$, $\varphi''(x_i) = 0$, $1 \le i \le n$, kde φ' je konstantní směrnice lineární funkce φ), platí $C(x) = 0$, takže $G(x) = J^T(x)J(x)$. Vrátíme-li se k případu, kdy dimenze n je libovolná, a předpokládáme-li, že funkce φ je lineární, můžeme usoudit, že matice $G(x) = J^T(x)J(x)$ má tvar (872) a její prvky splňují (874), kde nerovnosti přejdou v rovnosti. Tato matice je pentadiagonální, ale není diagonálně dominantní. Pokud $h^2|\varphi'| < 2$, platí

$$\begin{aligned} \tilde{\alpha}_i - |\tilde{\gamma}_{i-2}| - |\tilde{\beta}_{i-1}| - |\tilde{\beta}_i| - |\tilde{\gamma}_i| &= \psi_i^2 + 2 - 2 - \psi_{i-1} - 2\psi_i - \psi_{i+1} \\ &= (2 + h^2\varphi')^2 - 4(2 + h^2\varphi') = (h^2\varphi')^2 - 4 < 0 \end{aligned}$$

pro $3 \le i \le n-2$. Nemůžeme tedy použít větu 212. Jelikož funkce φ je lineární, platí $\psi_i = 2 + h^2\varphi'$, $1 \le i \le n$. Jestliže $(2 + h^2\varphi')^2 \ne 2$, lze psát $\psi_1\psi_2 - 2 \ne 0$, $\psi_{n-1}\psi_n - 2 \ne 0$. Jestliže $(2 + h^2\varphi')^2 \ne 4$, lze psát $\psi_i\psi_{i+1} - 4 \ne 0$, $2 \le i \le n-1$. Pokud $\delta_i = \delta$, $1 \le i \le n$, jsou splněny předpoklady věty 213 a matice $T(x)$ je pozitivně definitní. Je-i číslo $\varepsilon > 0$ dostatečně malé, je i matice $T_\varepsilon(x)$ pozitivně definitní.

Poznámka 323. Pokud se blížíme k minimu, kde $F(x) = 0$ (což odpovídá řešení soustavy rovnic (876)), platí $f_i \approx 0$, $1 \le i \le n$. Navíc absolutní hodnoty prvků matice $\mathrm{diag}(\psi_1', \ldots, \psi_n')$ jsou obvykle malé ve srovnání s absolutními hodnotami diagonálních prvků matice $J(x)^T J(x)$ (pokud $n \approx 1000$, je $h^2 \approx 10^{-6}$). Protože malá změna diagonálních prvků neporuší pozitivní definitnost matice (poznámka 316), můžeme očekávat, že v dostatečné blízkosti řešení je matice $T(x)$ (a tedy i matice $T_\varepsilon(x)$, je-li číslo $\varepsilon > 0$ dostatečně malé) pozitivně definitní, i když funkce $\varphi : R \to R$ není lineární.

V příkladu 6 jsme použili diagonálně dominantní toeplitzovskou matici sudého řádu. Je zajímavé, že je-li Hessova matice diagonálně dominantní toeplitzovskou maticí lichého řádu a platí-li $\delta_i = \delta$, $1 \le i \le n$, jsou matice $T(x)$ a $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní. Prvky toeplitzovské matice $G(x)$ označíme symboly c_i, $1 \le i \le n$, tedy

$$G(x) = \begin{bmatrix} c_1 & c_2, & c_3, & \ldots, & c_{n-2}, & c_{n-1}, & c_n \\ c_2 & c_1, & c_2, & \ldots, & c_{n-3}, & c_{n-2}, & c_{n-1} \\ c_3 & c_2, & c_1, & \ldots, & c_{n-4}, & c_{n-3}, & c_{n-2} \\ -- & -- & -- & -- & -- & -- & \ldots \\ c_{n-2}, & c_{n-3}, & c_{n-4}, & \ldots. & c_1, & c_2, & c_3 \\ c_{n-3}, & c_{n-2}, & c_{n-3}, & \ldots. & c_2, & c_1, & c_2 \\ c_n, & c_{n-1}, & c_{n-2}, & \ldots. & c_3, & c_2, & c_1 \end{bmatrix}. \tag{878}$$

Věta 214. *Nechť Hessova matice $G(x)$ je diagonálně dominantní toeplitzovskou maticí lichého řádu s kladnými prvky na hlavní diagonále. Pak, platí-li $\delta_i = \delta$, $1 \le i \le n$, je matice $T(x)$ pozitivně definitní a diagonálně dominantní. Je-li číslo $\varepsilon > 0$ dostatečně malé, je i matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní.*

Důkaz (a) Z rovností $G(x)v_1 = T(x)v_1$ a $G(x)v_2 = T(x)v_2$, plynou pro n liché vztahy

$$\begin{aligned}
\beta_1 = \frac{1}{\delta} e_1^T T(x) v_2 &= \frac{1}{\delta} e_1^T G(x) v_2 = \sum_{j=1}^{\frac{n-1}{2}} c_{2j}, \\
\beta_1 + \beta_2 = \frac{1}{\delta} e_2^T T(x) v_1 &= \frac{1}{\delta} e_2^T G(x) v_1 = c_2 + \sum_{j=1}^{\frac{n-1}{2}} c_{2j} \quad \Rightarrow \quad \beta_2 = c_2, \\
\beta_2 + \beta_3 = \frac{1}{\delta} e_3^T T(x) v_2 &= \frac{1}{\delta} e_3^T G(x) v_2 = c_2 + \sum_{j=1}^{\frac{n-3}{2}} c_{2j} \quad \Rightarrow \quad \beta_3 = \sum_{j=1}^{\frac{n-3}{2}} c_{2j}, \\
\beta_3 + \beta_4 = \frac{1}{\delta} e_4^T T(x) v_1 &= \frac{1}{\delta} e_4^T G(x) v_1 = c_2 + c_4 + \sum_{j=1}^{\frac{n-3}{2}} c_{2j} \quad \Rightarrow \quad \beta_4 = c_2 + c_4, \\
\beta_4 + \beta_5 = \frac{1}{\delta} e_5^T T(x) v_2 &= \frac{1}{\delta} e_5^T G(x) v_2 = c_2 + c_4 + \sum_{j=1}^{\frac{n-5}{2}} c_{2j} \quad \Rightarrow \quad \beta_5 = \sum_{j=1}^{\frac{n-5}{2}} c_{2j},
\end{aligned}$$

a tak dále, takže

$$\beta_i = \sum_{j=1}^{\frac{n-i}{2}} c_{2j}, \quad \mathrm{mod}(j,2) = 1, \qquad \beta_i = \sum_{j=1}^{\frac{i}{2}} c_{2j}, \quad \mathrm{mod}(j,2) = 0.$$

Odtud pro i liché dostaneme

$$|\beta_{i-1}| + |\beta_i| = \left| \sum_{j=1}^{\frac{i-1}{2}} c_{2j} \right| + \left| \sum_{j=1}^{\frac{n-i}{2}} c_{2j} \right| \le \sum_{j=1}^{\frac{i-1}{2}} |c_{2j}| + \sum_{j=1}^{\frac{n-i}{2}} |c_{2j}|,$$

kde na pravé straně je součet absolutních hodnot prvků se sudými indexy, které se nacházejí v i-tém řádku matice $G(x)$. Stejný výsledek dostaneme i pro i sudé.

(b) Z rovností $G(x)v_1 = T(x)v_1$ a $G(x)v_2 = T(x)v_2$ plyne, že diagonální prvek α_i matice $T(x)$ se rovná součtu prvků s lichými indexy, které se nacházejí v i-tém řádku matice $G(x)$. Spojíme-li tuto skutečnost s výsledkem uvedeným v (a) vidíme, že číslo $\alpha_i - |\beta_{i-1}| - |\beta_i|$ není větší než číslo, které dostaneme, odečteme-li od c_1 součet absolutních hodnot všech ostatních prvků nacházejících se v i-tém řádku matice $G(x)$. Jelikož matice $G(x)$ je diagonálně dominantní a $c_i > 0$, je toto číslo kladné. Matice $T(x)$ je tedy diagonálně dominantní a má kladné prvky na hlavní diagonále. Podle Geršgorinovy věty je tedy pozitivně definitní a diagonálně dominantní. Pro dostatečně malou hodnotu $\varepsilon > 0$ je i matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní. □

Ve větách týkajících se tridiagonálních předpodmiňovačů získaných směrovým derivováním jsme předpokládali, že všechny diference jsou stejné, tedy $\delta_i = \delta$, $1 \le i \le n$. Tento předpoklad je vždy splněn, používáme-li automatické derivování. Při numerickém derivování je však výhodnější volit diference podle (864). Ukážeme, že pokud $\tilde{\gamma}_i \ge 0$, $1 \le i \le n-2$, lze ve větě 212 vynechat požadavek diagonální dominance a použít libovolné diference $\delta_i > 0$, $1 \le i \le n$ (například $\delta_i = \max(|x_i|, 1)$, $1 \le i \le n$).

Věta 215. *Nechť Hessova matice $G(x)$ je pentadiagonální a má nezáporné prvky ve vnějších diagonálách (tedy $\tilde{\gamma}_i \ge 0$, $1 \le i \le n-2$). Nechť tridiagonální matice, která vznikne z $G(x)$ vynulováním těchto*

nezáporných prvků, je pozitivně definitní a diagonálně dominantní. Pak matice $T(x)$ je pozitivně definitní a diagonálně dominantní. Je-i číslo $\varepsilon > 0$ dostatečně malé, je i matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní.

Důkaz Nejsou-li diference δ_i, $1 \le i \le n$, stejné, implikují rovnosti $G(x)v_1 = T(x)v_1$ a $G(x)v_2 = T(x)v_2$ vztahy

$$\begin{aligned} \alpha_i\delta_i = e_i^T T(x)v_1 &= e_i^T G(x)v_1 = \tilde\gamma_{i-2}\delta_{i-2} + \tilde\alpha_i\delta_i + \tilde\gamma_i\delta_{i+2}, \\ \beta_{i-1}\delta_{i-1} + \beta_i\delta_{i+1} = e_i^T T(x)v_2 &= e_i^T G(x)v_2 = \tilde\beta_{i-1}\delta_{i-1} + \tilde\beta_i\delta_{i+1}, \end{aligned}$$

neboli

$$\alpha_i = \tilde\gamma_{i-2}\frac{\delta_{i-2}}{\delta_i} + \tilde\alpha_i + \tilde\gamma_i\frac{\delta_{i+2}}{\delta_i}, \qquad \beta_{i-1}\frac{\delta_{i-1}}{\delta_i} + \beta_i\frac{\delta_{i+1}}{\delta_i} = \tilde\beta_{i-1}\frac{\delta_{i-1}}{\delta_i} + \tilde\beta_i\frac{\delta_{i+1}}{\delta_i},$$

pro i liché (zde $\tilde\gamma_{-1} = \tilde\gamma_0 = \tilde\beta_0 = \beta_0 = 0$ a $\tilde\gamma_{n-1} = \tilde\gamma_n = \tilde\beta_n = \beta_n = 0$). Vyměníme-li vektory v_1 a v_2, dostaneme stejné rovnice pro i sudé. Platí tedy $\alpha_i \ge \tilde\alpha_i$, $1 \le i \le n$, a $\beta_i = \tilde\beta_i$, $1 \le i \le n-1$. Z pozitivní definitnosti a diagonální dominance tridiagonální matice s prvky $\tilde\alpha_i > 0$, $1 \le i \le n$, a $\tilde\beta_i$, $1 \le i \le n-1$, plyne, že

$$\alpha_i - |\beta_{i-1}| - |\beta_i| \ge \tilde\alpha_i - |\tilde\beta_{i-1}| - |\tilde\beta_i| > 0,$$

$1 \le i \le n$ (kde $\tilde\beta_0 = \tilde\beta_n = 0$). Odtud plyne, že matice $T(x)$ má kladné prvky na hlavní diagonále a je diagonálně dominantní. Podle Geršgorinovy věty je tedy pozitivně definitní a diagonálně dominantní. Pro dostatečně malou hodnotu $\varepsilon > 0$ je i matice $T_\varepsilon(x)$ pozitivně definitní a diagonálně dominantní. □

Je-li Hessova matice pentadiagonální a pozitivně definitní, je výhodné použít pentadiagonální předpodmiňovač vyšetřovaný v následující větě, jejíž důkaz je velmi podobný důkazu věty 211.

Věta 216. *Předpokládejme, že Hessova matice $G(x) = P(x)$ funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ je pentadiagonální maticí tvaru (856). Položme $v_1 = [\delta_1, 0, 0, \delta_4, 0, 0, \dots]$, $v_2 = [0, \delta_2, 0, 0, \delta_5, 0, \dots]$, $v_3 = [0, 0, \delta_3, 0, 0, \delta_6, \dots]$, kde $\delta_i > 0$, $1 \le i \le n$. Pak platí*

$$\begin{aligned}
&\alpha_i = \frac{e_i^T g'(x, v_1)}{\delta_i}, \quad \beta_i = \frac{e_i^T g'(x, v_2)}{\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \quad \gamma_i = \frac{e_i^T g'(x, v_3)}{\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 1, \\
&\alpha_i = \frac{e_i^T g'(x, v_2)}{\delta_i}, \quad \beta_i = \frac{e_i^T g'(x, v_3)}{\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \quad \gamma_i = \frac{e_i^T g'(x, v_1)}{\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 2, \\
&\alpha_i = \frac{e_i^T g'(x, v_3)}{\delta_i}, \quad \beta_i = \frac{e_i^T g'(x, v_1)}{\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \quad \gamma_i = \frac{e_i^T g'(x, v_2)}{\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 0,
\end{aligned}$$

kde veličiny s indexy $i < 1$ považujeme za nulové a veličiny, v jejichž vzorcích vystupují indexy $i > n$, nepočítáme.

Poznámka 324. V případě numerického derivování můžeme použít stejný postup jako v poznámce 322 a vyjádřit prvky pentadiagonálního předpodmiňovače $P_\varepsilon(x) \approx P(x)$ ve tvaru

$$\begin{aligned}
&\alpha_i = \frac{g_i(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_i}, \quad \beta_i = \frac{g_i(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \\
&\qquad\qquad \gamma_i = \frac{g_i(x + \varepsilon v_3) - g_i(x)}{\varepsilon\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 1, \\
&\alpha_i = \frac{g_i(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_i}, \quad \beta_i = \frac{g_i(x + \varepsilon v_3) - g_i(x)}{\varepsilon\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \\
&\qquad\qquad \gamma_i = \frac{g_i(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 2, \\
&\alpha_i = \frac{g_i(x + \varepsilon v_3) - g_i(x)}{\varepsilon\delta_i}, \quad \beta_i = \frac{g_i(x + \varepsilon v_1) - g_i(x)}{\varepsilon\delta_{i+1}} - \frac{\delta_{i-2}}{\delta_{i+1}}\gamma_{i-2}, \\
&\qquad\qquad \gamma_i = \frac{g_i(x + \varepsilon v_2) - g_i(x)}{\varepsilon\delta_{i+2}} - \frac{\delta_{i-1}}{\delta_{i+2}}\beta_{i-1}, \quad \mathrm{mod}(i,3) = 0.
\end{aligned}$$

Poznámka 325. Je-li matice $G(x)$ pentadiagonální a pozitivně definitní, je pentadiagonální předpodmiňovač $P(x)$ získaný podle věty 324 ideální v tom smyslu, že metoda sdružených gradientů najde řešení soustavy rovnic $G(s)s + g = 0$ již v prvním iteračním kroku. Není-li předpodmiňovač $P(x)$ pozitivně definitní, lze ho upravit podle pravidel uvedených v poznámce 319. V obecném případě je jistým nedostatkem tohoto předpodmiňovače nutnost počítat v každém kroku Newtonovy metody tři gradienty navíc (tridiagonální předpodmiňovač vyžaduje dva gradienty navíc a diagonální předpodmiňovač pouze jeden gradient navíc).

Známe-li analytické vzorce pro prvky ležící na hlavní diagonále, případně na několika vedlejších diagonálách Hessovy matice (což bývá splněno zejména u toeplitzovských matic), můžeme vypočítat prvky pásového předpodmiňovače analyticky. Tak jako v případě pásových matic určených standardní metodou BFGS odvozenou z předpodmíněné metody sdružených gradientů, nemusí být získaný předpodmiňovač pozitivně definitní. V tomto případě lze použít pravidla uvedená v poznámkách 318 a 319.

Nyní se budeme zabývat posledním typem předpodmínění zmíněným na začátku tohoto oddílu. Ukážeme, jak lze tridiagonální matici (594), získanou podle věty 142, použít ke konstrukci předpodmiňovače pro metodu sdružených gradientů. Ve větě 142 jsme ukázali, že symetrický Lanczosův proces je úzce spjatý s metodou sdružených gradientů a uvedli jsme převodní vztahy $\gamma_1 = 1/\alpha_1$ a

$$\delta_i^2 = \frac{\beta_i}{\alpha_i^2}, \qquad \gamma_{i+1} = \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}} \tag{879}$$

pro $1 \leq i \leq m$, kde m je číslo takové, že $\alpha_i > 0$ pro $1 \leq i \leq m$.

Věta 217. *Uvažujme metodu sdružených gradientů (aplikovanou na kvadratickou funkci s Hessovou maticí G) takovou, že $\alpha_i > 0$ pro $1 \leq i \leq m$. Pak tridiagonální matice T_m tvaru (594) řádu m získaná podle vzorců (879) je pozitivně definitní.*

Důkaz Označme $\tau_i = 1/\alpha_i$, $1 \leq i \leq m$. Pak $\tau_1 = \gamma_1$ a podle (879) platí

$$\delta_i^2 = \beta_i \tau_i^2, \qquad \gamma_{i+1} = \beta_i \tau_i + \tau_{i+1}$$

pro $1 \leq i \leq m$, takže

$$\tau_{i+1} = \gamma_{i+1} - \frac{\delta_i^2}{\tau_i}$$

pro $1 \leq i \leq m$, což je právě rekurentní vztah (853) (kde místo α_{i+1}, β_i píšeme γ_{i+1}, δ_i). Jelikož čísla $\tau_i = 1/\alpha_i$, $1 \leq i \leq m$, jsou podle předpokladu kladná, je podle věty 206 matice T_m pozitivně definitní. □

Poznámka 326. Tridiagonální matice T_m má dimenzi $m \leq n$. Abychom dostali předpodmiňovač dimenze n, položíme

$$C = (I - Q_m Q_m^T) + Q_m T_m Q_m^T \tag{880}$$

kde Q_m je matice s m ortonormálními sloupci získaná symetrickým Lanczosovým procesem (poznámka 227 a věta 142). Abychom zdůvodnili tuto volbu, ukážeme, že platí

$$C = [Q_m, Q_{n-m}] \begin{bmatrix} T_m, & 0 \\ 0, & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T, \tag{881}$$

kde Q_{n-m} je matice s $n - m$ ortonormálními sloupci taková, že matice $[Q_m, Q_{n-m}]$ je čtvercová a ortogonální. Nechť $v = v_1 + v_2$, kde $v_1 = Q_m \tilde{v_1}$ a $v_2 = Q_{n-m} \tilde{v_2}$ (takže $Q_{m-n}^T v_1 = 0$ a $Q_m^T v_2 = 0$). Pak platí

$$\left((I - Q_m Q_m^T) + Q_m T_m Q_m^T\right) v = Q_m T_m \tilde{v_1} + v_2 \tag{882}$$

a

$$[Q_m, Q_{n-m}] \begin{bmatrix} T_m, & 0 \\ 0, & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T v = [Q_m, Q_{n-m}] \begin{bmatrix} T_m, & 0 \\ 0 & I_{n-m} \end{bmatrix} [\tilde{v_1}, \tilde{v_2}]^T$$
$$= Q_m T_m \tilde{v_1} + Q_{n-m} \tilde{v_2} = Q_m T_m \tilde{v_1} + v_2$$

a jelikož vektor v lze volit libovolně, jsou obě matice stejné. Dále podle (882) platí

$$Cv = Q_m T_m \tilde{v_1} + v_2 = Q_m (Q_m^T G Q_m) \tilde{v_1} + v_2 = Q_m Q_m^T G v_1 + v_2$$

(neboť $T_m = Q_m^T G Q_m$ podle poznámky 228), takže předpodmiňovač C působí na složku v_1 jako matice G následovaná projekcí do $\mathcal{L}(Q_m)$ a na složku v_2 jako jednotková matice.

Věta 218. *Nechť jsou splněny předpoklady věty 217. Pak předpodmiňovač (880) je pozitivně definitní a platí*

$$C^{-1} = (I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T. \tag{883}$$

Důkaz (a) Jelikož $Q_m^T Q_m = I$, je matice $(I - Q_m Q_m^T)$ idempotentní a tudíž pozitivně semidefinitní (pro libovolný vektor v platí $v^T(I - Q_m Q_m^T)v = v^T(I - Q_m Q_m^T)(I - Q_m Q_m^T)v \geq 0$). Jelikož matice T_m je podle věty 217 pozitivně definitní a matice Q_m má lineárně nezávislé sloupce, je matice $Q_m T_m Q_m^T$ pozitivně definitní. Matice C je tedy (jako součet pozitivně semidefinitní a pozitivně definitní matice) pozitivně definitní.

(b) Jelikož $Q_m^T Q_m = I$ a matice $I - Q_m Q_m^T$ je idempotentní, platí

$$\left[(I - Q_m Q_m^T) + Q_m T_m Q_m^T\right] \left[(I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T\right]$$
$$= I - Q_m Q_m^T + Q_m T_m Q_m^T - Q_m T_m Q_m^T + Q_m T_m^{-1} Q_m^T - Q_m T_m^{-1} Q_m^T + Q_m Q_m^T = I.$$

□

Nevýhoda předpodmiňovače (880) spočívá v tom, že tuto matici lze definovat pouze v nepředpodmíněném kroku Newtonovy metody. Pokud krok Newtonovy metody předpodmíníme, nejsou sloupce matice Q_m ortonormální (poznámka 233) a matice (880) nemá požadované vlastnosti. Abychom se těmto potížím vyhnuli, museli bychom použitý předpodmiňovač zahrnout do výrazu (880) (místo jednotkové matice). To znamená, že bychom museli ukládat předpodmiňovače ze všech předchozích kroků, což je nepraktické. Proto se postupuje tak, že se provede $m \ll n$ kroků nepředpodmíněné metody sdružených gradientů, zkonstruuje se předpodmiňovač (880) a ten se použije v dalších krocích metody sdružených gradientů (také se lze vrátit na začátek iteračního procesu).

Je třeba se také zmínit o způsobu, který nám dovolí rozhodnout, zda máme předpodmiňovač použít nebo odmítnout (ne vždy je nalezený předpodmiňovač vhodný k použití). To se týká hlavně pásových předpodmiňovačů určených metodou BFGS nebo numerickým derivováním, které mohou být indefinitní. Je třeba zdůraznit, že indefinitní předpodmiňovač je nevhodný i v případě, že Hessova matice není pozitivně definitní, neboť v takovém případě není účelné hledat řešení soustavy $Gs + g = 0$, které charakterizuje sedlový bod a ne minimum. K testování pozitivní definitnosti a špatné podmíněnosti matice se hodí Gillův–Murrayův rozklad popsaný v oddílu 2.7. Pokud v některém eliminačním kroku je pivot menší než $\delta \max(1, \max_{1 \leq i \leq n}(|\alpha_i|))$, kde δ je předepsaná mez, rozklad předpodmiňovače ukončíme a předpodpodmiňovač odmítneme. Provést Gillův-Murrayův rozklad do konce a použít získanou pozitivně definitní matici jako předpodmiňovač se nevyplácí (dokládají to numerické experimenty). Číslo δ se obvykle volí tak, že $\delta = 10^{-12}$. Někdy je však třeba zvolit větší hodnotu (například $\delta = 10^{-2}$).

Poznámka 327. Závěrem uvedeme několik poznámek ke konstrukci předpodmiňovačů pro vektorové diferenční verze Newtonovy metody.

- Předpodmiňovače založené na metodách s proměnnou metrikou s omezenou pamětí nevyžadují žádné korekce. Jsou poměrně robustní, ale nejsou nejefektivnější.

- Pásové předpodmiňovače určené standardní metodou BFGS odvozenou z předpodmíněné metody sdružených gradientů je třeba předem upravit, jinak jsou při určování Gillova–Murrayova rozkladu většinou odmítnuty. Velmi se osvědčily úpravy založené na větě 207, kdy se nevhodné mimodiagonální prvky zmenšují tak, aby se záporné determinanty matic (854) vynulovaly, a úpravy založené na větě 208, kdy se indefinitní matice (857) upravují tak, že se prvky β_i, β_{i+1} zmenší na dvě třetiny a prvek γ_i na třetinu. Použití věty 209 a vzorce (859) je méně výhodné. Ukazuje se, že takto získané předpodmiňovače je třeba častěji odmítat (například volbou $\delta = 10^{-2}$).
- Pásové předpodmiňovače získané směrovým derivováním stačí upravit tak, že diagonální prvky nahradíme jejich absolutními hodnotami. Úpravy založené na větách 207 a 208 snižují efektivitu předpodmínění. Pro odmítání stačí většinou volba $\delta = 10^{-12}$ (kromě diagonálních předpodmiňovačů, které jsou citlivější na odmítání).
- Pásové předpodmiňovače získané směrovým derivováním významně snižují počet iterací metody sdružených gradientů. V případě numerického derivování je však třeba počítat několik gradientů navíc, takže celkový počet použitých gradientů může být vyšší v porovnání s nepředpodmíněnou metodou sdružených gradientů. Tato situace nastává zejména tehdy, je-li počet iterací obou metod velmi malý. Abychom zlepšili účinnost metod používajících předpodmiňovače získané numerickým derivováním, můžeme tyto metody kombinovat s nepředpodmíněnou metodou sdružených gradientů. Výsledné kombinované metody lze popsat například takto:
  (1) Položíme $L = 0$ a $M = 10$ v první iteraci Newtonovy metody (hodnota $M = 10$ byla získána experimentálně).
  (2) V každé iteraci Newtonovy metody postupujeme takto:
    (a) Pokud $L = 0$ položíme $C = I$, jinak vypočteme tridiagonální matici T podle poznámky 322 a položíme $C = T$.
    (b) Pokud $L = 1$ a matice C není pozitivně definitní, položíme $C = I$ a $L = 0$.
    (c) Určíme směrový vektor metodou sdružených gradientů s předpodmiňovačem C.
    (d) Pokud $L = 0$ a počet iterací metody sdružených gradientů použité v (c) je vyšší než M, položíme $L = 1$.
- Předpodmiňovače získané Lanczosovou metodou není třeba korigovat (jsou podle věty 218 vždy pozitivně definitní). Jejich použití však ztěžuje to, že je nelze určovat v předpodmíněném kroku Newtonovy metody. To vyvolává řadu technických potíží (musí se upravovat iterační proces metody sdružených gradientů).

9.8 Numerické porovnání

K testování metod pro rozsáhlé husté úlohy byly použity úlohy ze sbírky TEST25 zmíněné v oddílu 1.5 a popsané v práci [106]. Nejprve porovnáme účinnost metod s proměnnou metrikou s omezenou pamětí pomocí 73 testovacích úloh s 1000 proměnnými (9 úloh ze sbírky TEST25 bylo vynecháno, protože je některá z testovaných metod nevyřešila). V tabulce 8 jsou uvedeny výsledky získané těmito metodami:

`LMVM-17` - vektorová metoda BFGS s omezenou pamětí (algoritmus 17),
`LMVM-18` - modifikovaná vektorová metoda s omezenou pamětí (algoritmus 18),
`LMVM-19` - maticová metoda BFGS s omezenou pamětí (algoritmus 19),
`LMVM-20` - maticová metoda s omezenou pamětí z Broydenovy třídy (algoritmus 20 s $\eta = 1.2$),
`LMVM-21` - modifikovaná maticová metoda s omezenou pamětí (algoritmus 21),
`LMRH-23` - metoda redukovaných Hessiánů s omezenou pamětí (algoritmus 23 upravený podle poznámek 306 a 306),
`LMSA-24` - posunutá metoda s omezenou pamětí (algoritmus 24 s volbou (844)),
`LMSB-24` - posunutá metoda s omezenou pamětí (algoritmus 24 s volbou (841) a (845)–(846)).

Pro srovnání jsou též uvedeny výsledky získané metodou sdružených gradientů CG, což je algoritmus 5, kde používáme volbu HST+ (vzorce (172), 173), (174)) a proceduru pro výběr délky kroku převzatou z programu CG-DESCENT. Výsledky pro metodu CG jsou též uvedeny oddílu 3.7 (ve sloupci HS a řádku MT+ tabulky 2.

Metoda	NIT	NFV	NFG	čas	počet
LMVM-17	126650	132891	132801	15.41	74
LMVM-18	116142	119273	119273	15.55	76
LMVM-19	124904	131269	131269	15.38	74
LMVM-20	122178	131037	131037	14.91	74
LMVM-21	107387	111364	111364	14.05	78
LMRH-23	124830	142264	142264	17.23	74
LMSA-24	137536	140846	140846	23.36	75
LMSB-24	132207	135014	135014	22.49	75
CG	169096	333431	176767	38.23	74

Tabulka 8: TEST25 – 73 úloh

Tabulka 8 obsahuje celkový počet iterací NIT, celkový počet použitých funkčních hodnot NFV, celkový počet použitých gradientů NFG, celkový čas výpočtu a počet úloh (z celkového počtu 82), které daná metoda vyřešila. Poslední údaj vyjadřuje robustnost dané metody. Z údajů uvedených v této tabulce lze vyvodit několik závěrů:

- Metody s proměnnou metrikou (s omezenou pamětí) jsou obvykle účinnější než metody redukovaných Hessiánů a posunuté metody s proměnnou metrikou. Je to způsobeno tím, že u metod s proměnnou metrikou lze volit $\bar{m} = 5$, zatímco metody redukovaných Hessiánů a posunuté metody s proměnnou metrikou vyžadují více aktualizací (obvykle $\bar{m} = 10$) a proto jsou pomalejší.
- Metodu BFGS realizovanou pomocí Strangových rekurencí (algoritmus 17) lze překonat vhodnými úpravami (algoritmus 18).
- Metody s proměnnou metrikou s omezenou pamětí realizované pomocí maticových reprezentací jsou velmi efektivní. Rekurzivní postup založený na větě 186 se zdá být stabilnější, než použití explicitního vzorce (760). Velmi účinná je modifikace použitá v algoritmu 21.
- Metody s proměnnou metrikou s omezenou pamětí jsou účinnější než metoda sdružených gradientů.

Nyní porovnáme účinnost vektorových diferenčních verzí Newtonovy metody s různými předpodmiňovači pomocí 71 testovacích úloh s 1000 proměnnými (11 úloh ze sbírky TEST25 bylo vynecháno, protože je některá z testovaných metod nevyřešila). V tabulce 9 jsou uvedeny výsledky získané těmito metodami spádových směrů:

LSTN - nepředpodmíněná vektorová diferenční Newtonova metoda,

LSTNL-V - metoda předpodmíněná třemi BFGS aktualizacemi realizovanými pomocí Strangových rekurencí (700)–(701),

LSTNL-M - metoda předpodmíněná třemi BFGS aktualizacemi realizovanými podle věty 186,

LSTNB-D - metoda s diagonálním předpodmiňovačem určeným BFGS aktualizacemi (850),

LSTNB-T - metoda s tridiagonálním předpodmiňovačem určeným BFGS aktualizacemi (850) s korekcí podle poznámky 327,

LSTNB-P - metoda s pentadiagonálním předpodmiňovačem, určeným BFGS aktualizacemi (850) s korekcí podle poznámky 327,

LSTND-D - metoda s diagonálním předpodmiňovačem určeným numerickým derivováním s náhradou diagonálních prvků jejich absolutními hodnotami,

LSTND-T - metoda s tridiagonálním předpodmiňovačem určeným numerickým derivováním s náhradou diagonálních prvků jejich absolutními hodnotami,

LSTND-P - metoda s pentadiagonálním předpodmiňovačem určeným numerickým derivováním s náhradou diagonálních prvků jejich absolutními hodnotami,

LSTNI - metoda s předpodmiňovačem (880) získaným symetrickým Lanczosovým procesem.

Pro srovnání jsou též uvedeny výsledky získané metodami LMVM-18, LMVM-21 a CG použitými v tabulce 8.

Metoda	NIT	NFV	NFG	NCG	NIP	NCP	čas	počet
LSTN	7409	11810	369496	356234	-	-	22.91	71
LSTNL-V	7247	12476	230881	217011	7247	-	15.67	77
LSTNL-M	7336	12710	243530	229352	7336	-	17.19	74
LSTNB-D	7082	10290	271970	260499	4432	37	18.12	71
LSTNB-T	6736	9235	137735	127706	4252	37	10.23	74
LSTNB-P	6782	8833	221622	211967	4009	36	18.29	73
LSTND-D	6484	8437	342404	326845	3856	40	20.92	71
LSTND-T	7614	11997	128785	99928	5650	2	9.56	74
LSTND-P	7083	10691	125080	86693	4922	5	9.01	74
LSTNI	7380	11656	349286	336197	4205	1	20.60	72
LMVM-18	107007	109673	109673	-	-	-	14.46	76
LMVM-21	107387	111364	111364	-	-	-	13.13	78
CG	130657	256358	138063	-	-	-	29.15	74

Tabulka 9: TEST25 – 71 úloh

Tabulka 9 obsahuje celkový počet iterací NIT, celkový počet použitých funkčních hodnot NFV, celkový počet použitých gradientů NFG, celkový počet vnitřních iterací (iterací metody sdružených gradientů) NCG, celkový počet předpodmíněných vnějších iterací (iterací Newtonovy metody) NIP, počet problémů, pro které bylo nutné zvýšit mez pro odmítání (poznámka 327) NCP, celkový čas výpočtu a počet úloh (z celkového počtu 82), které daná metoda vyřešila. Poslední údaj vyjadřuje robustnost dané metody. Z údajů uvedených v této tabulce lze vyvodit několik závěrů:

- Diferenční verze Newtonovy metody konvergují velmi rychle, vyžadují však velký počet gradientů minimalizované funkce k určení diferencí použitých v metodě sdružených gradientů.
- Nepředpodmíněná diferenční verze Newtonovy metody nemůže konkurovat metodám s proměnnou metrikou s omezenou pamětí.
- Diagonální předpodmiňovače a předpodmiňovače získané Lanczosovou metodou nejsou příliš efektivní. Lepší výsledky dávají tridiagonální a pentadiagonální předpodmiňovače.
- Pásové předpodmiňovače určené metodou BFGS je třeba upravovat podle vět 207 a 208. Často je také třeba zvýšit mez pro odmítání δ.
- Pásové předpodmiňovače určené numerickým derivováním vyžadují pouze minimální korekce. Dávají dobré výsledky zejména tehdy, jsou-li Hessovy matice pásové. Diferenční verze Newtonovy metody používající tyto předpodmiňovače jsou obvykle účinnější, než metody s proměnnou metrikou s omezenou pamětí.

10 Metody pro rozsáhlé řídké a separovatelné úlohy

10.1 Řídké matice a grafy

Rozsáhlé optimalizační úlohy se obvykle vyznačují tím, že jejich Jacobiovy matice (při minimalizaci součtu čtverců) a Hessovy matice obsahují málo (typicky $O(n)$) nenulových prvků. Protože počet nulových prvků je obvykle mnohem větší, je účelné ukládat pouze nenulové prvky a jen s nimi provádět aritmetické operace. Nejjednodušším způsobem jak ukládat nenulové prvky obecné řídké matice A je použití tří polí $num(A)$, $row(A)$, $col(A)$ délky m_A, kde m_A je počet nenulových prvků matice A. V poli $num(A)$ jsou v nějakém (obecně libovolném) pořadí uloženy numerické hodnoty nenulových prvků matice A, jejichž řádkové a sloupcové indexy jsou (ve stejném pořadí) uloženy v polích $row(A)$ a $col(A)$. Tento způsob je velmi vhodný pro zadávání obecných řídkých matic, ale není vhodný pro provádění operací s těmito maticemi. Proto byly vyvinuty jiné úspornější a pro práci s řídkými maticemi výhodnější reprezentace řídké struktury. V tomto oddílu popíšeme pouze komprimované ukládání po řádcích používající tři pole $num(A)$, $adr(A)$, $col(A)$, kde pole $num(A)$ a $col(A)$ mají stejný význam jako v předchozím případě pouze s tím rozdílem, že nenulové prvky jsou povinně uspořádány po řádcích (se vzrůstajícími indexy řádků a sloupců). Pole $adr(A)$ délky $n+1$ obsahuje ukazatele do polí $num(A)$ a $col(A)$ na začátky uložených řádků. Poslední prvek pole $adr(A)$ obsahuje počet nenulových prvků matice A zvětšený o 1 (tedy číslo $m_A + 1$).

Ještě úsporněji lze ukládat nenulové prvky řídké symmetrické matice B. Je zřejmé, že stačí ukládat pouze horní polovinu této matice, čili prvky matice B_U, která vznikne z B vynulováním poddiagonálních prvků. Horní trojúhelníkovou matici, která tím vznikne, reprezentujeme pomocí komprimovaného ukládání po řádcích. Pole $num(B)$ a $col(B)$ obsahují nenulové prvky matice B_U a jejich indexy uspořádané po řádcích. Pole $adr(B)$ obsahuje ukazatele do polí $num(B)$ a $col(B)$ na diagonální prvky matice B_U (a tedy i B). Pro matici (885) platí

$$\begin{aligned} num(G) &= [G_{11}, G_{12}, G_{22}, G_{23}, G_{33}, G_{34}, G_{44}, G_{45}, G_{55}], \\ col(G) &= [1, 2, 2, 3, 3, 4, 4, 5, 5], \\ adr(G) &= [1, 3, 5, 7, 9, 10]. \end{aligned}$$

Pro matici (886) platí

$$\begin{aligned} num(G) &= [G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{22}, G_{33}, G_{44}, G_{55}], \\ col(G) &= [1, 2, 3, 4, 5, 2, 3, 4, 5], \\ adr(G) &= [1, 6, 7, 8, 9, 10]. \end{aligned}$$

Kombinatorické problémy týkající se struktury řídkých matic se nejlépe popisují v řeči teorie grafů. Grafová formulace je názornější než maticový popis, neboť většinu grafových pojmů a operací si lze snadno představit. Řídkou strukturu obecné čtvercové matice lze reprezentovat pomocí orientovaného grafu.

Definice 58. *Orientovaným grafem $\vec{\mathcal{G}} = \vec{\mathcal{G}}(V, \vec{E})$ nazveme dvojici množin $V = \{v_1, \dots v_n\}$ a $\vec{E} \subset V \times V$. Prvky $v_i \in V$ nazveme vrcholy a dvojice $(v_i, v_j) \in \vec{E}$ orientovanými hranami orientovaného grafu $\vec{\mathcal{G}}$. Pokud $(v_i, v_i) \in \vec{E}$ nazveme tuto hranu vlastní smyčkou orientovaného grafu $\vec{\mathcal{G}}$.*

Definice 59. *Nechť A je řídká čtvercová matice řádu n. Pak orientovaný graf $\vec{\mathcal{G}}[A] = \mathcal{G}(V, \vec{E})$, kde $V = \{v_1, \dots v_n\}$ a $(v_i, v_j) \in \vec{E} \Leftrightarrow A_{ij} \neq 0$, nazveme grafem vyjadřujícím řídkou strukturu obecné čtvercové matice A.*

Poznámka 328. Pomocí orientovaného grafu lze reprezentovat i řídkou strukturu obdélníkové matice typu $n \times m$. Jestliže $n < m$ přidáme k matici $m - n$ nulových řádků. Jestliže $n > m$ přidáme k matici $n - m$ nulových sloupců. Vzniklou čtvercovou matici pak reprezentujeme pomocí orientovaného grafu. Jinou možností je použití bipartitního grafu (definice 71).

Řídkou strukturu čtvercové symetrické matice lze reprezentovat pomocí neorientovaného grafu.

Definice 60. *Nechť $\vec{\mathcal{G}} = \vec{\mathcal{G}}(V, \vec{E})$ a $\overleftarrow{E} = \{(v_j, v_i) : (v_i, v_j) \in \vec{E}\}$. Pak dvojici množin $\mathcal{G} = \mathcal{G}(V, E)$, kde $E = \vec{E} \cup \overleftarrow{E} \setminus \bigcup_{i=1}^{n}(v_i, v_i)$, nazveme neorientovaným grafem vzniklým zrušením orientace grafu $\vec{\mathcal{G}}$. Dvojice $(v_i, v_j) \in E$ a $(v_j, v_i) \in E$ považujeme za totožné (dva vrcholy jsou spojeny nanejvýš jednou hranou).*

Poznámka 329. Neorientovaný graf $\mathcal{G}$ získáme z orientovaného grafu $\vec{\mathcal{G}}$, odstraníme-li všechny vlastní smyčky, nahradíme-li orientované hrany neorientovanými hranami a případné dvojice hran spojující tytéž vrcholy nahradíme jedinou hranou (dvojice $(v_i, v_j) \in E$ a $(v_j, v_i) \in E$ považujeme za totožné).

Definice 61. *Nechť B je řídká symetrická matice, jejíž všechny diagonální prvky jsou (strukturálně) nenulové. Pak graf $\mathcal{G}[B] = \mathcal{G}(V, E)$, kde $V = \{v_1, \dots v_n\}$ a $(v_i, v_j) \equiv (v_j, v_i) \in E \Leftrightarrow B_{ij} \neq 0$, nazveme grafem vyjadřujícím řídkou strukturu symetrické matice B.*

V této kapitole se budeme zabývat převážně symetrickými maticemi, takže se zaměříme na vyšetřování vlastností neorientovaných grafů.

Definice 62. *Nechť $\mathcal{G} = \mathcal{G}(V, E)$ a $v_i \in V$. Pak množinu $\mathrm{adj}(v_i) = \{v_j \in V : (v_i, v_j) \in E\}$ nazveme okolím (množinou sousedů) vrcholu v_i a vrcholy $v_j \in \mathrm{adj}(v_i)$ nazveme sousedy vrcholu v_i. Stupněm vrcholu v_i nazveme počet prvků (kardinalitu) množiny $\mathrm{adj}(v_i)$ a budeme psát $\deg(v_i) = |\mathrm{adj}(v_i)|$.*

Definice 63. *Graf $\mathcal{G} = \mathcal{G}(V, E)$, kde pro libovolný vrchol $v \in V$ platí $V \setminus v = \mathrm{adj}(v)$, nazveme úplným grafem (v úplném grafu jsou každé dva vrcholy spojené hranou). Symbolem $\overline{\mathcal{G}}$ budeme označovat zúplnění grafu $\mathcal{G}$, neboli graf $\overline{\mathcal{G}}(V, \overline{E})$, který je úplný.*

Definice 64. *Nechť $\mathcal{G} = \mathcal{G}(V, E)$ a $\mathcal{G}' = \mathcal{G}'(V', E')$, kde $V' \subset V$, $E' \subset E$. Jestliže pro libovolnou hranu $(v_i, v_j) \in E'$ platí $v_i \in V'$, $v_j \in V'$, řekneme, že $\mathcal{G}'$ je podgrafem grafu $\mathcal{G}$ a píšeme $\mathcal{G}' \subset \mathcal{G}$. Graf $\mathcal{G}(V') = \mathcal{G}(V', E(V'))$, kde $V' \subset V$, $E(V') \subset E$ a $(v_i, v_j) \in E(V') \Leftrightarrow v_i \in V'$, $v_j \in V'$, nazveme podgrafem grafu $\mathcal{G}$ určeným množinou vrcholů V'. Podgraf $\mathcal{G}(V')$, který je úplný nazveme klikou. Kliku, která není podgrafem žádné větší kliky nazveme maximální klikou.*

Definice 65. *Nechť $\mathcal{G}_1 = \mathcal{G}_1(V_1, E_1)$ a $\mathcal{G}_2 = \mathcal{G}_2(V_2, E_2)$ jsou dva podgrafy grafu $\mathcal{G} = \mathcal{G}(V, E)$. Pak graf $\mathcal{G}_3 = \mathcal{G}_3(V_1 \cup V_2, E_2 \cup E_2)$ nazveme sjecnocením podgrafů $\mathcal{G}_1$ a $\mathcal{G}_2$ a píšeme $\mathcal{G}_3 = \mathcal{G}_1 \cup \mathcal{G}_1$ (sjednocení podgrafů dostaneme, sjednotíme-li jejich množiny vrcholů a hran).*

Definice 66. *Nechť $\mathcal{G} = \mathcal{G}(V, E)$. Pak posloupnost vrcholů $v_{i_j} \in V$, $1 \leq j \leq k$, takovou že $v_{i_{j+1}} \in \mathrm{adj}(v_{i_j})$, $1 \leq j < k$, nazveme sledem v grafu $\mathcal{G}$. Jestliže se vrcholy (s výjimkou případu, kdy $v_{i_k} = v_{i_1}$) neopakují, nazveme tento sled cestou v grafu $\mathcal{G}$. Cestu, pro kterou platí $v_{i_k} = v_{i_1}$ nazveme smyčkou v grafu $\mathcal{G}$.*

Poznámka 330. Sled (cestu, smyčku) $v_{i_j} \in V$, $1 \leq j \leq k$, lze též vyjádřit jako posloupnost hran $(v_{i_j}, v_{i_{j+1}})$, $1 \leq j < k$. Délkou sledu (cesty, smyčky) nazveme počet těchto hran, tedy číslo $k - 1$.

Definice 67. *Nechť $\mathcal{G} = \mathcal{G}(V, E)$. Jestliže pro libovolné dva vrcholu $v_i \in V$ a $v_j \in V$ existuje cesta, která obsahuje $v_i \in V$ a $v_j \in V$, řekneme, že graf $\mathcal{G} = \mathcal{G}(V, E)$ je souvislý.*

Definice 68. *Souvislý podgraf $\mathcal{G}'(V, E') \subset \mathcal{G}(V, E)$, který neobsahuje žádnou smyčku, nazveme stromem grafu $\mathcal{G}$.*

Definice 69. *Obarvením grafu $\mathcal{G}(V, E)$ nazveme rozklad $V = \bigcup_{i=1}^{k} V_i$. takový, že $V_i \cap V_j = \emptyset$, pokud $1 \leq i < j \leq k$. Množinám V_i, $1 \leq i \leq k$, přiřazujeme různé barvy (například vcísla $1 \leq i \leq k$). Pokud neexistuje hrana, jejíž oba koncové vrcholy jsou obarveny stejnou barvou, řekneme, že obarvení je dobré. Minimalní počet barev potřebných k dobrému obarvení grafu nazveme chromatickým číslem grafu.*

Poznámka 331. Pojem dobrého obarvení grafu lze použít i pro orientované grafy, neuvažujeme-li vlastní smyčky. Ani v dobře obarveném orientovaném grafu neexistuje hrana, jejíž oba koncové vrcholy jsou obarveny stejnou barvou. Je tedy možné zrušit orientaci (definice 60), aniž to má vliv na vlastnosti obarvení.

Definice 70. *Graf $\hat{\mathcal{G}} = \hat{\mathcal{G}}(V, \hat{E})$, který má chromarické číslo 2, nazveme bipartitním grafem. V tomto případě $V = V' \cup V''$, $V' \cap V'' = \emptyset$ a $\hat{E} \subset V' \times V''$.*

Bipartitní graf lze použít k reprezentaci řídké obdélníkové matice.

Definice 71. *Nechť A je řídká obdélníková matice typu $n \times m$. Pak bipartitní graf $\hat{\mathcal{G}}[A] = \hat{G}(V, \hat{E})$, kde $V = V' \cup V''$, $V' = \{v'_1, \dots v'_n\}$, $V'' = \{v''_1, \dots v''_m\}$ a $(v'_i, v''_j) \in \hat{E} \Leftrightarrow A_{ij} \neq 0$, nazveme bipartitním grafem vyjadřujícím řídkou strukturu obdelníkové matice A.*

10.2 Diferenční verze Newtonovy metody pro řídké úlohy

Diferenční verze Newtonovy metody pro řídké úlohy jsou založeny na aproximaci sloupců Ge_i, $1 \leq i \leq n$, Hessovy matice G pomocí diferenčních vzorců

$$G(x)e_i \approx \frac{g(x + \delta e_i) - g(x)}{\delta}, \qquad 1 \leq i \leq n, \tag{884}$$

kde δ je malá diference (obvykle $\delta = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost). Je-li však Hessova matice G řídká, může nastat případ, kdy pomocí jedné diference gradientů určíme více sloupců této matice. Jako příklad uvedeme pásovou matici

$$G = \begin{bmatrix} G_{11}, & G_{12}, & 0, & 0, & 0 \\ G_{21}, & G_{22}, & G_{23}, & 0, & 0 \\ 0, & G_{32}, & G_{33}, & G_{34}, & 0 \\ 0, & 0, & G_{43}, & G_{44}, & G_{45} \\ 0, & 0, & 0, & G_{54}, & G_{55} \end{bmatrix}. \tag{885}$$

Nechť

$$v_1 = [1, 0, 0, 1, 0]^T, \qquad v_2 = [0, 1, 0, 0, 1]^T, \qquad v_3 = [0, 0, 1, 0, 0]^T.$$

Pak platí

$$Gv_1 = [G_{11}, G_{21}, G_{34}, G_{44}, G_{54}]^T, \quad Gv_2 = [G_{12}, G_{22}, G_{32}, G_{45}, G_{55}]^T, \quad Gv_3 = [0, G_{23}, G_{33}, G_{43}, 0]^T,$$

takže všechny prvky matice G můžeme určit pomocí tří diferenčních vzorců

$$\frac{g(x + \delta v_1) - g(x)}{\delta} \approx Gv_1, \qquad \frac{g(x + \delta v_2) - g(x)}{\delta} \approx Gv_2, \qquad \frac{g(x + \delta v_3) - g(x)}{\delta} \approx Gv_3.$$

Postup, který jsme použili v tomto konkrétním případě můžeme snadno zobecnit [26]. Rozdělme sloupce matice G do k disjunktních skupin $\mathcal{S}_i \subset \{1, \dots, n\}$, $1 \leq i \leq k$, tak aby submatice $G(\mathcal{S}_i)$, složené ze sloupců matice G patřících do skupin $\mathcal{S}_i$, měly v každém řádku nanejvýš jeden nenulový prvek (tyto sloupce nazveme strukturálně ortogonálními sloupci matice G). Pak můžeme všechny sloupce matice G určit pomocí k diferencí

$$\frac{g(x + \delta v_i) - g(x)}{\delta} \approx Gv_i, \qquad 1 \leq i \leq k,$$

kde v_i, $1 \leq i \leq k$, jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \quad \Longleftrightarrow \quad j \in \mathcal{S}_i$$

(pomocí vektoru v_i určíme prvky submatice $G(\mathcal{S}_i)$). Takto lze postupovat pro libovolnou čtvercovou matici G a dokonce i pro libovolnou obdélníkovou matici J. Bližší podrobnosti jsou uvedeny v oddílu 12.3.

Je-li matice G symetrická, můžeme její symetrii využít k dalšímu snížení počtu potřebných diferencí. Uvažujme matici

$$G = \begin{bmatrix} G_{11}, & G_{12}, & G_{13}, & G_{14}, & G_{15} \\ G_{21}, & G_{22}, & 0, & 0, & 0 \\ G_{31}, & 0, & G_{33}, & 0, & 0 \\ G_{41}, & 0, & 0, & G_{44}, & 0 \\ G_{51}, & 0, & 0, & 0, & G_{55} \end{bmatrix}. \tag{886}$$

Použijeme-li předchozí postup, potřebujeme k určení prvků matice G pět diferencí gradientů. Položíme-li však

$$v_1 = [1,0,0,0,0]^T, \qquad v_2 = [0,1,1,1,1]^T,$$

platí

$$Gv_1 = [G_{11}, G_{21}, G_{31}, G_{41}, G_{51}]^T, \qquad Gv_2 = [*, G_{22}, G_{33}, G_{44}, G_{55}]^T,$$

kde hvězdičkou je označen prvek, který nás nezajímá. Určili jsme tedy prvky G_{11}, G_{21}, G_{31}, G_{41}, G_{51}, G_{22}, G_{33}, G_{44}, G_{55} a protože matice G je symetrická i prvky $G_{12} = G_{21}$, $G_{13} = G_{31}$, $G_{14} = G_{41}$, $G_{15} = G_{51}$, to vše pomocí dvou diferencí gradientů.

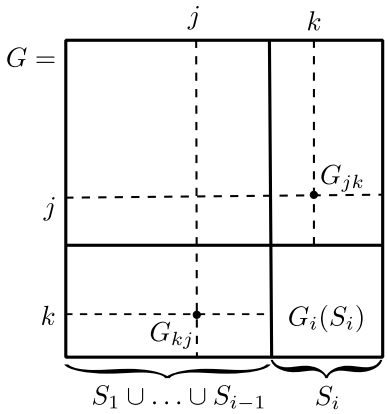
Postup, který jsme použili v tomto konkrétním případě můžeme opět zobecnit [27], [139]. Sloupce matice G rozdělíme opět do k disjunktních skupin $\mathcal{S}_i$, $1 \le i \le k$. Při určování těchto skupin však nebudeme pracovat s celou maticí G, ale pouze s jejími submaticemi, které dostaneme vyškrtnutím známých řádků a sloupců. Nechť G_i je submatice matice G, kterou dostaneme, vyškrtneme-li v matici G řádky a sloupce s indexy z $S_1 \cup \cdots \cup S_{i-1}$, a nechť $G_i(\mathcal{S}_i)$ je submatice matice G_i, která obsahuje sloupce této matice s indexy z $\mathcal{S}_i$, takže $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$ (zde $G_i \cap G(\mathcal{S}_i)$ je submatice obsahující prvky společné submaticím G_i a $G(S_i)$). Rozdělíme-li sloupce matice G do k disjunktních skupin $\mathcal{S}_i$, $1 \le i \le k$, tak aby submatice $G_i(\mathcal{S}_i)$, $1 \le i \le k$, měly v každém řádku nanejvýš jeden nenulový prvek, můžeme sloupce matice G určit pomocí k diferencí

$$\frac{g(x + \delta v_i) - g(x)}{\delta} \approx Gv_i, \quad 1 \le i \le k, \tag{887}$$

kde v_i, $1 \le i \le k$, jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \quad \Longleftrightarrow \quad j \in \mathcal{S}_i.$$

Pomocí vektoru v_i určíme prvky submatice $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$, ostatní prvky submatice $G(\mathcal{S}_i)$ jsou určeny symetrií matice G, jak je ukázáno na následujícím obrázku (kde pro jednoduchost předpokládáme, že $S_1 \cup \cdots \cup S_i = \{1, \ldots, n\}$). V tomto obrázku platí $G_{jk} \in G(S_i) \backslash G_i(S_i) \Rightarrow G_{jk} = G_{kj} \in G(S_1 \cup \cdots \cup S_{i-1})$, takže prvek G_{jk} je již znám.



Zatím jsme se nezabývali určováním skupin $\mathcal{S}_i$, $1 \le i \le k$. Je účelné volit tyto skupiny tak, aby jejich počet byl minimální. To je však NP těžká úloha, kterou nelze v obecném případě vyřešit v rozumném čase.

Proto se, tak jako ve většině kombinatorických úloh používají algoritmy, které jsou poměrně jednoduché, rychlé a dávají dobrou aproximaci optimálního řešení.

Při určování skupin $\mathcal{S}_i$, $1 \le i \le k$, lze použít sekvenční postup. Sloupce submatice G_i se nejprve přerovnají podle nějakého pravidla a potom se probírají postupně podle vzrůstajících indexů. Index $j \in \{1, \ldots, n\} \backslash (S_1 \cup \cdots \cup S_{i-1})$ se přidá do skupiny $\mathcal{S}_i$ pouze tehdy, neporuší-li se přitom požadavek, aby submatice $G_i(\mathcal{S}_i)$ měla v každém řádku nanejvýš jeden nenulový prvek.

Poznámka 332. Na přerovnání sloupců submatice G_i obvykle dosti záleží. Následující matice se liší pouze pořadím řádků a sloupců (nenulové prvky jsou znázorněny symbolem $*$).

$$\begin{bmatrix} * & & & & * \\ & * & & & * \\ & & * & & * \\ & & & * & * \\ * & * & * & * & * \end{bmatrix}, \qquad \begin{bmatrix} * & * & * & * & * \\ * & * & & & \\ * & & * & & \\ * & & & * & \\ * & & & & * \end{bmatrix}.$$

Probíráme-li sloupce první matice sekvenčně podle vzrůstajících indexů, potřebujeme k určení všech nenulových prvků celkem pět diferencí gradientů. Probíráme-li sloupce druhé matice sekvenčně podle vzrůstajících indexů, stačí k určení všech nenulových prvků pouze dvě diference gradientů. Z této ukázky plyne, že je vhodné volit takové přerovnání, aby řádky a sloupce s nejnižšími indexy měly co nejvíce nenulových prvků.

Následující jednoduchý algoritmus je shrnutím dosavadních úvah.

Algoritmus 25. Data: Řídká struktura Hessovy matice G řádu n.

Krok 1 Položíme $G_1 = G$ a $i = 1$.

Krok 2 Pokud $G_i = \emptyset$, ukončíme výpočet (množiny $S_1, \ldots S_{i-1}$ tvoří hledaný rozklad).

Krok 3 Položíme $S_i = \emptyset$ a uspořádáme sloupce submatice G_i podle počtu nenulových prvků sestupně. Pak probíháme sloupce submatice G_i v daném pořadí a do S_i vždy přidáme index sloupce, který je strukturálně ortogonální ke všem sloupcům, jejichž indexy jsou již v S_i obsaženy.

Krok 4 Určíme matici G_{i+1} vyškrtnutím řádků a sloupců submatice G_i s indexy obsaženými v S_i. Zvětšíme i o 1 a přejdeme na krok 2.

kromě tohoto jednoduchého algoritmu existují další složitější algoritmy popsané v pracích [27], [60], [154]. Je také možné použít volně dostupný zdrojový program v jazyce Fortran [25] popsaný v [24].

Zatím jsme se zabývali přímými metodami pro výpočet prvků řídké Hessovy matice pomocí diferencí. Nyní obrátíme pozornost na substituční metody, které obvykle vyžadují menší počet diferencí než přímé metody. Uvažujme opět matici (885) a položme

$$v_1 = [1, 0, 1, 0, 1]^T, \qquad v_2 = [0, 1, 0, 1, 0]^T.$$

Pak platí

$$\frac{g(x + \delta v_1) - g(x)}{\delta} \approx G v_1 = \begin{bmatrix} G_{11} \\ G_{21} + G_{23} \\ G_{33} \\ G_{43} + G_{45} \\ G_{55} \end{bmatrix}, \qquad \frac{g(x + \delta v_2) - g(x)}{\delta} \approx G v_2 = \begin{bmatrix} G_{12} \\ G_{22} \\ G_{32} + G_{34} \\ G_{44} \\ G_{54} \end{bmatrix}.$$

Z těchto rovnic určíme přímo hodnoty G_{11}, G_{33}, G_{55}, G_{12}, G_{22}, G_{44}, G_{54} a protože matice G je symetrická i hodnoty $G_{21} = G_{12}$, $G_{45} = G_{54}$. Dosadíme-li hodnoty G_{21}, G_{45} zpět do uvedených rovnic, určíme hodnoty $G_{23} = e_2^T G v_1 - G_{21}$, $G_{43} = e_2^T G v_1 - G_{45}$ a protože matice G je symetrická i hodnoty $G_{32} = G_{23}$,

$G_{34} = G_{43}$. Potřebujeme k tomu pouze dvě diference gradientů (přímá metoda používá tři diference gradientů). Tento způsob výpočtu prvků tridiagonální Hessovy matice byl použit v oddílu 9.7 (věta 211).

Postup, který jsme použili v tomto konkrétním případě můžeme snadno zobecnit [27], [139]. Nechť G_U je horní trojúhelníková matice, jejíž horní trojúhelníková část má stejnou strukturu (rozložení nenulových prvků) jako horní trojúhelníková část matice G. Rozdělme sloupce matice G_U do k disjunktních skupin $\mathcal{S}_i \subset \{1, \ldots, n\}$, $1 \leq i \leq k$, tak, aby submatice $G_U(\mathcal{S}_i)$ složené ze sloupců matice G_U patřících do skupin $\mathcal{S}_i$, měly v každém řádku nanejvýš jeden nenulový prvek. Pak můžeme všechny sloupce matice G určit pomocí k diferencí

$$\frac{g(x + \delta v_i) - g(x)}{\delta} \approx G v_i, \quad 1 \leq i \leq k,$$

kde v_i, $1 \leq i \leq k$ jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \quad \Longleftrightarrow \quad j \in \mathcal{S}_i$$

Výpočet prvků Hessovy matice se provádí po řádcích vzestupně. První prvky vektorů Gv_i, $1 \leq i \leq k$, určují nenulové prvky prvního řádku matice G_U a zároveň nenulové prvky prvního sloupce matice G. Druhé prvky vektorů Gv_i, $1 \leq i \leq k$, určují nenulové prvky druhého řádku matice G_U s tím, že se od jednoho z těchto prvků odečte prvek G_{21} (pokud je nenulový). Tím dostaneme zároveň druhý sloupec matice G. Takto postupujeme dále. Pomocí j-tých prvků vektorů Gv_i, $1 \leq i \leq k$, a poddiagonálních prvků j-tého řádku matice G určíme nenulové prvky j-tého řádku matice G_U a zároveň nenulové prvky j-tého sloupce matice G. Substituční metody jsou algoritmicky složitější než přímé metody reprezentované algoritmem 25 a používají se především k určování prvků pásových matic jak je ukázáno v oddílu 9.7.

Smyslem těchto úvah bylo ukázat, že určení Hessovy matice pomocí diferencí gradientů může být časově nenáročné, je-li tato matice řídká. To staví diferenční verze Newtonovy metody do zcela jiného světla, neboť pro řídké úlohy mohou konkurovat metodám s proměnnou metrikou a metodám sdružených gradientů nebo je i překonat.

Diferenční verze Newtonovy metody pro řídké úlohy se obvykle realizují jako metody s optimálním lokálně omezeným krokem (algoritmus 10) nebo jako nepřesné metody s lokálně omezeným krokem (algoritmus 11). Metody s optimálním lokálně omezeným krokem vyžadují opakované řešení soustavy rovnic $(G + \lambda I)s + g = 0$ (pro různé hodnoty parametru $\lambda \geq 0$). Používá se přitom řídký Choleského rozklad

$$R^T R = P(G + \lambda I)P^T,$$

kde R je regulární horní trojúhelníková matice a P je permutační matice, jejíž jediným účelem je přerovnat řádky a sloupce matice $G + \lambda I$ tak, aby počet nově vzniklých nenulových prvků byl co nejmenší. Nalezení permutační matice P a následné určení struktury horní trojúhelníkové matice R se nazývá symbolickou faktorizací. Symbolická faktorizace se provádí pouze jednou (na začátku iteračního procesu) a proto je možné používat časově náročnější algoritmy, které přibližně minimalizují počet nově vzniklých nenulových prvků. Výpočet prvků horní trojúhelníkové matice R (numerická faktorizace) se provádí v každém iteračním kroku podle vzorců uvedených v oddílu 2.7.

Nepřesné metody s lokálně omezeným krokem používají metodu sdružených gradientů popsanou v oddílu 3.8 (Algoritmus 6), kde se řídká Hessova matice G používá pouze k výpočtu součinů $q_i = Gp_i$, $1 \leq i \leq n$, a není ji tudíž třeba rozkládat. V souvislosti s diferenční verzí Newtonovy metody pro řídké úlohy se osvědčilo předpodmiňování pomocí neúplného Choleského rozkladu. Princip tohoto postupu spočívá v provádění Choleského rozkladu, při němž se zanedbávají všechny nově vznikající nenulové prvky (někdy se nově vznikajícími nenulovými prvky modifikuje diagonála rozkládané matice). Získaná horní trojúhelníková matice R má stejnou strukturu jako horní btrojúhelníková část matice G a aproximace $RR^T \approx G$ je často velmi dobrá, což dává velmi účinné předpodmínění.

V závěru tohoto oddílu se budeme podrobněji zabývat symbolickou faktorizací a zaplňováním řídké matice při provádění Choleského rozkladu $B = R^T R$. Z popisu algoritmu 3, ve kterém se konstruuje posloupnost matic $\bar{B}_k$, $1 \leq k \leq n$, řádu $n - k + 1$ s řádkovými a sloupcovými indexy $k \leq i \leq n$ (kde $\bar{B}_1 = B$) je zřejmé, že v kroku 3 vznikne nenulový prvek právě tehdy, když $\bar{B}_{ki}\bar{B}_{kj} \neq 0$ a $\bar{B}_{ij} = 0$.

Vyjádřeno v řeči grafů to znamená, že $\mathcal{G}[\bar{B}_k] = \mathcal{G}(V_k, E_k)$, kde $V_k = \{v_i : k \leq i \leq n\}$, a nový nenulový prvek vznikne právě tehdy, když $(v_k, v_i) \in E_k$, $(v_k, v_j) \in E_k$ a $(v_i, v_j) \notin E_k$. Označíme-li $\mathcal{G}_k = \mathcal{G}[\bar{B}_k]$, platí $\mathcal{G}_1 = \mathcal{G}[B] = \mathcal{G}(V_1, E_1)$ a

$$\mathcal{G}[\bar{B}_{k+1}] = \mathcal{G}(V_k \setminus \{v_k\}) \cup \overline{\mathcal{G}}(\mathrm{adj}(v_k))$$

(používáme označení z definic 62–65. Graf $\mathcal{G}[\bar{B}_{k+1}]$ tedy vznikne z grafu $\mathcal{G}[\bar{B}_k]$ eliminací vrcholu v_k podle následující definice.

Definice 72. *Nechť $\mathcal{G} = \mathcal{G}(V, E)$ a $v \in V$. Pak řekneme, že graf*

$$\mathcal{G}^{(v)} = \mathcal{G}(V \setminus \{v\}) \cup \overline{\mathcal{G}}(\mathrm{adj}(v))$$

vznikl z grafu $\mathcal{G}$ eliminací vrcholu v. Eliminaci vrcholu v lze popsat tak, že vyjmeme vrchol v a všechny hrany s ním incidentní a přidáme (pomocí operace sjednocení) kliku tvořenou vrcholy z $\mathrm{adj}(v)$.

Aplikujeme-li na řídkou symetrickou matici Choleského rozklad, je žádoucí, aby vznikalo co nejméně nenulových prvků. Zaplnění Choleského faktoru velmi závisí na pořadí v jakém eliminujeme diagonální prvky, čili na permutaci řádků a sloupců, pokud provádíme eliminaci v přirozeném pořadí. Aplikujeme-li Choleského rozklad na první matici v poznámce 332, nevznikne žádný nový nenulový prvek, zatímco druhá matice se zcela zaplní. Snažíme se tedy uspořádat řádky a sloupce tak aby toto uspořádání bylo perfektní podle následující definice.

Definice 73. *Řádky a sloupce řídké symetrické matice B jsou perfektně eliminačně uspořádány, jestliže při symbolickém Choleského rozkladu $B = R^T R$ nevznikají žádné nové nenulové prvky.*

Perfektní uspořádání obvykle neexistuje. Proto se nabízí volit uspořádání tak, aby počet nově vzniklých nenulových prvků byl minimální. To je však NP těžká úloha, kterou nelze v obecném případě vyřešit v rozumném čase. Proto se, tak jako ve většině kombinatorických úloh používají algoritmy, které jsou poměrně jednoduché, rychlé a dávají dobrou aproximaci optimálního řešení. Jedním takovým algoritmem je následující algoritmus uspořádání podle minimálního stupně.

Algoritmus 26. Data: Řídká struktura symetrické matice B řádu n.

Krok 1 Položíme $\mathcal{G}_1 = \mathcal{G}[B] = \mathcal{G}(V_1, E_1)$ a $k = 1$.

Krok 2 Pokud $k = n$, ukončíme výpočet (pořadí nalezených vrcholů tvoří uspořádání řádků a sloupců pro určení Choleského rozkladu).

Krok 3 Nalezneme vrchol $v_k \in V_k$ takový, že $\deg(v_k) = \min_{v_i \in V_k} \deg(v_i)$.

Krok 4 Sestrojíme graf $\mathcal{G}_{k+1} = \mathcal{G}_k^{(v_k)}$ (definice 72).

Krok 5 Zvětšíme k o 1 a přejdeme na krok 2.

V kroku 3 algoritmu 26 není výběr vrcholu v_k určen jednoznačně. I když se používají různá heuristická pravidla, teoreticky podložený způsob výběru není nikde popsán. I když algoritmus 26 vypadá jednoduše, jeho implementace není triviální. Grafové operace použité v kroku 4 jsou časově i operačně náročné a velmi záleží na tom, jaké datové struktury se použijí.

K nalezení vhodného uspořádání řádků a sloupců pro provádění Choleského rozkladu řídké symetrické matice (spolu s určením struktury Choleského faktoru, který obsahuje nové nenulové prvky) by bylo možné použít pole $adr(B)$, $col(B)$ a patřičně je upravovat (doplňovat a posouvat), tento značně neefektivní postup se však nepoužívá. Problematice symbolické faktorizace je věnována velká pozornost v [62] a v učebním textu [T9]. Zde se těmito algoritmickými detaily zabývat nebudeme a to zejména proto, že jsou dostupné kvalitní programy pro řešení lineárních soustav s řídkou symetrickou maticí (například CHOLMOD dostupný na `http://faculty.cse.tamu.edu/davis/suitesparse.html`). Podobný program, napsaný v jazyce Fortran, je uveden v [44].

10.3 Metody s proměnnou metrikou pro řídké úlohy

Metody s proměnnou metrikou pro řídké úlohy používají aktualizace, které zachovávají řídkou strukturu (rozložení nenulových prvků) Hessovy matice. Toto zachovávání řídké struktury je násilným omezením, které eliminuje některé jiné důležité vlastnosti metod s proměnnou metrikou (například nalezení minima kvadratické funkce po konečném počtu kroků), nicméně lze získat metody, které jsou Q-superlineárně konvergentní. Nastávají však potíže s globální konvergencí, neboť získaná aproximace Hessovy matice nemusí být pozitivně definitní.

Od metod s proměnnou metrikou pro řídké úlohy požadujeme, aby použité aktualizace splňovaly kvazinewtonovskou podmínku, neporušovaly symetrii a zachovávaly řídkou strukturu Hessovy matice. Označme

$$\begin{aligned}\mathcal{V}_Q &= \{B \in R^{n\times n} : Bd = y\},\\ \mathcal{V}_S &= \{B \in R^{n\times n} : B^T = B\},\\ \mathcal{V}_G &= \{B \in R^{n\times n} : B_{ij} = 0, \text{ pokud } G_{ij} = 0\}.\end{aligned}$$

Symbolem $G_{ij} = 0$ budeme v tomto oddílu označovat pouze strukturální nuly, tedy prvky pro které platí $G_{ij}(x) = 0\ \forall x \in R^n$. Ze symetrie Hessovy matice je zřejmé, že $G_{ij} = 0 \Leftrightarrow G_{ji} = 0$. Protože se budeme snažit, aby approximace Hessovy matice byla pozitivně definitní, budeme předpokládat, že $G_{ii} \neq 0$, $1 \leq i \leq n$ (diagonální prvky Hessovy matice budou strukturálně nenulové). Zřejmě $\mathcal{V}_Q \subset R^{n\times n}$, $\mathcal{V}_S \subset R^{n\times n}$, $\mathcal{V}_G \subset R^{n\times n}$ jsou lineární variety ($\mathcal{V}_S$ a $\mathcal{V}_G$ jsou podprostory) v $R^{n\times n}$. Jelikož Frobeniova norma matice je euklidovskou normou v $R^{n\times n}$ (chápeme-li matice řádu n jako vektory dimenze $n \times n$), můžeme definovat operátory ortogonální projekce $\mathcal{P}_Q$, $\mathcal{P}_S$, $\mathcal{P}_G$ do $\mathcal{V}_Q$, $\mathcal{V}_S$, $\mathcal{V}_G$ předpisem

$$\begin{aligned}\mathcal{P}_Q B &= \arg\min_{\tilde{B}\in\mathcal{V}_Q} \|\tilde{B} - B\|_F,\\ \mathcal{P}_S B &= \arg\min_{\tilde{B}\in\mathcal{V}_S} \|\tilde{B} - B\|_F,\\ \mathcal{P}_G B &= \arg\min_{\tilde{B}\in\mathcal{V}_G} \|\tilde{B} - B\|_F.\end{aligned}$$

Podobně můžeme definovat operátory ortogonální projekce $\mathcal{P}_{QS}$, $\mathcal{P}_{QG}$, $\mathcal{P}_{SG}$ a $\mathcal{P}_{QSG}$ do lineárních variet $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$ a $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Zřejmě $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G \neq \emptyset$, neboť $\tilde{G} \in \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$, kde $\tilde{G}$ je matice definovaná vztahem (375). Je zřejmé, že naše požadavky na řídkou aktualizaci splňuje matice $B_+ = \mathcal{P}_{QSG}B$. V tomto oddílu ukážeme, že i jednoduché aktualizace založené na skládání projekcí mohou vést k superlineárně konvergentním metodám.

Věta 219. *Nechť $B \in R^{n\times n}$ a nechť $\mathcal{P}_Q$, $\mathcal{P}_S$, $\mathcal{P}_G$ jsou operátory ortogonální projekce do $\mathcal{V}_Q$, $\mathcal{V}_S$, $\mathcal{V}_G$. Pak*

$$\begin{aligned}\mathcal{P}_Q B &= B + \frac{(y - Bd)d^T}{d^T d},\\ \mathcal{P}_S B &= \frac{1}{2}(B + B^T),\\ (\mathcal{P}_G B)_{ij} &= B_{ij}, \quad G_{ij} \neq 0,\\ (\mathcal{P}_G B)_{ij} &= 0, \quad G_{ij} = 0,\end{aligned}$$

kde $1 \leq i \leq n$, $1 \leq j \leq n$.

Důkaz (a) Vztah pro $\mathcal{P}_Q B$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned}L(\tilde{B}, u) &= \frac{1}{2}\|\tilde{B} - B\|_F^2 + u^T(y - \tilde{B}d)\\ &= \frac{1}{2}\sum_{i=1}^n\sum_{j=1}^n(\tilde{B}_{ij} - B_{ij})^2 + \sum_{i=1}^n u_i\left(y_i - \sum_{j=1}^n \tilde{B}_{ij}d_j\right).\end{aligned}$$

Derivujeme-li Lagrangeovu funkci podle prvků matice $\tilde{B}$ a položíme-li derivace rovny nule, dostaneme

$$\frac{\partial L(\tilde{B},u)}{\partial \tilde{B}_{kl}} = (\tilde{B}_{kl} - B_{kl}) - u_k d_l = 0, \qquad 1 \le k \le n, \quad 1 \le l \le n.$$

Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $\tilde{B} = B + ud^T$, což po dosazení do kvazinewtonovské podmínky $\tilde{B}d = y$ dává $u\, d^T d = y - Bd$, neboli

$$u = \frac{y - Bd}{d^T d}.$$

Dosadíme-li tento výraz do vzorce $\tilde{B} = B + ud^T$, dostaneme vztah pro $\mathcal{P}_Q B$.

(b) Vztah pro $\mathcal{P}_S B$ odvodíme pomocí Lagrangeovy funkce

$$L(\tilde{B},V) = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(\tilde{B}_{ij} - B_{ij})^2 + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}(\tilde{B}_{ij} - \tilde{B}_{ji}).$$

Podmínky optimality mají tvar

$$\begin{aligned}
\frac{\partial L(\tilde{B},V)}{\partial \tilde{B}_{kl}} &= \frac{1}{2}(\tilde{B}_{kl} - B_{kl}) + v_{kl} - v_{lk} = 0,\\
\frac{\partial L(\tilde{B},V)}{\partial \tilde{B}_{lk}} &= \frac{1}{2}(\tilde{B}_{lk} - B_{lk}) + v_{lk} - v_{kl} = 0,
\end{aligned}$$

kde $1 \le k \le n$, $1 \le l \le n$. Sečteme-li obě rovnosti a přihlédneme-li k symetrii matice $\tilde{B}$, dostaneme

$$\tilde{B}_{kl} - \frac{1}{2}(B_{kl} + B_{lk}) = 0, \qquad 1 \le k \le n, \quad 1 \le l \le n,$$

což maticově zapsáno dává vztah pro $\mathcal{P}_S B$. Poznamenejme, že stejný výsledek dostaneme nepodmíněnou minimalizací funkce

$$\frac{1}{2}\|\mathcal{P}_S(\tilde{B} - B)\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\tilde{B}_{ij} - \frac{1}{2}(B_{ij} + B_{ji})\right)^2.$$

(c) Vztah pro $\mathcal{P}_G B$ odvodíme minimalizací funkce

$$\frac{1}{2}\|\tilde{B} - B\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\tilde{B}_{ij} - B_{ij})^2 = \frac{1}{2}\sum_{G_{ij}\neq 0}(\tilde{B}_{ij} - B_{ij})^2 + \frac{1}{2}\sum_{G_{ij}=0} B_{ij}^2,$$

neboť pro $\tilde{B} \in \mathcal{V}_G$ platí $\tilde{B}_{ij} = 0$, pokud $G_{ij} = 0$. Jelikož poslední člen na pravé straně nezávisí na prvcích matice $\tilde{B}$, dostaneme stejný výsledek minimalizací funkce

$$\frac{1}{2}\|\mathcal{P}_G(\tilde{B} - B)\|_F^2 = \frac{1}{2}\sum_{G_{ij}\neq 0}(\tilde{B}_{ij} - B_{ij})^2.$$

Derivujeme-li tuto funkci podle nenulových prvků matice $\tilde{B}$ a položíme-li derivace rovny nule dostaneme

$$\begin{aligned}
\tilde{B}_{kl} - B_{kl} &= 0, \quad G_{kl} \neq 0,\\
\tilde{B}_{kl} &= 0, \quad G_{kl} = 0,
\end{aligned}$$

kde $1 \le k \le n$, $1 \le l \le n$, což jsme měli dokázat. □

V důkazu předchozí věty jsme ukázali, že pokud $\tilde{B} \in \mathcal{V}_S$, můžeme funkci $\|\tilde{B} - B\|_F^2$ nahradit funkcí $\|\mathcal{P}_S(\tilde{B} - B)\|_F^2$ a pokud $\tilde{B} \in \mathcal{V}_G$, můžeme funkci $\|\tilde{B} - B\|_F^2$ nahradit funkcí $\|\mathcal{P}_G(\tilde{B} - B)\|_F^2$. V dalším

výkladu bude vhodné upravit kvazinewtonovskou podmínku použitím vektorů $d^i = \mathcal{P}_G(de_i^T)e_i$, $1 \le i \le n$, jejichž prvky jsou určeny vztahy

$$\begin{aligned} d_j^i &= d_j, \quad G_{ij} \neq 0, \\ d_j^i &= 0, \quad G_{ij} = 0, \end{aligned}$$

Význam těchto vektorů spočívá v tom, že pro libovolnou matici $B \in R^{n\times n}$ platí $e_i^T \mathcal{P}_G B\, d = e_i^T B\, d^i$, $1 \le i \le n$ (lze odstranit operátor projekce).

Věta 220. *Nechť $B \in R^{n\times n}$ a nechť $\mathcal{P}_{QS}$, $\mathcal{P}_{QG}$, $\mathcal{P}_{SG}$ jsou operátory orthogonální projekce do $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$. Pak*

$$\begin{aligned} \mathcal{P}_{QS}B &= \mathcal{P}_S B + \frac{(y - \mathcal{P}_S B\, d)d^T + d(y - \mathcal{P}_S B\, d)^T}{d^T d} - \frac{(y - \mathcal{P}_S B\, d)^T d}{d^T d}\ \frac{dd^T}{d^T d}, \\ \mathcal{P}_{QG}B &= \mathcal{P}_G(B + ud^T), \\ \mathcal{P}_{SG}B &= \mathcal{P}_S\mathcal{P}_G B = \mathcal{P}_G\mathcal{P}_S B, \end{aligned}$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = y - \mathcal{P}_G B\, d$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \|d^i\|^2 e_i e_i^T.$$

Důkaz (a) Vztah pro $\mathcal{P}_{QS}$ odvodíme pomocí Lagrangeovy funkce

$$L(\tilde{B}, u, V) = \frac{1}{4}\|\tilde{B} - B\|_F^2 + \sum_{i=1}^{n} u_i \left(y_i - \sum_{j=1}^{n} \tilde{B}_{ij} d_j \right) + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}(\tilde{B}_{ij} - \tilde{B}_{ji}).$$

Podmínky optimality mají tvar

$$\begin{aligned} \frac{\partial L(\tilde{B}, u, V)}{\partial \tilde{B}_{kl}} &= \frac{1}{2}(\tilde{B}_{kl} - B_{kl}) - u_k d_l + v_{kl} - v_{lk} = 0, \\ \frac{\partial L(\tilde{B}, u, V)}{\partial \tilde{B}_{lk}} &= \frac{1}{2}(\tilde{B}_{lk} - B_{lk}) - u_l d_k^l + v_{lk} - v_{kl} = 0, \end{aligned}$$

kde $1 \le k \le n$, $1 \le l \le n$. Sečteme-li obě rovnice a přihlédneme-li k symetrii matice $\tilde{B}$, dostaneme

$$\tilde{B}_{kl} - \frac{1}{2}(B_{kl} + B_{lk}) = u_k d_l^k + u_l d_k^l, \qquad 1 \le i \le n, \quad 1 \le j \le n,$$

což maticově zapsáno dává $\tilde{B} - \mathcal{P}_S B = ud^T + du^T$. Dosadíme-li tento vztah do kvazinewtonovské podmínky $(\tilde{B} - \mathcal{P}_S B)d = y - \mathcal{P}_S B\, d$, dostaneme

$$(d^T d\ I + dd^T)u = (ud^T + du^T)d = (\tilde{B} - \mathcal{P}_S B)d = y - \mathcal{P}_S B\, d,$$

takže $u = (d^T d\ I + dd^T)^{-1}(y - \mathcal{P}_S B\, d)$, což s pomocí Shermanova-Morrisonova vzorce (poznámka 106) dává vztah pro $\mathcal{P}_{QS}$ (podrobněji je tento postup popsán v důkazu věty 89). Poznamenejme, že je-li matice B symetrická, dostaneme metodu PSB uvedenou v poznámce 147.

(b) Předpokládáme-li, že $\tilde{B} \in \mathcal{V}_G$, takže $\mathcal{P}_G\tilde{B} = \tilde{B}$, můžeme kvazinewtonovskou podmínku vyjádřit ve tvaru

$$\tilde{B}d - y = \mathcal{P}_G(\tilde{B} - B)d - (y - \mathcal{P}_G B\, d) = 0,$$

neboli $\mathcal{P}_G(\tilde{B}-B)d = z$, kde $z = y - \mathcal{P}_G B\ d$, což zapsáno po složkách dává

$$e_i^T \mathcal{P}_G(\tilde{B}-B)d = e_i^T(\tilde{B}-B)d^i = \sum_{j=1}^{n}(\tilde{B}_{ij}-B_{ij})d_j^i = z_i, \qquad 1 \le i \le n. \tag{888}$$

Vztah pro $\mathcal{P}_{QG}$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned} L(\tilde{B},u) &= \frac{1}{2}\|\mathcal{P}_G(\tilde{B}-B)\|_F^2 + \sum_{i=1}^{n} u_i \left(z_i - e_i^T \mathcal{P}_G(\tilde{B}-B)d\right) \\ &= \frac{1}{2}\sum_{G_{ij}\neq 0}(\tilde{B}_{ij}-B_{ij})^2 + u^T z - \sum_{G_{ij}\neq 0} u_i(\tilde{B}_{ij}-B_{ij})d_j^i \end{aligned}$$

obsahující kvazinewtonovskou podmínku (888). Derivováním Lagrangeovy funkce podle nenulových prvků matice $\tilde{B}$ dostaneme

$$\frac{\partial L(\tilde{B},u)}{\partial \tilde{B}_{kl}} = \tilde{B}_{kl} - B_{kl} - u_k d_l^k, \qquad G_{kl} \neq 0$$

(derivace podle nulových prvků matice $\tilde{B}$ jsou nulové). Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar

$$\tilde{B}_{kl} - B_{kl} = u_k d_l^k, \quad G_{kl} \neq 0, \tag{889}$$

což lze spolu s požadavkem $\tilde{B}_{kl} = 0$ pro $G_{kl} = 0$ zapsat v maticovém tvaru $\tilde{B} = \mathcal{P}_G(B + ud^T)$. Dosadíme-li vztah (889) do kvazinewtonovské podmínky (888), dostaneme

$$z_i = \sum_{j=1}^{n} u_i d_j^i d_j^i = u_i \|d^i\|^2, \qquad 1 \le i \le n,$$

neboli $Qu = z$, kde Q je diagonální matice vystupující v tvrzení věty (pozitivní semidefinitnost je zřejmá).

(c) Vztahy pro $\mathcal{P}_{SG}B$ plynou ze symetrie Hessovy matice, neboť operátor $\mathcal{P}_S$ zachovává symetrickou řídkou strukturu a operátor $\mathcal{P}_G$ symetrii matice na kterou působí. □

Věta 221. *Nechť $B \in R^{n\times n}$ a nechť $\mathcal{P}_{QSG}$ je operátor ortogonální projekce do $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Pak*

$$\mathcal{P}_{QSG}B = \mathcal{P}_{SG}(B + ud^T + du^T) = \mathcal{P}_G(\mathcal{P}_S B + ud^T + du^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = y - \mathcal{P}_{SG}B\ d$ se symetrickou pozitivně semidefinitní maticí

$$Q = \mathcal{P}_G(dd^T) + \sum_{i=1}^{n} \|d^i\|^2 e_i e_i^T,$$

která má stejnou strukturu jako matice G.

Důkaz Předpokládáme-li, že $\tilde{B} \in \mathcal{V}_S \cap \mathcal{V}_G$, takže $\mathcal{P}_{SG}\tilde{B} = \tilde{B}$, můžeme kvazinewtonovskou podmínku vyjádřit ve tvaru

$$\tilde{B}d - y = \mathcal{P}_{SG}(\tilde{B}-B)d - (y - \mathcal{P}_{SG}B\ d) = 0,$$

neboli $\mathcal{P}_{SG}(\tilde{B}-B)d = \mathcal{P}_G(\tilde{B}-\mathcal{P}_S B)d = z$, kde $z = y - \mathcal{P}_{SG}B\ d$, což zapsáno po složkách dává

$$e_i^T \mathcal{P}_G(\tilde{B}-\mathcal{P}_S B)d = e_i^T(\tilde{B}-\mathcal{P}_S B)d^i = \sum_{j=1}^{n}(\tilde{B}_{ij}-(\mathcal{P}_S B)_{ij})d_j^i = z_i, \qquad 1 \le i \le n \tag{890}$$

(neboť podle věty 220 platí $\mathcal{P}_{SG} = \mathcal{P}_G\mathcal{P}_S$). Vztah pro $\mathcal{P}_{QSG}$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned} L(\tilde{B}, u, V) &= \frac{1}{4}\|\mathcal{P}_G(\tilde{B} - B)\|_F^2 + \sum_{i=1}^{n} u_i \left(z_i - e_i^T \mathcal{P}_G(\tilde{B} - B)d\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} v_{ij}(\tilde{B}_{ij} - \tilde{B}_{ji}) \\ &= \frac{1}{4}\sum_{G_{ij}\neq 0}(\tilde{B}_{ij} - B_{ij})^2 + u^T z - \sum_{G_{ij}\neq 0} u_i(\tilde{B}_{ij} - B_{ij})d_j^i + \sum_{G_{ij}\neq 0} v_{ij}(\tilde{B}_{ij} - \tilde{B}_{ji}) \end{aligned}$$

obsahující kvazinewtonovskou podmínku (890). Derivujeme-li Lagrangeovu funkci podle nenulových prvků matice $\tilde{B}$ a položíme-li tyto derivace rovny nule, dostaneme

$$\begin{aligned} \frac{\partial L(\tilde{B}, u, V)}{\partial \tilde{B}_{kl}} &= \frac{1}{2}(\tilde{B}_{kl} - B_{kl}) - u_k d_l^k + v_{kl} - v_{lk} = 0, \qquad G_{kl} \neq 0, \\ \frac{\partial L(\tilde{B}, u, V)}{\partial \tilde{B}_{lk}} &= \frac{1}{2}(\tilde{B}_{lk} - B_{lk}) - u_l d_k^l + v_{lk} - v_{kl} = 0, \qquad G_{lk} \neq 0. \end{aligned}$$

Sečteme-li obě rovnice a přihlédneme-li k symetrii matice $\tilde{B}$, můžeme podmínky optimality zapsat ve tvaru

$$\tilde{B}_{kl} - (\mathcal{P}_S B)_{kl} = u_k d_l^k + u_l d_k^l, \qquad G_{kl} \neq 0, \tag{891}$$

neboli $\tilde{B} = \mathcal{P}_G(\mathcal{P}_S B + ud^T + du^T)$. Dosadíme-li vztah (891) do kvazinewtonovské podmínky (890), dostaneme

$$z_i = \sum_{j=1}^{n}(u_i d_j^i + u_j d_i^j)d_j^i = \|d^i\|^2 u_i + \sum_{G_{ij}\neq 0} d_i d_j u_j = \|d^i\|^2 u_i + \sum_{j=1}^{n} e_i^T \mathcal{P}_G(dd^T)u_j,$$

neboli $Qu = z$, kde Q je symetrická matice vystupující v tvrzení věty. Matice Q má zřejmě stejnou strukturu jako matice G. Nechť $v \in R^n$ je libovolný vektor. Pak platí

$$v^T Q v = \sum_{i=1}^{n}\sum_{j=1}^{n} d_i^j d_j^i v_i v_j + \sum_{i=1}^{n}\|d^i\|^2 v_i^2 = \sum_{G_{ij}\neq 0} d_i d_j v_i v_j + \sum_{G_{ij}\neq 0} d_j^2 v_i^2 = \frac{1}{2}\sum_{G_{ij}\neq 0}(d_i v_j + d_j v_i)^2 \geq 0 \tag{892}$$

(matice G je symetrická), takže matice Q je pozitivně semidefinitní. Zbývá dokázat, že rovnice $Qu = z$ má řešení. Předpokládejme nejprve, že $\|d^i\| \neq 0$, $1 \leq i \leq n$. Ukážeme, že v tomto případě je matice Q pozitivně definitní. Kdyby matice Q nebyla pozitivně definitní, existoval by vektor $v \neq 0$ takový, že $v^T Q v = 0$. Pak by podle vyjádření (892) musel existovat index $1 \leq i \leq n$ takový, že $v_i \neq 0$ a

$$d_i v_j + d_j v_i = 0, \qquad G_{ij} \neq 0.$$

Jelikož předpokládáme, že $G_{ii} \neq 0$ muselo by nutně platit $d_i v_i = 0$, neboli $d_i = 0$, což po dosazení do poslední rovnosti dává $d_j v_i = 0$ $\forall G_{ij} \neq 0$, neboli $d_j = 0$ $\forall G_{ij} \neq 0$. To je ale ve sporu s předpokladem, že

$$\|d^i\|^2 = \sum_{j=1}^{n}(d_j^i)^2 = \sum_{G_{ij}\neq 0} d_j^2 \neq 0.$$

Předpokládejme nyní, že pro nějaký index $1 \leq i \leq n$ platí $\|d^i\| = 0$. Pak matice Q má nulový i-tý řádek a i-tý sloupec a platí

$$z_i = y_i - \sum_{G_{ij}\neq 0} B_{ij} d_j = \sum_{G_{ij}\neq 0} (\tilde{G}_{ij} - B_{ij})d_j^i = 0$$

(matice $\tilde{G}$ je definovaná vztahem (375)). Můžeme tedy i-tou rovnici vypustit a položit $u_i = 0$. Tímto způsobem můžeme eliminovat všechny nadbytečné rovnice. Zbylá soustava rovnic má pozitivně definitní matici. □

Poznámka 333. V dalších úvahách budeme předpokládat, že $B \in \mathcal{V}_S \cap \mathcal{V}_G$. Pak lze vztah pro $\mathcal{P}_{QSG}$ zapsat ve tvaru

$$\mathcal{P}_{QSG}B = B + \mathcal{P}_G(ud^T + du^T) \tag{893}$$

kde $Qu = y - Bd$.

Metoda s proměnnou metrikou, která používá aktualizaci

$$B_+ = \mathcal{P}_{QSG}B \tag{894}$$

se nazývá Tointovou metodou. Její realizace je poměrně pracná, neboť je třeba řešit dodatečnou soustavu lineárních rovnic $Qu = v$, která může mít nulové řádky a sloupce, což je třeba v eliminačním procesu ohlídat (takže nelze použít stejnou proceduru jako pro určení směrového vektoru). V případě, že matice B je hustá, je tato metoda ekvivalentní metodě PSB, která je neefektivní. Proto byly navrženy další aktualizace, které však v jistém smyslu narušují splnění kvazinewtonovské podmínky. V této práci se budeme zabývat Marwilovou metodou s aktualizací

$$B_+ = \mathcal{P}_S\mathcal{P}_{QG}B, \tag{895}$$

Powellovou metodou s aktualizací

$$B_+ = \mathcal{P}_G\mathcal{P}_{QS}B, \tag{896}$$

a Steihaugovou metodou s aktualizací

$$B_+ = \mathcal{P}_{SG}\mathcal{P}_QB. \tag{897}$$

Lemma 95. *Nechť B_+ je matice určená pomocí některé z aktualizací (894)–(897). Pak platí*

$$B_+ \in \mathcal{V}_S \cap \mathcal{V}_G.$$

Důkaz Pro aktualizaci (894) a (897) je toto tvrzení zřejmé. V případě aktualizace (895) tvrzení plyne z toho, že projekce $\mathcal{P}_S$, určená symetrií matice, neovlivní symetrickou řídkou strukturu. V případě aktualizace (896) tvrzení plyne z toho, že projekce $\mathcal{P}_G$, určená symetrickou řídkou strukturou neovlivní symetrii matice. □

Ve vzorcích (894)–(897) vystupují vždy dva operátory ortogonální projekce $\mathcal{P}_A$, $\mathcal{P}_B$ do lineárních variet $\mathcal{V}_A$, $\mathcal{V}_B$ (v případě Tointovy aktualizace je druhý operátor indentickým operátorem), přičemž platí $\mathcal{V}_A \subset \mathcal{V}_Q$ a $\mathcal{V}_A \cap \mathcal{V}_B = \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$.

Lemma 96. *Nechť $B_+ = \mathcal{P}_B\mathcal{P}_AB$, kde $\mathcal{P}_A, \mathcal{P}_B$ jsou operátory ortogonální projekce do $\mathcal{V}_A$, $\mathcal{V}_B$, přičemž $\mathcal{V}_A \subset R^{n\times n}$, $\mathcal{V}_B \subset R^{n\times n}$ jsou lineární variety takové, že $\mathcal{V}_A \subset \mathcal{V}_Q$ a $\mathcal{V}_A \cap \mathcal{V}_B = \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. Pak pro libovolnou matici $\tilde{G} \in \mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$ platí*

$$\|B_+ - \tilde{G}\|_F^2 \leq \|B - \tilde{G}\|_F^2 - \frac{\|y - Bd\|^2}{\|d\|^2}.$$

Důkaz Jelikož $\tilde{G} \in \mathcal{V}_B$ a $\mathcal{P}_B$ je operátor ortogonální projekce, můžeme použít Pythagorovu větu

$$\|\mathcal{P}_B\mathcal{P}_AB - \tilde{G}\|_F^2 = \|\mathcal{P}_AB - \tilde{G}\|_F^2 - \|\mathcal{P}_AB - \mathcal{P}_B\mathcal{P}_AB\|_F^2 \leq \|\mathcal{P}_AB - \tilde{G}\|_F^2.$$

Jelikož $\mathcal{P}_AB \in \mathcal{V}_A \subset \mathcal{V}_Q$, můžeme psát $\mathcal{P}_ABd = y$, takže platí

$$\|y - Bd\| = \|(\mathcal{P}_AB - B)d\| \leq \|\mathcal{P}_AB - B\|\|d\| \leq \|\mathcal{P}_AB - B\|_F\|d\|.$$

Jelikož $\tilde{G} \in \mathcal{V}_A$ a $\mathcal{P}_A$ je operátor ortogonální projekce, můžeme psát

$$\|\mathcal{P}_AB - \tilde{G}\|_F^2 = \|B - \tilde{G}\|_F^2 - \|B - \mathcal{P}_AB\|_F^2.$$

Spojením všech uvedených vztahů dostaneme

$$\begin{aligned}\|B_+ - \tilde{G}\|_F^2 &= \|\mathcal{P}_B\mathcal{P}_A B - \tilde{G}\|_F^2 \leq \|\mathcal{P}_A B - \tilde{G}\|_F^2 = \|B - \tilde{G}\|_F^2 - \|B - \mathcal{P}_A B\|_F^2 \\ &\leq \|B - \tilde{G}\|_F^2 - \frac{\|y - Bd\|^2}{\|d\|^2}.\end{aligned}$$

□

Nyní se budeme zabývat konvergencí metod s proměnnou metrikou pro řídké úlohy. Omezíme se pouze na metody s lokálně omezeným krokem neboť řídké aktualizace nezaručují pozitivní definitnost aktualizovaných matic.

Věta 222. *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s lokálně omezeným krokem (T1)–(T3) (definice 38). Nechť $B_{i+1} = \mathcal{P}_B\mathcal{P}_A(B_i)$, $i \in N_2$, a $B_{i+1} = B_i$, $i \notin N_2$ ($\mathcal{P}_B\mathcal{P}_A(B_i)$ značí některou z řídkých aktualizací (894)–(897) a množiny N_1, N_2 jsou definovány v poznámce 199). Pak jestliže funkce $F : R^n \to R$ splňuje předpoklady F1, F4 a F6, platí*

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Důkaz (a) nejprve ukážeme, že matice B_i, $i \in N$, jsou dostatečně omezené, neboli že platí $\|B_i\| \leq C_i$, $i \in N$, kde C_i, $i \in N$, jsou čísla splňující rekurentní nerovnosti

$$C_{i+1} \leq C_i + \overline{C}\|d_i\| \leq C_i + \overline{C}\|s_i\|, \tag{898}$$

kde $C_1 > 1$ a $\overline{C} \geq 0$. Nechť $i \in N_2$ a nechť $\tilde{G}_i$ je matice definovaná vztahem (375). Pak platí

$$\begin{aligned}\|\tilde{G}_i - G_i\|_F &= \left\|\int_0^1 (G(x_i + \lambda d_i) - G(x_i))d\lambda\right\|_F \leq \sqrt{n}\int_0^1 \|G(x_i + \lambda d_i) - G(x_i)\|d\lambda \\ &\leq \overline{L}\sqrt{n}\|d_i\|\int_0^1 \lambda d\lambda = \frac{1}{2}\overline{L}\sqrt{n}\|d_i\|\end{aligned} \tag{899}$$

(používáme předpoklad (F6) a skutečnost, že Frobeniova norma není větší než $\sqrt{n}$ násobek spektrální normy). Podobným způsobem dostaneme

$$\|\tilde{G}_i - G_{i+1}\|_F \leq \frac{1}{2}\overline{L}\sqrt{n}\|d_i\|. \tag{900}$$

Použijeme-li nerovnost $\|B_{i+1} - \tilde{G}_i\|_F \leq \|B_i - \tilde{G}_i\|_F$, která plyne z lemmatu 96, můžeme podle (899) a (900) psát

$$\begin{aligned}\|B_{i+1} - G_{i+1}\|_F &\leq \|B_{i+1} - \tilde{G}_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \leq \|B_i - \tilde{G}_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \\ &\leq \|B_i - G_i\|_F + \|\tilde{G}_i - G_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \\ &\leq \|B_i - G_i\|_F + \overline{L}\sqrt{n}\|d_i\|.\end{aligned}$$

Tato nerovnost je splněna i pro $i \notin N_2$, neboť v tomto případě platí $G_{i+1} = G_i$ a $B_{i+1} = B_i$ Použijeme-li tuto nerovnost několikrát po sobě, dostaneme

$$\|B_{i+1} - G_{i+1}\|_F \leq \|B_1 - G_1\|_F + \overline{L}\sqrt{n}\sum_{j=1}^{i}\|d_j\|.$$

neboli

$$\|B_{i+1}\| \leq \|B_{i+1}\|_F \leq 2\overline{G}\sqrt{n} + \|B_1\|\sqrt{n} + \overline{L}\sqrt{n}\sum_{j=1}^{i}\|d_j\|.$$

Položíme-li $C_1 = (2\,\overline{G} + \|B_1\|)\sqrt{n}$ a $\overline{C} = \overline{L}\sqrt{n}$, můžeme psát $\|B_i\| \leq C_i$, $i \in N$, kde čísla C_i, $i \in N$, splňují nerovnosti (898) (neboť podle (T2) platí $\|d_i\| \leq \|s_i\|$, $i \in N$).

(b) Označíme-li

$$M_i = \max_{1 \leq j \leq i} \|B_j\|,$$

platí $M_i \leq C_i$, $i \in N$, a podle poznámky 202 dostaneme

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty,$$

takže můžeme použít větu 118. □

Věta 223. *Nechť jsou splněny předpoklady věty 222 a nechť $x_i \to x^*$ a $\|\omega_i(s_i)\| \to 0$. Jestliže funkce $F : R^n \to R$ splňuje předpoklady F4, F5 a F6, pak $x_i \to x^*$ Q-superlineárně.*

Důkaz Nechť $i \in N_2$. Použijeme-li lemma 96 a nerovnosti (899), (900), můžeme psát

$$\begin{aligned}
\|B_{i+1} - G_{i+1}\|_F^2 &\leq \left(\|B_{i+1} - \tilde{G}_i\|_F + \|G_{i+1} - \tilde{G}_i\|_F\right)^2 \\
&\leq \|B_{i+1} - \tilde{G}_i\|_F^2 + \frac{1}{4}\overline{L}^2 n\|d_i\|^2 + \overline{L}\sqrt{n}\|B_{i+1} - \tilde{G}_i\|_F\|d_i\| \\
&\leq \|B_i - \tilde{G}_i\|_F^2 - \frac{\|y_i - B_i d_i\|^2}{\|d_i\|^2} + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}n\left(\overline{B} + \overline{G}\right)\right)\|d_i\|
\end{aligned}$$

a

$$\begin{aligned}
\|B_i - \tilde{G}_i\|_F^2 &\leq \left(\|B_i - G_i\|_F + \|G_i - \tilde{G}_i\|_F\right)^2 \\
&\leq \|B_i - G_i\|_F^2 + \frac{1}{4}\overline{L}^2 n\|d_i\|^2 + \overline{L}\sqrt{n}\|B_i - G_i\|_F\|d_i\| \\
&\leq \|B_i - G_i\|_F^2 + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}n\left(\overline{B} + \overline{G}\right)\right)\|d_i\|
\end{aligned}$$

(existence konstanty $\overline{\Delta}$ plyne z (T3), existence konstanty $\overline{B}$ plyne z důkazu věty 120 a existence konstanty $\overline{G}$ plyne z (F5)). Spojením obou nerovností dostaneme

$$\frac{\|y_i - B_i d_i\|^2}{\|d_i\|^2} \leq \|B_i - G_i\|_F^2 - \|B_{i+1} - G_{i+1}\|_F^2 + \overline{M}\|d_i\|,$$

kde $2\overline{M} = \overline{L}^2 n\overline{\Delta} + 4\overline{L}n(\overline{B} + \overline{G})$. Tato nerovnost je splněna i pro $i \notin N_2$, neboť v tomto případě platí $d_i = 0$, $y_i = 0$, $G_{i+1} = G_i$ a $B_{i+1} = B_i$. Použijeme-li tuto nerovnost a větu 120, dostaneme

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{\|y_i - B_i d_i\|^2}{\|d_i\|^2} &\leq \sum_{i=1}^{\infty} \left(\|B_i - G_i\|_F^2 - \|B_{i+1} - G_{i+1}\|_F^2\right) + \overline{M}\sum_{i=1}^{\infty}\|d_i\| \\
&\leq \|B_1 - G_1\|_F^2 + \overline{M}\sum_{i=1}^{\infty}\|d_i\| \leq \|B_1 - G_1\|_F^2 + \overline{M}\sum_{i=1}^{\infty}\|s_i\| < \infty. \qquad (901)
\end{aligned}$$

Dále podle (F5) a (899) platí

$$\begin{aligned}
\frac{\|\left(G_i - B_i\right)d_i\|}{\|d_i\|} &\leq \frac{\|(G_i - \tilde{G}_i)d_i\|}{\|d_i\|} + \frac{\|y_i - B_i d_i\|}{\|d_i\|} \\
&\leq \frac{1}{2}\overline{L}\sqrt{n}\|d_i\| + \frac{\|y_i - B_i d_i\|}{\|d_i\|},
\end{aligned}$$

takže

$$\frac{\|\left(G_i - B_i\right) d_i\|}{\|d_i\|} \to 0,$$

neboť $\|d_i\| \to 0$ podle věty 120 a $\|y_i - B_i d_i\|/\|d_i\| \to 0$ podle (901). Jelikož $\|\omega_i(s_i)\| \to 0$ jsou splněny předpoklady věty 122 a $x_i \to x^*$ Q-superlineárně. □

Metody s proměnnou metrikou pro řídké úlohy můžeme také realizovat jako metody spádových směrů, kdy se soustava lineárních rovnic $Bs + g = 0$ řeší nepřesně metodou sdružených gradientů (Algoritmus 6). Použití metody sdružených gradientů je velmi výhodné, neboť tato metoda, aplikovaná na kvadratickou funkci $Q(s)$ s maticí B dává spádové směry bez ohledu na to, jak přesně se řeší soustava rovnic $Bs+g = 0$ (věta 72). I když konvergenční teorie, kterou jsme se dosud zabývali, není aplikovatelná na metody s proměnnou metrikou realizované jako metody spádových směrů (protože matice B nemusí být pozitivně definitní, není zaručeno, že vyřešíme soustavu $Bs + g = 0$ s požadovanou přesností), jsou tyto metody obvykle účinnější než metody s proměnnou metrikou realizované jako metody s lokálně omezeným krokem.

Poznámka 334. Tointovu metodu (894) můžeme upravit tak, že položíme

$$B_+ = \mathcal{P}_{QSG} B^{VM}, \tag{902}$$

kde B^{VM} je symetrická pozitivně definitní matice získaná z matice B metodou s proměnnou metrikou z Broydenovy třídy (vzorec (306) s $\gamma = 1$, $\rho = 1$ a $\beta > \beta^*$), například metodou BFGS. Použijeme-li větu 221 a symetrii matice B^{VM}, dostaneme

$$\mathcal{P}_{QSG} B = \mathcal{P}_G B^{VM} + \mathcal{P}_G(ud^T + du^T), \tag{903}$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = y - \mathcal{P}_G B^{VM} d$ se symetrickou pozitivně semidefinitní maticí Q definovanou ve větě 221. Je zřejmé, že stačí počítat prvky matice $\mathcal{P}_G B^{VM}$, tedy prvky B^{VM}_{ij}, kde $G_{ij} \neq 0$. Numerické testy však ukazují, že tato úprava Tointovy metody není efektivnější než základní metoda (894).

Jak již bylo poznamenáno, Tointova metoda (894) nezaručuje pozitivní definitnost matice B_+. V oddílu 4.3 jsme ukázali, že použitím funkce $\psi(X) = \operatorname{Tr} X - \ln\det X$ lze zajistit pozitivní definitnost matice X (pro jiné než pozitivně definitní matice není tato funkce definována), a že řešením úlohy

$$B_+ = \arg \min_{\tilde{B} \in \mathcal{V}_Q \cap \mathcal{V}_S,\, \tilde{B} \succeq 0} \psi(B^{-1/2} \tilde{B} B^{-1/2})$$

je metoda BFGS. Proto se nabízí určovat řídkou symetrickou matici B_+ řešením úlohy

$$B_+ = \arg \min_{\tilde{B} \in \mathcal{B}_F} \psi(B^{-1/2} \tilde{B} B^{-1/2}), \quad \text{kde} \quad \mathcal{B}_F = \{\tilde{B} \succeq 0 :\ \tilde{B}d = y,\ \tilde{B} = \tilde{B}^T,\ \tilde{B} = \mathcal{P}_G \tilde{B}\}. \tag{904}$$

V dalších úvahách využijeme toho, že matice $B^{-1/2} \tilde{B} B^{-1/2}$ má stejná vlastní čísla jako matice $\tilde{B}H$, kde $H = B^{-1}$, takže $\psi(B^{-1/2} \tilde{B} B^{-1/2}) = \psi(\tilde{B}H)$.

Věta 224. *Nechť symetrická pozitivně definitní matice B_+ je řešením úlohy*

$$B_+ = \arg \min_{\tilde{B} \in \mathcal{B}_F} \psi(\tilde{B}H), \tag{905}$$

kde $H = B^{-1}$ je symetrická pozitivně definitní matice, a nechť $H_+ = B_+^{-1}$. Pak existuje vektor Lagrangeových multiplikátorů $u \in R^n$ takový, že

$$\mathcal{P}_G H_+ = \mathcal{P}_G(H + ud^T + du^T). \tag{906}$$

Důkaz Lagrangeova funkce úlohy (905) má tvar

$$L(\tilde{B},u,V,W)=\psi(\tilde{B}H)+2\sum_{i=1}^{n}u_i\left(y_i-\sum_{j=1}^{n}\tilde{B}_{ij}d_j\right)+\sum_{i=1}^{n}\sum_{j=1}^{n}v_{ij}(\tilde{B}_{ij}-\tilde{B}_{ji})+2\sum_{i=1}^{n}\sum_{j=1}^{n}w_{ij}\tilde{B}_{ij},$$

kde $w_{ij}=0$, pokud $G_{ij}\neq 0$. Označíme-li W matici, jejímiž prvky jsou čísla w_{ij}, můžeme podmínku $w_{ij}=0$, pokud $G_{ij}\neq 0$, zapsat ve tvaru $\mathcal{P}_G W=0$.

(a) Abychom určili parciální derivace Lagrangeovy funkce podle prvků matice $\tilde{B}$, je třeba znát parciální derivace funkce $\psi(\tilde{B}H)=Tr(\tilde{B}H)-\ln\det(\tilde{B}H)=Tr(\tilde{B}H)-\ln\det\tilde{B}-\ln\det H$. Platí

$$\frac{\partial Tr(\tilde{B}H)}{\partial\tilde{B}_{kl}}=\frac{\partial}{\partial\tilde{B}_{kl}}\sum_{i=1}^{n}\sum_{j=1}^{n}\tilde{B}_{ij}H_{ji}=H_{lk},$$

takže $\partial Tr(\tilde{B}H)/\partial\tilde{B}=H^T$. Dále, tak jako v části (b) důkazu lemmatu 42, platí $\partial\ln\det\tilde{B}/\partial\tilde{B}=\tilde{H}^T$, kde $\tilde{H}=\tilde{B}^{-1}$.

(b) Derivujeme-li, tak jako v důkazu lematu 38, zbylé členy Lagrangeovy funkce podle prvků matice $\tilde{B}$, dostaneme

$$\frac{\partial L(\tilde{B},u,V,W)}{\partial\tilde{B}_{kl}}=H_{lk}-\tilde{H}_{lk}+2u_kd_l+v_{kl}-v_{lk}+2w_{kl}.$$

Úplně stejným způsobem dostaneme

$$\frac{\partial L(\tilde{B},u,V,W)}{\partial\tilde{B}_{lk}}=H_{kl}-\tilde{H}_{kl}+2u_ld_k+v_{lk}-v_{kl}+2w_{lk}.$$

Matice H je podle předpokladu symetrická, takže $H_{lk}=H_{kl}$, a nutné podmínky, které musí splňovat řešení úlohy 905, vyžadují symetrii matice $\tilde{B}$, takže $\tilde{H}_{lk}=\tilde{H}_{kl}$. Sečteme-li obě rovnosti, využijeme-li zmíněnou symetrii a položíme-li výsledek roven nule, dostaneme nutné podmínky pro extrém ve tvaru

$$2(H_{kl}-\tilde{H}_{kl})+2(u_kd_l+u_ld_k)+2(w_{kl}+w_{lk})=0,$$

což maticově zapsáno dává $H-\tilde{H}+ud^T+du^T+W+W^T=0$. Převedeme-li matici $\tilde{H}$ na pravou stranu této rovnosti a použijeme-li vztah $\mathcal{P}_G W=0$, dostaneme (906). □

Z rovnice (906) nelze explicitně určit vektor u (nutný ke konstrukci matice B_+) tak, aby byla splněna kvazinewtonovská podmínka $B_+d=y$. V práci [55] je ukázáno, jak lze podmínky (906) a $B_+d=y$ splnit řešením soustavy nelineárních rovnic

$$B_+(u)d-y=0,\quad \mathcal{P}_G B_+^{-1}(u)=\mathcal{P}_G(H+ud^T+du^T),\tag{907}$$

kde u je neznámý vektor a $\mathcal{P}_G B_+=B_+$, v případě, že řádky a sloupce matice B_+ jsou perfektně eliminačně uspořádány (definice 73). Za tohoto předpokladu lze pro daný vektor u sestrojit matice L_+ a D_+ takové, že $\mathcal{P}_G(L_+D_+L_+^T)^{-1}=\mathcal{P}_G(H+ud^T+du^T)$ a navíc získat Jacobiovu matici nelineární soustavy rovnic (907). Tento způsob je však výpočetně velmi náročný, neboť se řeší soustava nelineárních rovnic o n neznámých, která má stejnou dimenzi jako soustava nelineárních rovnic $g(x)=0$ vyjadřující nutné podmínky pro extrém funkce F.

Podle poznámky 150 je metoda používající matici (905) zobecněním metody BFGS. Podobným způsobem můžeme zobecnit metodu DFP tak, že řešíme úlohu

$$H_+=\arg\min_{\tilde{H}\in\mathcal{H}_F}\psi(\tilde{H}H^{-1}),\quad\text{kde}\quad \mathcal{H}_F=\{\tilde{H}\succeq 0:\ \tilde{H}y=d,\ \tilde{H}=\tilde{H}^T,\ \tilde{H}^{-1}=\mathcal{P}_G\tilde{H}^{-1}\}.\tag{908}$$

Vzhledem k tomu, že v definici matice $\mathcal{H}_F$ vystupuje požadavek $\tilde{H}^{-1}=\mathcal{P}_G\tilde{H}^{-1}$ místo požadavku na řídkou strukturu matice $\tilde{H}$, nelze odvodit vztah podobný podmínce (906), takže tento přístup není prakticky

použitelný. V práci [169] je použita podobná myšlenka jako v poznámce 334 a matice H_+ je určována řešením úlohy

$$H_+ = \arg\min_{\tilde{H}\in\mathcal{H}} \psi(\tilde{H}H^{-1}), \quad \text{kde} \quad \mathcal{H} = \{\tilde{H} \succeq 0 : \mathcal{P}_G\tilde{H} = \mathcal{P}_G H^{VM},\ \tilde{H} = \tilde{H}^T,\ \tilde{H}^{-1} = \mathcal{P}_G\tilde{H}^{-1}\} \tag{909}$$

kde H^{VM} je symetrická pozitivně definitní matice získaná z matice H metodou s proměnnou metrikou z Broydenovy třídy (vzorec (286) s $\gamma = 1$, $\rho = 1$ a $\eta > \eta^*$), například metodou BFGS.

Věta 225. *Nechť H je symetrická pozitivně definitní matice taková, že $H^{-1} = \mathcal{P}_G H^{-1}$. Pak úloha (909) je ekvivalentní úloze*

$$H_+ = \arg\max_{\tilde{H}\in\tilde{\mathcal{H}}} \det(\tilde{H}), \quad \text{kde} \quad \tilde{\mathcal{H}} = \{\tilde{H} \succeq 0 : \mathcal{P}_G\tilde{H} = \mathcal{P}_G H^{VM},\ \tilde{H} = \tilde{H}^T\}. \tag{910}$$

Důkaz (a) Dokážeme nejprve, že úloha (909) je ekvivalentní úloze

$$H_+ = \arg\max_{\tilde{H}\in\mathcal{H}} \det(\tilde{H}).$$

Protože $H^{-1} = \mathcal{P}_G H^{-1}$, $\mathcal{P}_G\tilde{H} = \mathcal{P}_G H^{VM}$ a matice H je symetrická, můžeme psát

$$\mathrm{Tr}(\tilde{H}H^{-1}) = \sum_{i=1}^n\sum_{j=1}^n \tilde{H}_{ij}H^{-1}_{ji} = \sum_{i=1}^n\sum_{G_{ij}\neq 0} \tilde{H}_{ij}H^{-1}_{ij} = \sum_{i=1}^n\sum_{G_{ij}\neq 0} H^{VM}_{ij}H^{-1}_{ij},$$

takže stopa $\mathrm{Tr}(\tilde{H}H^{-1})$ nezávisí na prvcích matice $\tilde{H}$ (je konstantní na množině $\mathcal{H}$). Funkce $\psi(\tilde{H}H^{-1}) = \mathrm{Tr}(\tilde{H}H^{-1}) - \ln\det(\tilde{H}H^{-1})$ tedy nabývá svého minima právě tehdy, když $\ln\det(\tilde{H}H^{-1})$ a tedy i $\det(\tilde{H}H^{-1})$ nabývá svého maxima. Protože

$$\det(\tilde{H}H^{-1}) \leq \left(\frac{1}{n}\mathrm{Tr}(\tilde{H}H^{-1})\right)^n$$

a výraz na pravé straně nezávisí na $\tilde{H}$, je $\det(\tilde{H}H^{-1})$ omezený. Jelikož $\det(\tilde{H}H^{-1}) = \det\tilde{H}\det H^{-1}$, je i $\det\tilde{H}$ omezený a $\det(\tilde{H}H^{-1})$ je maximální právě tehdy, když $\det\tilde{H}$ je maximální.

(b) Množina $\tilde{\mathcal{H}}$ je uzavřená (na její hranici platí $\det\tilde{H} = 0$) a má neprázdný vnitřek, neboť $H^{VM} \in \tilde{\mathcal{H}}$ a $\det H^{VM} > 0$. Nechť matice H_+ je řešením úlohy (910). Protože podle (910) platí $\det H_+ \geq \det H^{VM} > 0$, musí matice H_+ ležet v $\tilde{\mathcal{H}}^o$, takže musí splňovat nutné podmínky pro lokální extrém (věta 3). Výraz $\det\tilde{H}$ závisí pouze na proměnných $\tilde{H}_{ij}$, kde $G_{ij} = 0$ (neboť $\tilde{H}_{ij} = H^{VM}_{ij}$, pokud $G_{ij} \neq 0$) a je maximální, je-li $\ln\det\tilde{H}$ maximální. Podle vzorce (361) v důkazu lemmatu 42 platí

$$\frac{\partial\ln\det\tilde{H}}{\partial\tilde{H}_{ij}} = e_j^T\tilde{H}^{-1}e_i = e_i^T\tilde{H}^{-1}e_j$$

(neboť matice $\tilde{H}$ je symetrická), takže $\ln\det\tilde{H}$ (a tedy i $\det\tilde{H}$) nabývá svého maxima pokud $e_i^T\tilde{H}^{-1}e_j = 0$ pro $G_{ij} = 0$, neboli $\tilde{H}^{-1} = \mathcal{P}_G\tilde{H}^{-1}$. □

10.4 Diferenční verze Newtonovy metody pro separovatelné úlohy

Rozsáhlé úlohy jsou často formulovány tak, že platí

$$F(x) = \sum_{k=1}^m f_k(x), \tag{911}$$

kde $m = O(n)$ a kde každá z dílčích funkcí $f_k : R^n \to R$, $1 \leq k \leq m$, závisí na $n_k = O(1)$ proměnných. Pak výpočet hodnoty a gradientu funkce $F(x)$ spotřebuje $O(m) = O(n)$ operací a Hessova matice této

funkce obsahuje $O(m) = O(n)$ nenulových prvků. Gradient a Hessovu matici funkce $F : R^n \to R$ můžeme vyjádřit ve tvaru

$$g(x) = \sum_{k=1}^{m} g_k(x), \qquad G(x) = \sum_{k=1}^{m} G_k(x),$$

kde gradienty $g_k(x)$ a Hessovy matice $G_k(x)$ funkcí $f_k : R^n \to R$, $1 \leq k \leq m$, obsahují $O(1)$ nenulových prvků. Označme

$$\begin{aligned} f(x) &= [f_1(x), \ldots, f_m(x)]^T, \\ J(x) &= [g_1(x), \ldots, g_m(x)]^T. \end{aligned}$$

Pak platí $F(x) = f^T(x)e$, $g(x) = J^T(x)e$, kde $e = [1, \ldots, 1]^T$ je vektor, který obsahuje samé jednotky. Jacobiova matice $J(x)$ je řídká (její k-tý řádek $g_k^T(x)$ obsahuje $O(1)$ nenulových prvků). Hessova matice $G(x)$ má stejnou strukturu jako matice $J^T(x)J(x)$. Struktura Hessovy matice je tedy určena strukturou Jacobiovy matice.

Definiční obory funkcí f_k, $1 \leq k \leq m$, leží v podprostorech $R_k^n \subset R^n$ dimenze $n_k \ll n$. Jelikož podprostor R_k^n je izomorfní s prostorem R^{n_k}, je vhodné zavést redukované veličiny, čímž odpadne nutnost uvažovat strukturálně nulové prvky gradientů a Hessových matic.

Definice 74. *Nechť N_k, $1 \leq k \leq m$, jsou množiny indexů proměnných definujících podprostory R_k^n (v nichž leží definiční obory funkcí $f_k(x)$), a nechť $Z_k \in R^{n \times n_k}$ jsou matice, jejichž sloupce tvoří ortonormální báze v R_k^n (jsou to sloupce jednotkové matice s indexy z N_k). Pak vektory $\hat{x}_k = Z_k^T x$ dimenze n_k nazveme redukovanými vektory proměnných, funkce $\hat{f}_k : R^{n_k} \to R$, pro které platí $\hat{f}_k(\hat{x}_k) = f_k(x)$, nazveme redukovanými dílčími funkcemi, vektory $\hat{g}_k(\hat{x}_k) = Z_k^T g_k(x)$ dimenze n_k nazveme redukovanými gradienty funkcí $f_k(x)$ a symetrické matice $\hat{G}_k(\hat{x}_k) = Z_k^T G_k(x) Z_k$ řádu n_k nazveme redukovanými Hessovými maticemi funkcí $f_k(x)$.*

Poznámka 335. Z praktických důvodů budeme předpokládat, že $n_k > 0$ (neboli $N_k \neq \emptyset$), $1 \leq k \leq m$, a $N_1 \cup \cdots \cup N_m = \{1, \ldots, n\}$. Všechny matice Z_k, $1 \leq k \leq m$, jsou tedy neprázdné (mají alespoň jeden sloupec) a vyskytují se v nich všechny sloupce jednotkové matice řádu n.

Poznámka 336. Redukované gradienty $\hat{g}_k(\hat{x}_k)$ a redukované Hessovy matice $\hat{G}_k(\hat{x}_k)$, jednoznačně určují gradient $g(x)$ a řídkou Hessovu matici $G(x)$ funkce $F(x)$. Platí

$$F(x) = \sum_{k=1}^{m} \hat{f}_k(\hat{x}_k), \qquad g(x) = \sum_{k=1}^{m} Z_k \hat{g}_k(\hat{x}_k), \tag{912}$$

$$G(x) = \sum_{k=1}^{m} Z_k \hat{G}_k(\hat{x}_k) Z_k^T. \tag{913}$$

Pro další úvahy je vhodné klást na redukované funkce $\hat{f}_k$, $1 \leq k \leq m$, předpoklady F3–F6 analogické předpokladům F3–F6 kladeným na funkci F. Budeme přitom předpokládat, že funkce $\hat{f}_k$ jsou definované na otevřených množinách $\hat{\mathcal{D}}_k \subset R^{n_k}$. Při výkladu využijeme toho, že matice Z_k, mají ortonormální sloupce, takže $\|Z_k\| = 1$, $1 \leq k \leq m$.

Předpoklad F3. *Funkce $\hat{f}_k \in \mathcal{C}^1 : \hat{\mathcal{D}}_k \to R$, $1 \leq k \leq m$, mají lipschitzovské první derivace na $\hat{\mathcal{D}}_k$, takže existují konstanty $\overline{G}_k > 0$, $1 \leq k \leq m$, takové, že pro $1 \leq k \leq m$ platí*

$$\|\hat{g}_k(\hat{y}_k) - \hat{g}_k(\hat{x}_k)\| \leq \overline{G}_k \|\hat{y}_k - \hat{x}_k\| \quad \forall \hat{y}_k, \hat{x}_k \in \hat{\mathcal{D}}_k. \tag{914}$$

Poznámka 337. Splňují-li redukované funkce předpoklad F3 a položíme-li $\overline{G} = m \max_{1\le k\le m} \overline{G}_k$, splňuje funkce F předpoklad F3. Platí totiž

$$\begin{aligned}\|g(y) - g(x)\| &= \|\sum_{k=1}^{m} Z_k(\hat{g}_k(\hat{y}_k) - \hat{g}_k(\hat{x}_k))\| \le \sum_{k=1}^{m} \|Z_k\|\overline{G}_k\|\hat{y}_k - \hat{x}_k\| \\ &\le \sum_{k=1}^{m} \overline{G}_k\|y - x\| \le m \max_{1\le k\le m} \overline{G}_k\|y - x\|,\end{aligned}$$

neboť $\|Z_k\| = 1$ a $\|\hat{y}_k - \hat{x}_k\| = \|Z_k^T(y - x)\| \le \|Z_k\|\|y - x\| = \|y - x\|$, $1 \le k \le m$.

Předpoklad F4. *Funkce $\hat{f}_k \in \mathcal{C}^2 : \hat{\mathcal{D}}_k \to R$, $1 \le k \le m$, mají omezené druhé derivace na $\hat{\mathcal{D}}_k$, takže existují konstanty $\overline{G}_k > 0$, $1 \le k \le m$, takové, že pro $1 \le k \le m$ platí*

$$|\hat{d}_k^T \hat{G}_k(\hat{x}_k)\hat{d}_k| \le \overline{G}_k\|\hat{d}_k\|^2 \quad \forall \hat{x}_k \in \hat{\mathcal{D}}_k \quad \forall \hat{d}_k \in R^{n_k}. \tag{915}$$

Poznámka 338. Splňují-li redukované funkce předpoklad F4 a položíme-li $\overline{G} = m \max_{1\le k\le m} \overline{G}_k$, splňuje funkce F předpoklad F4. Plyne to z věty 226, kde $B = G$ a $\hat{B}_k = \hat{G}_k$, a z toho, že $\hat{d}_k = Z_k^T d$ pro $1 \le k \le m$.

Předpoklad F5. *Funkce $\hat{f}_k \in \mathcal{C}^2 : \hat{\mathcal{D}}_k \to R$, $1 \le k \le m$, jsou stejnoměrně silně konvexní na $\hat{\mathcal{D}}_k$, takže existují konstanty $\underline{G}_k > 0$, $1 \le k \le m$, takové, že pro $1 \le k \le m$ platí*

$$\hat{d}_k^T \hat{G}_k(\hat{x}_k)\hat{d}_k \ge \underline{G}_k\|\hat{d}_k\|^2 \quad \forall \hat{x}_k \in \hat{\mathcal{D}}_k \quad \forall \hat{d}_k \in R^{n_k}. \tag{916}$$

Poznámka 339. Splňují-li redukované funkce předpoklad F5 a položíme-li $\underline{G} = m \min_{1\le k\le m} \underline{G}_k$, splňuje funkce F předpoklad F5. Plyne to z věty 226, kde $B = G$ a $\hat{B}_k = \hat{G}_k$, a z toho, že $\hat{d}_k = Z_k^T d$ pro $1 \le k \le m$.

Předpoklad F6. *Funkce $\hat{f}_k \in \mathcal{C}^2 : \hat{\mathcal{D}}_k \to R$, $1 \le k \le m$, mají lipschitzovské druhé derivace na $\hat{\mathcal{D}}_k$, takže existují konstanty $\overline{L}_k > 0$, $1 \le k \le m$, takové, že pro $1 \le k \le m$ platí*

$$\|\hat{G}_k(\hat{y}_k) - \hat{G}_k(\hat{x}_k)\| \le \overline{L}_k\|\hat{y}_k - \hat{x}_k\| \quad \forall \hat{y}_k, \hat{x}_k \in \hat{\mathcal{D}}_k. \tag{917}$$

Poznámka 340. Splňují-li redukované funkce předpoklad F6 a položíme-li $\overline{L} = m \max_{1\le k\le m} \overline{L}_k$, splňuje funkce F předpoklad F6. Platí totiž

$$\begin{aligned}\|G(y) - G(x)\| &= \|\sum_{k=1}^{m} Z_k(\hat{G}_k(\hat{y}_k) - \hat{G}_k(\hat{x}_k))Z_k^T\| \le \sum_{k=1}^{m} \|Z_k\|^2\overline{L}_k\|\hat{y}_k - \hat{x}_k\| \\ &\le \sum_{k=1}^{m} \overline{L}_k\|y - x\| \le m \max_{1\le k\le m} \overline{L}_k\|y - x\|,\end{aligned}$$

neboť neboť $\|Z_k\| = 1$ a $\|\hat{y}_k - \hat{x}_k\| = \|Z_k^T(y - x)\| \le \|Z_k\|\|y - x\| = \|y - x\|$, $1 \le k \le m$.

Diferenční verze Newtonovy metody pro separovatelné úlohy používají místo redukovaných Hessových matic $\hat{G}_k(\hat{x}_k)$, $1 \le k \le m$, jejich aproximace $\hat{B}_k$, $1 \le k \le m$, získané numerickým derivováním. Používají se přitom diferenční vzorce

$$\hat{B}_k\hat{e}_k^j = \frac{\hat{g}_k(\hat{x}_k + \delta\,\hat{e}_k^j) - \hat{g}_k(\hat{x}_k)}{\delta} \approx \hat{G}_k(\hat{x}_k)\hat{e}_k^j, \tag{918}$$

kde $\hat{e}_k^j$, $1 \le j \le n_k$, jsou sloupce jednotkové matice řádu n_k a obvykle $\delta = \sqrt{\varepsilon_M}$. K určení prvků redukovaných Hessových matic je tedy zapotřebí

$$\sum_{k=1}^{m} O(n_k^2) = m\, O(1) = O(m) = O(n)$$

operací. Aproximaci $B \approx G(x)$ Hessovy matice pak vypočteme podle vzorce

$$B = \sum_{k=1}^{m} Z_k \hat{B}_k Z_k^T, \tag{919}$$

který je analogií vzorce (913). Jelikož stopa je lineární maticovou funkcí a $Tr\,(Z_k \hat{B}_k Z_k^T) = Tr\,\hat{B}_k$, můžeme s použitím (919) psát

$$Tr\,B = \sum_{k=1}^{m} Tr\,\hat{B}_k. \tag{920}$$

Věta 226. *Jsou-li matice $\hat{B}_k$, $1 \le k \le m$, pozitivně definitní, je i matice (919) pozitivně definitní a platí*

$$\|B\| \le m \max_{1\le k\le m} \|\hat{B}_k\|, \qquad \|B^{-1}\| \le m \max_{1\le k\le m} \|\hat{B}_k^{-1}\|.$$

Důkaz (a) Jsou-li matice $\hat{B}_k$, $1 \le k \le m$, pozitivně definitní, můžeme pro libovolný nenulový vektor $v \in R^n$ psát

$$v^T B v = \sum_{k=1}^{m} v^T Z_k \hat{B}_k Z_k^T v = \sum_{k=1}^{m} \hat{v}_k^T \hat{B}_k \hat{v}_k > 0,$$

neboť podle poznámky 335 je alespoň jeden z vektorů $\hat{v}_k = Z_k^T v$, $1 \le k \le m$, nenulový.

(b) Zřejmě

$$\|B\| \le \sum_{k=1}^{m} \|Z_k\|\|\hat{B}_k\|\|Z_k\| = \sum_{k=1}^{m} \|\hat{B}_k\| \le m \max_{1\le k\le m} \|\hat{B}_k\|$$

(neboť $\|Z_k\| = 1$, $1 \le k \le m$). Nechť $v \in R^n$ je vlastní vektor matice B, příslušný jejímu nejmenšímu vlastnímu číslu a nechť l je index, pro který platí $\hat{v}_l^T \hat{v}_l = \max_{1\le k\le m} \hat{v}_k^T \hat{v}_k$. Jelikož podle poznámky 335 obsahují sloupce matic Z_k, $1 \le k \le m$, všechny sloupce jednotkové matice, platí

$$v^T v = \sum_{k=1}^{m} v^T e_k e_k^T v \le \sum_{k=1}^{m} v^T Z_k Z_k^T v = \sum_{k=1}^{m} v_k^T v_k \le m v_l^T v_l$$

a můžeme psát

$$\frac{1}{\|B^{-1}\|} = \frac{v^T B v}{v^T v} = \sum_{k=1}^{m} \frac{\hat{v}_k^T \hat{B}_k \hat{v}_k}{v^T v} \ge \frac{1}{m} \frac{\hat{v}_l^T \hat{B}_l \hat{v}_l}{v_l^T v_l} \ge \frac{1}{m} \min_{1\le k\le m} \frac{\hat{v}_k^T \hat{B}_k \hat{v}_k}{\hat{v}_k^T \hat{v}_k} \ge \frac{1}{m} \min_{1\le k\le m} \frac{1}{\|\hat{B}_k^{-1}\|},$$

neboli $\|B^{-1}\| \le m \max_{1\le k\le m} \|\hat{B}_k^{-1}\|$. □

Věta 227. *Nechť redukované funkce $\hat{f}_k$, $1 \le k \le m$, splňují předpoklad F6 a nechť pro $1 \le k \le m$ platí (918), kde $\hat{e}_k^j$, $1 \le j \le n$, jsou sloupce jednotkové matice řádu n_k. Pak lze psát*

$$\|B - G(x)\| \le \frac{1}{2} \overline{L} \sqrt{n} \delta, \tag{921}$$

kde $\overline{L} \le m \max_{1\le k\le m} \overline{L}_k$.

Důkaz Jelikož je splněn předpoklad F6, můžeme pro každou funkci $\hat{f}_k : R^{n_k} \to R$ použít větu 127, takže pro $1 \leq k \leq m$ platí

$$\|\hat{B}_k - \hat{G}_k(\hat{x}_k)\| \leq \frac{1}{2}\overline{L}_k\sqrt{n}\delta.$$

Odtud plyne, že

$$\|B - G(x)\| = \|\sum_{k=1}^{m} Z_k(\hat{B}_k - \hat{G}_k(\hat{x}_k))Z_k^T\| \leq \frac{1}{2}\sqrt{n}\delta\sum_{k=1}^{m}\|Z_k\|^2\overline{L}_k = \frac{1}{2}\sqrt{n}\delta\sum_{k=1}^{m}\overline{L}_k \leq \frac{1}{2}\sqrt{n}\delta\, m \max_{1\leq k\leq m}\overline{L}_k.$$

□

Diferenční verze Newtonovy metody pro separovatelné úlohy se liší od diferenčních verzí Newtonovy metody pro řídké úlohy pouze způsobem získání řídké Hessovy matice $G(x)$ (vzorce (918)–(919)). Všechny ostatní úvahy zůstávají stejné. Lze opět použít realizaci ve formě metody s optimálním lokálně omezeným krokem (oddíl 6.1) nebo realizaci ve formě nepřesné metody s lokálně omezeným krokem (oddíl 6.3).

Numerickým porovnáním diferenčních verzí Newtonovy metody pro separovatelné úlohy s diferenčními verzemi Newtonovy metody pro řídké úlohy lze zjistit, že oba dva typy metod vyžadují přibližně stejný počet operací na jednu iteraci a dávají přibližně stejně přesnou aproximaci Hessovy matice. Metody pro řídké úlohy jsou algoritmicky náročnější (je třeba hledat rozklady sloupců Hessovy matice), ale vzhledem k tomu, že se tyto náročné operace provádějí pouze jednou před zahájením iteračního procesu, je celková doba řešení o něco kratší než u metod pro separovatelné úlohy.

Separovatelné úlohy se obvykle vyznačují tím, že jejich Jacobiovy matice obsahují málo (typicky $O(n)$) nenulových prvků. Protože počet nulových prvků je obvykle mnohem větší, je účelné ukládat pouze nenulové prvky a jen s nimi provádět aritmetické operace. Obdélníkové řídké matice lze ukládat různým způsobem. V tomto oddílu budeme používat pouze komprimované ukládání po řádcích, které používá tři pole $num(A)$, $adr(A)$, $col(A)$, jejichž význam je popsán v oddílu 10.1.

Známe-li řídkou reprezentaci Jacobiovy matice (pole $adr(A)$, $col(A)$) a numerické hodnoty prvků redukovaných Hessových matic, můžeme snadno určit řídkou reprezentaci symetrické Hessovy matice, tedy pole $num(G)$, $adr(G)$, $col(G)$, jejichž význam je popsán v oddílu 10.1. V tomto případě lze redukované Hessovy matice určovat a zpracovávat sekvenčně (není třeba je ukládat současně v paměti počítače).

10.5 Metody s proměnnou metrikou pro separovatelné úlohy

Metody s proměnnou metrikou pro separovatelné úlohy používají místo redukovaných Hessových matic $\hat{G}_k(\hat{x}_k)$, $1 \leq k \leq m$, jejich aproximace $\hat{B}_k$, $1 \leq k \leq m$, které se generují pomocí aktualizací z Broydenovy třídy metod s proměnnou metrikou

$$\hat{B}_k^+ = \frac{1}{\hat{\gamma}_k}\left(\hat{B}_k + \frac{\hat{\gamma}_k}{\hat{\rho}_k}\frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T + \frac{\hat{\beta}_k}{\hat{c}_k}\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)^T\right), \qquad (922)$$

kde $\hat{y}_k = \hat{g}_k^+ - \hat{g}_k = Z_k^T y$ a $\hat{d}_k = \hat{x}_k^+ - \hat{x}_k = Z_k^T d$ jsou vektory dimenze n_k. Přitom $\hat{b}_k = \hat{y}_k^T\hat{d}_k$, $\hat{c}_k = \hat{d}_k^T\hat{B}_k\hat{d}_k$ a $\hat{\gamma}_k > 0$, $\hat{\rho}_k > 0$, $\hat{\beta}_k$ jsou volné parametry. Vzhledem k tomu, že se matice $B \approx G(x)$ konstruuje pomocí redukovaných matic podle vzorce (919), je účelné aby platilo $\hat{B}_k \approx \hat{G}_k(\hat{x}_k)$, takže se většinou pokládá $\hat{\gamma}_k = 1$, $\hat{\rho}_k = 1$, $1 \leq k \leq m$ (jiné volby těchto volných parametrů obyčejně zhoršují rychlost konvergence).

Při vyšetřování konvergence metod s proměnnou metrikou pro separovatelné úlohy budeme používat dvojí indexování. V tom případě budeme index iteračního kroku i umisťovat vpravo dole a index dílčí funkce k vpravo nahoře. Pokud bude možné index i vynechat, budeme index k umisťovat opět vpravo dole (jako ve vzorci (922)). Při dvojím indexování zapíšeme vzorec (922) (s $\rho_i^k = 1$ a $\gamma_i^k = 1$) ve tvaru

$$\hat{B}_{i+1}^k = \hat{B}_i^k + \frac{1}{\hat{b}_i^k}\hat{y}_i^k(\hat{y}_i^k)^T - \frac{1}{\hat{c}_i^k}\hat{B}_i^k\hat{d}_i^k\left(\hat{B}_i^k\hat{d}_i^k\right)^T + \frac{\hat{\beta}_i^k}{\hat{c}_i^k}\left(\frac{\hat{c}_i^k}{\hat{b}_i^k}\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)\left(\frac{\hat{c}_i^k}{\hat{b}_i^k}\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)^T, \qquad (923)$$

Poznamenejme, že aktualizace (923) je definována pouze tehdy, když $\hat{b}_i^k \neq 0$, přičemž rovnost $\hat{b}_i^k = 0$ nastane například tehdy, když $\hat{d}_i^k = Z_k d_i = 0$, neboli $d_i \perp R_k^n$, což nemusí být nikterak výjimečný případ.

U metod s proměnnou metrikou pro separovatelné úlohy nastává problém se zajištěním pozitivní definitnosti generovaných matic. Souvisí to s tím, že není obecně zaručena platnost nerovností $\hat{b}_i^k = (\hat{y}_i^k)^T \hat{d}_i^k > 0$, $1 \leq k \leq m$, takže některé z matic $\hat{B}_{i+1}^k$, $1 \leq k \leq m$, nemusí být pozitivně definitní a tudíž ani matice B_{i+1} nemusí být pozitivně definitní. Tento případ nenastane, jsou-li všechny funkce $\hat{f}_k$, $1 \leq k \leq m$, ryze konvexní a vektory $\hat{d}_i^k$, $1 \leq k \leq m$, nenulové, což budeme v dalších úvahách předpokládat (pokud $\hat{b}_i^k = 0$, můžeme aktualizaci (923) vynechat a položit $\hat{B}_{i+1}^k = \hat{B}_i^k$, aniž by to ovlivnilo důkaz věty 228). Nejprve se budeme zabývat lokální konvergencí metod s proměnnou metrikou pro separovatelné úlohy. Dokážeme větu, která je analogií věty 104.

Věta 228. *(Lokální konvergence) Nechť bod $x^* \in R^n$ je izolovaným lokálním minimem funkce (912), přičemž funkce $\hat{f}_k$, $1 \leq k \leq m$, jsou v bodě x^* dvakrát spojitě diferencovatelné a splňují v jeho okolí předpoklady F4–F6. Uvažujme metodu spádových směrů takovou, že $1/\cos^2\theta_i \leq \kappa(B_i)$ (poznámka 19), kde $B_i = (Z_i^k)^T \hat{B}_i^k Z_i^k$ a $\hat{B}_i^k$, $i \in N$, $1 \leq k \leq m$ jsou posloupnosti pozitivně definitních matic získané aktualizacemi z Broydenovy třídy (923), kde $\hat{b}_i^k > 0$, $i \in N$, $1 \leq k \leq m$. Pak existuje číslo $\delta > 0$ takové, že pokud $\|x_1 - x^*\| < \delta$, platí $x_i \to x^*$ a $\sum_{i=1}^{\infty} \|x_i - x^*\| < \infty$. Jestliže navíc $\|B_i s_i + g_i\|/\|g_i\| \to 0$ a $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2a) a (S3a), pak $x_i \to x^*$ superlineárně.*

Důkaz Označme $e_i = x_i - x^*$ a $\hat{e}_i^k = Z_k^T(x_i - x^*)$, $1 \leq k \leq m$ (vektory $\hat{e}_i^k$, $1 \leq k \leq m$, zde mají jiný význam než vektory $\hat{e}_k^j$ použité ve větě 227).

(a) Jelikož pro libovolný index $1 \leq k \leq m$ splňují funkce $\hat{f}_k$ a posloupnost matic $\hat{B}_i^k$, $i \in N$, předpoklady lemmatu 52, platí

$$\|R_{i+1}^{*\,k}\| \leq \max(1, \|R_i^{*\,k}\|)(1 + O(\|\hat{e}_i^k\|)), \qquad \|(R_{i+1}^{*\,k})^{-1}\| \leq \max(1, \|(R_i^{*\,k})^{-1}\|)(1 + O(\|\hat{e}_i^k\|))$$

a můžeme, tak jako v části (a) důkazu věty 104, psát

$$\|\hat{B}_i^k\| \leq \overline{B}_k \exp\left(\frac{C_k}{2}\sum_{j=1}^{i-1}\|\hat{e}_j^k\|\right), \qquad \|\hat{H}_i^k\| \leq \overline{H}_k \exp\left(\frac{C_k}{2}\sum_{j=1}^{i-1}\|\hat{e}_j^k\|\right)$$

(kde $\hat{H}_i^k = (\hat{B}_i^k)^{-1}$). Položíme-li nyní

$$\overline{B} = m \max_{1\leq k\leq m} \overline{B}_k, \qquad \overline{H} = m \max_{1\leq k\leq m} \overline{H}_k, \qquad C = \max_{1\leq k\leq m} C_k,$$

a použijeme-li větu 226 a vztah

$$\max_{1\leq k\leq m} \|\hat{e}_j^k\| = \max_{1\leq k\leq m} \|Z_k^T e_j\| \leq \max_{1\leq k\leq m} \|Z_k\|\|e_j\| = \|e_j\|$$

(neboť $\|Z_k\| = 1$), můžeme psát

$$\|B_i\| \leq m \max_{1\leq k\leq m} \|\hat{B}_i^k\| \leq \overline{B}\exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right), \qquad \|H_i\| \leq m \max_{1\leq k\leq m} \|\hat{H}_i^k\| \leq \overline{H}\exp\left(\frac{C}{2}\sum_{j=1}^{i-1}\|e_j\|\right),$$

což po vynásobení dává (51) (kde $\kappa_i = \kappa(B_i)$ a $\overline{\kappa} = \overline{B}\,\overline{H}$). Podle poznámky 37 a věty 18 tedy platí $x_i \to x^*$ a $\sum_{i=1}^{\infty}\|e_i\| < \infty$, pokud $x_1 \in \mathcal{B}(x^*, \delta)$, kde číslo δ je určeno vztahem (53).

(b) Jelikož pro libovolný index $1 \leq k \leq m$ splňují funkce $\hat{f}_k$ a posloupnost matic $\hat{B}_i^k$, $i \in N$, předpoklady lemmatu 51 a podle (a) platí

$$\sum_{i=1}^{\infty}\|\hat{e}_i^k\| = \sum_{i=1}^{\infty}\|Z_k^T e_i\| \leq \sum_{i=1}^{\infty}\|e_i\| < \infty,$$

jsou splněny předpoklady věty 103 (v prostoru R^{n_k}), takže

$$\lim_{i\to\infty} \frac{\|(\hat{B}_i^k - \hat{G}_i^k)\hat{d}_i^k\|}{\|\hat{d}_i^k\|} = 0.$$

Použijeme-li (913) a (919) dostaneme

$$\|(B_i - G_i)d_i\| = \|\sum_{k=1}^m Z_k(\hat{B}_i^k - \hat{G}_i^k)Z_k^T d_i\| \le \sum_{k=1}^m \|Z_k(\hat{B}_i^k - \hat{G}_i^k)\hat{d}_i^k\| \le \sum_{k=1}^m \|(\hat{B}_i^k - \hat{G}_i^k)\hat{d}_i^k\|$$

a jelikož platí $\|\hat{d}_i^k\| = \|Z_k^T d_i\| \le \|d_i\|$, můžeme psát

$$\lim_{i\to\infty} \frac{\|(B_i - G_i)d_i\|}{\|d_i\|} \le \sum_{k=1}^m \frac{\|(\hat{B}_i^k - \hat{G}_i^k)\hat{d}_i^k\|}{\|\hat{d}_i^k\|} = 0.$$

Superlineární konvergence je pak bezprostředním důsledkem věty 20. □

Podstatně složitější než dokázat lokální konvergenci je zajistit globální konvergenci metod s proměnnou metrikou pro separovatelné úlohy. Je to způsobeno tím že neplatí $\hat{B}_i^k \hat{d}_i^k = -\alpha_i \hat{g}_i^k$, $i \in N$, $1 \le k \le m$ (vztahy $B_i d_i = -\alpha_i g_i$, $i \in N$, tvoří podstatu důkazu věty 101). V dalších úvahách budeme předpokládat, že funkce $\hat{f}_k$, $1 \le k \le m$, splňují předpoklady F4–F5. Omezíme se na metodu BFGS pro separovatelné úlohy, realizovanou pomocí spádových směrů (definice 17), kde matice B_1^k, $1 \le k \le m$, jsou pozitivně definitní a aktualizace se provádějí podle vzorců

$$\begin{aligned}
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k}\hat{y}_i^k(\hat{y}_i^k)^T - \frac{1}{\hat{c}_i^k}\hat{B}_i^k\hat{d}_i^k\left(\hat{B}_i^k\hat{d}_i^k\right)^T, \quad k \in J(i),\\
\hat{B}_{i+1}^k &= \hat{B}_i^k \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\quad , \quad k \notin J(i),
\end{aligned}$$

$i \in N$, $1 \le k \le m$, kde $J(i) \subset \{1,\dots,m\}$ je vhodně zvolená indexová množina taková, že pokud $k \in J(i)$, platí $(\hat{y}_i^k)^T\hat{d}_i^k > 0$. V důkazu globální konvergence budeme předpokládat, že

$$J(i) = \{k \in N : 1 \le k \le m,\ \|\hat{d}_i^k\|^2 \ge \underline{c}_1 \min(1, \|g_i\|)\,\|d_i\|^2\} \tag{924}$$

a že platí

$$\sum_{k\in J(i)} \|\hat{B}_i^k\hat{d}_i^k\|^2 \ge \underline{c}_2 \min(1, \|g_i\|)\,\|B_i d_i\|^2, \tag{925}$$

kde $0 < \underline{c}_1 \le 1/m$ a $0 < \underline{c}_2 \le 1/m$ jsou vhodně zvolené konstanty.

Poznámka 341. Podle poznámky 335 platí $N_1 \cup \cdots \cup N_m = \{1,\dots,n\}$, takže $\sum_{k=1}^m \|\hat{d}_i^k\|^2 \ge \|d_i\|^2$. Jelikož všechny členy v uvedeném součtu nemohou být menší než $\|d_i\|^2/m$, platí $J(i) \ne \emptyset$. Podobně lze psát

$$\|B_i d_i\| = \left\|\sum_{k=1}^m Z_k\hat{B}_i^k\hat{d}_i^k\right\| \le \sum_{k=1}^m \|\hat{B}_i^k\hat{d}_i^k\|,$$

takže alespoň jeden člen v součtu na pravé straně je větší než $\|B_i d_i\|/\sqrt{m}$ a jeho druhá mocnina je větší než $\|B_i d_i\|^2/m$. Podmínka (925) říká, že $J(i)$ obsahuje indexy členů jejichž součet čtverců není menší než $\underline{c}_2\|B_i d_i\|^2$, kde $\underline{c}_2 \le 1/m$. Platnost této podmínky není automaticky zaručena, ale její porušení je (pro dostatečně malé $\underline{c}_2$) značně nepravděpodobné. Obecnější věta, ve které je podmínka (925) nahražena jinou definicí množiny $J(i)$, je uvedena v [153]. Její důkaz je však formálně komplikovanější než důkaz věty 229 (i když je veden ve stejném duchu).

Věta 229. *Uvažujme metodu BFGS pro separovatelné úlohy, pro kterou platí (924) a (925). Splňují-li funkce $\hat{f}_k$, $1 \le k \le m$, předpoklady F4–F5, je tato metoda globálně konvergentní (definice 14).*

Důkaz Předpokládejme že uvažovaná metoda není globálně konvergentní. Pak podle poznámky 16 existuje číslo $0 < \underline{\varepsilon} \leq 1$ takové, že $\|g_j\| \geq \underline{\varepsilon}\ \forall j \in N$, a podle (924) a (925) pro každý index $i \in N$ platí

$$\|\hat{d}_i^k\|^2 \geq c_1 \|d_i\|^2, \quad k \in J(i), \tag{926}$$

$$\sum_{k \in J(i)} \|\hat{B}_i^k \hat{d}_i^k\|^2 \geq c_2 \|B_i d_i\|^2, \tag{927}$$

kde $c_1 = \underline{c}_1 \underline{\varepsilon}$ a $c_2 = \underline{c}_2 \underline{\varepsilon}$.

(a) Podobně jako v části (a) důkazu věty 100 lze s použitím (920) a (915) psát

$$\begin{aligned} Tr\, B_{i+1} &= \sum_{k=1}^{m} Tr\, \hat{B}_{i+1}^k = \sum_{k=1}^{m} Tr\, \hat{B}_i^k + \sum_{k \in J(i)} \frac{(\tilde{G}_i^k \hat{d}_i^k)^T \tilde{G}_i^k \hat{d}_i^k}{(\hat{d}_i^k)^T \tilde{G}_i^k \hat{d}_i^k} - \sum_{k \in J(i)} \frac{(\hat{B}_i^k \hat{d}_i)^T \hat{B}_i^k \hat{d}_i^k}{(\hat{d}_i^k)^T \hat{B}_i^k \hat{d}_i^k} \\ &\leq Tr\, B_1 + i \sum_{k=1}^{m} \overline{G}_k - \sum_{j=1}^{i} \sum_{k \in J(j)} \frac{(\hat{B}_j^k \hat{d}_j)^T \hat{B}_j^k \hat{d}_j^k}{(\hat{d}_j^k)^T \hat{B}_j^k \hat{d}_j^k}, \end{aligned}$$

takže platí

$$Tr\, B_{i+1} \leq 2i\overline{G}, \qquad \sum_{j=1}^{i} \sum_{k \in J(j)} \frac{(\hat{B}_j^k \hat{d}_j)^T \hat{B}_j^k \hat{d}_j^k}{(\hat{d}_j^k)^T \hat{B}_j^k \hat{d}_j^k} \leq 2i\overline{G}, \tag{928}$$

kde $\overline{G} = \max(Tr\, B_1, \sum_{k=1}^{m} \overline{G}_k)$. Jelikož pro libovolný index $1 \leq k \leq m$ platí $d_j^T B_j d_j \geq (\hat{d}_j^k)^T \hat{B}_j^k \hat{d}_j^k$, můžeme s použitím (927) psát

$$\sum_{j=1}^{i} \sum_{k \in J(j)} \frac{(\hat{B}_j^k \hat{d}_j)^T \hat{B}_j^k \hat{d}_j^k}{(\hat{d}_j^k)^T \hat{B}_j^k \hat{d}_j^k} \geq c_2 \sum_{j=1}^{i} \frac{(B_j d_j)^T B_j d_j}{d_j^T B_j d_j},$$

což spolu s (928) dává

$$\sum_{j=1}^{i} \sum_{k \in J(j)} \frac{d_j^T B_j^2 d_j}{d_j^T B_j d_j} \leq \frac{2im\overline{G}}{c_2}.$$

Označíme-li $\tau(i)$ počet členů v tomto součtu, takže $i \leq \tau(i) \leq mi$, můžeme podle lemmatu 2 psát

$$\prod_{j=1}^{i} \prod_{k \in J(j)} \frac{d_j^T B_j^2 d_j}{d_j^T B_j d_j} \leq \left(\frac{2im\overline{G}}{c_2 \tau(i)} \right)^{\tau(i)} \leq \left(\frac{2m\overline{G}}{c_2} \right)^{\tau(i)}. \tag{929}$$

(b) Jelikož předpokládáme, že uvažovaná metoda není globálně konvergentní, není splněna podmínka (38) a existuje tedy číslo $c > 0$ takové, že pro libovolný index $i \in N$ platí

$$\sum_{j=1}^{i} \cos^2 \theta_j \leq \frac{c}{m} \quad \Rightarrow \quad \sum_{j=1}^{i} \sum_{k \in J(j)} \cos^2 \theta_j \leq c$$

a podle lemmatu 2 též

$$\prod_{j=1}^{i} \prod_{k \in J(j)} \cos^2 \theta_j \leq \left(\frac{c}{\tau(i)} \right)^{\tau(i)}. \tag{930}$$

Jelikož

$$\cos^2 \theta_j = \frac{(g_j^T d_j)^2}{\|g_j\|^2 \|d_j\|^2} = \frac{d_j^T B_j d_j}{d_j^T B_j^2 d_j} \frac{d_j^T B_j d_j}{d_j^T d_j}$$

(neboť $B_j d_j = -\alpha_j g_j$), můžeme podle (929) a (930) psát

$$\prod_{j=1}^{i}\prod_{k\in J(j)}\frac{d_j^T B_j d_j}{d_j^T d_j} = \prod_{j=1}^{i}\prod_{k\in J(j)}\frac{d_j^T B_j^2 d_j}{d_j^T B_j d_j}\cos^2\theta_j \le \left(\frac{2mc\overline{G}}{c_2\tau(i)}\right)^{\tau(i)}. \tag{931}$$

(c) Pro námi uvažovanou metodu BFGS podle (312) (kde $\gamma_i^k = 1$, $\rho_i^k = 1$ a $\beta_i^k = 0$) platí

$$\begin{aligned}\frac{\det \hat{B}_{i+1}^k}{\det \hat{B}_i^k} &= \frac{(\hat{d}_i^k)^T\tilde{G}_i^k\hat{d}_i^k}{(\hat{d}_i^k)^T\hat{B}_i^k\hat{d}_i^k}, \qquad k\in J(i),\\ \frac{\det \hat{B}_{i+1}^k}{\det \hat{B}_i^k} &= 1, \qquad k\notin J(i),\end{aligned}$$

takže tak jako v části (c) důkazu věty 100 s použitím (928) dostaneme

$$\begin{aligned}\prod_{j=1}^{i}\prod_{k\in J(j)}\frac{(\hat{d}_j^k)^T\tilde{G}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{B}_j^k\hat{d}_j^k} &= \prod_{j=1}^{i}\prod_{k\in J(j)}\frac{\det \hat{B}_{j+1}^k}{\det \hat{B}_j^k} = \prod_{k=1}^{m}\frac{\det \hat{B}_{i+1}^k}{\det \hat{B}_1^k} \le \prod_{k=1}^{m}\frac{1}{\det \hat{B}_1^k}\left(\frac{Tr\,\hat{B}_{i+1}^k}{n}\right)^n\\ &\le \prod_{k=1}^{m}\frac{1}{\det \hat{B}_1^k}\left(\frac{2i\overline{G}_k}{n}\right)^n \triangleq \prod_{k=1}^{m}\tilde{c}_k i^n \le \prod_{k=1}^{m}\overline{c}_k^i \le \overline{c}^{\tau(i)},\end{aligned} \tag{932}$$

kde $\log\overline{c}_k \ge \log\tilde{c}_k + n/e$, $1\le k\le m$, a $\overline{c} = \max(1, \prod_{k=1}^m \overline{c}_k)$. Spojíme-li (931) a (932) a přihlédneme-li k tomu, že $(\hat{d}_j^k)^T\hat{B}_j^k\hat{d}_j^k \le d_j^T B_j d_j$ a že podle (926) pro $k\in J(j)$ platí $(\hat{d}_j^k)^T\hat{d}_j^k \ge c_1 d_j^T d_j$, můžeme psát

$$\begin{aligned}\prod_{j=1}^{i}\prod_{k\in J(j)}\frac{(\hat{d}_j^k)^T\tilde{G}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{d}_j^k} &= \prod_{j=1}^{i}\prod_{k\in J(j)}\frac{(\hat{d}_j^k)^T\tilde{G}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{B}_j^k\hat{d}_j^k}\frac{(\hat{d}_j^k)^T\hat{B}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{d}_j^k} \le \overline{c}^{\tau(i)}\prod_{j=1}^{i}\prod_{k\in J(j)}\frac{(\hat{d}_j^k)^T\hat{B}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{d}_j^k}\\ &\le \overline{c}^{\tau(i)}\prod_{j=1}^{i}\prod_{k\in J(j)}\frac{1}{c_1}\frac{d_j^T B_j d_j}{d_j^T d_j} \le \left(\frac{2mc\overline{c}\overline{G}}{c_1c_2\tau(i)}\right)^{\tau(i)} \le \left(\frac{2mc\overline{c}\overline{G}}{c_1c_2 i}\right)^i,\end{aligned}$$

neboť funkce $(a/t)^t$ (kde $a>0$) je pro $t\ge 1$ klesající. Jelikož předpokládáme, že funkce $\hat{f}_k$, $1\le k\le m$, splňují poředpoklad F5, platí $(\hat{d}_j^k)^T\tilde{G}_j^k\hat{d}_j^k/(\hat{d}_j^k)^T\hat{d}_j^k \ge \underline{G}_k$, $1\le k\le m$, takže podle poslední nerovnosti pro libovolný index $i\in N$, dostaneme

$$\underline{G}^{mi} \le \underline{G}^{\tau(i)} \le \prod_{j=1}^{i}\prod_{k\in J(j)}\frac{(\hat{d}_j^k)^T\tilde{G}_j^k\hat{d}_j^k}{(\hat{d}_j^k)^T\hat{d}_j^k} \le \left(\frac{2mc\overline{c}\overline{G}}{c_1c_2 i}\right)^i \quad\Rightarrow\quad \underline{G}^m \le \frac{2mc\overline{c}\overline{G}}{c_1c_2 i},$$

kde $\underline{G} = \min(1, \underline{G}_1, \ldots, \underline{G}_m)$, což je však spor, neboť pravá strana poslední nerovnosti konverguje k nule pokud $i\to\infty$. □

Věta 229 je teoreticky velmi zajímavá, ale obsahuje předpoklady, které v praktických úlohách nebývají splněny. Zejména nejsou zahrnuty případy, kdy některé z funkcí $\hat{f}_k$, $1\le k\le m$, jsou nekonvexní. Ukazuje se však, že při vhodně volbě množin $J(i)$, $i\in N$, jsou podmínky pro globální konvergenci (věta 11) v praktických úlohách většinou splněny. Proto je vhodné zajišťovat globální konvergenci jinak než ve větě 229, například pomocí modifikací matice B, popsaných v oddílu 2.7, nebo pomocí přerušování iteračního procesu (poznámka 32), kdy pokládáme $\hat{B}_k = I$, $1\le k\le m$, pokud $-s^Tg < \varepsilon_0\|s\|\|g\|$ (poznámka 32). Zbývá tedy popsat jak efektivně volit množiny $J(i)$, $i\in N$, v obecném případě.

Nejsou-li všechny funkce f_k, $1\le k\le m$, ryze konvexní, může se stát, že pro $i\in N$ a pro některé indexy $1\le k\le m$ platí $\hat{b}_i^k = (\hat{y}_i^k)^T\hat{d}_i^k \le 0$ a odpovídající matice $\hat{B}_{i+1}^k$ nejsou pozitivně definitní. Proto se nabízí možnost volit množiny $J(i)$, $i\in N$, tak, že

$$J(i) = \{k\in N : 1\le k\le m,\ \hat{b}_i^k \ge c_M,\ \hat{c}_i^k \ge c_M\}, \tag{933}$$

kde c_M je nějaké malé číslo (strojová nula). Numerické experimenty ukazují, že tato metoda (kterou označíme VMB) funguje velmi dobře, lze ji však ještě vylepšit tím, že se aktualizace provádějí i pro některé indexy $k \notin J(i)$. Použití aktualizací BFGS pro $k \notin J(i)$ se ukazuje jako nevhodné, lepší výsledky lze docílit aktualizacemi R1, které jsou méně citlivé na porušení pozitivní definitnosti aktualizované matice. Jednou z možností je použití aktualizací

$$\begin{aligned}
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k}\hat{y}_i^k(\hat{y}_i^k)^T - \frac{1}{\hat{c}_i^k}\hat{B}_i^k\hat{d}_i^k\left(\hat{B}_i^k\hat{d}_i^k\right)^T, && k \in J(i),\\
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k - \hat{c}_i^k}\left(\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)\left(\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)^T, && k \notin J(i), \quad |\hat{b}_i^k - \hat{c}_i^k| \geq c_M,\\
\hat{B}_{i+1}^k &= \hat{B}_i^k, && k \notin J(i), \quad |\hat{b}_i^k - \hat{c}_i^k| < c_M,
\end{aligned}$$

$i \in N$, $1 \leq k \leq m$, kde

$$\begin{aligned}
J(i) &= \{k \in N : 1 \leq k \leq m,\ \hat{b}_i^k \geq c_M,\ \hat{c}_i^k \geq c_M\}, && i = 1\\
J(i) &= \{k \in N : 1 \leq k \leq m,\ \hat{b}_i^k \geq c_M,\ \hat{c}_i^k \geq c_M\} \cap J(i-1), && i > 1.
\end{aligned}$$

Tato metoda (kterou označíme VMC) je většinou horší než základní metoda s indexovou množinou (933), nicméně mnohem lepší než kdybychom místo aktualizací R1 používali aktualizace BFGS. Ukazuje se, že aktualizace R1 jsou výhodné zejména tehdy, když funkce F je součtem většího počtu nekonvexních dílčích funkcí. Nechť $K_i = 0$, pokud $i \leq \underline{i}$, a $K_i = 1$, pokud $i > \underline{i}$, kde $\underline{i} \in N$ je nejmenší číslo, pro které nemá množina $J(\underline{i})$ (vzorec (933)) alespon $m/2$ prvků (pokud takové číslo neexistuje, pokládáme $K_i = 0$, $i \in N$). Pak je výhodné použít aktualizace

$$\begin{aligned}
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k}\hat{y}_i^k(\hat{y}_i^k)^T - \frac{1}{\hat{c}_i^k}\hat{B}_i^k\hat{d}_i^k\left(\hat{B}_i^k\hat{d}_i^k\right)^T, && K_i = 0, \quad k \in J(i)\\
\hat{B}_{i+1}^k &= \hat{B}_i^k, && K_i = 0, \quad k \notin J(i),\\
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k - \hat{c}_i^k}\left(\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)\left(\hat{y}_i^k - \hat{B}_i^k\hat{d}_i^k\right)^T, && K_i = 1, \quad |\hat{b}_i^k - \hat{c}_i^k| \geq c_M,\\
\hat{B}_{i+1}^k &= \hat{B}_i^k + \frac{1}{\hat{b}_i^k}\hat{y}_i^k(\hat{y}_i^k)^T - \frac{1}{\hat{c}_i^k}\hat{B}_i^k\hat{d}_i^k\left(\hat{B}_i^k\hat{d}_i^k\right)^T, && K_i = 1, \quad |\hat{b}_i^k - \hat{c}_i^k| < c_M, \quad k \in J(i)\\
\hat{B}_{i+1}^k &= \hat{B}_i^k, && K_i = 1, \quad |\hat{b}_i^k - \hat{c}_i^k| < c_M, \quad k \notin J(i),
\end{aligned}$$

kde $i \in N$ a $1 \leq k \leq m$ (tuto metodu označíme VMP).

V následující tabulce jsou uvedeny výsledky porovnání metod s aktualizacemi VMB, VMC, VMP, které byly realizovány jako metody spádových směrů s přímým řešením řídkých soustav lineárních rovnic. Je ukázán celkový počet iterací `NIT`, celkový počet použitých funkčních hodnot `NFV`, počet nevyřešených problémů (selhání) a celkový čas výpočtu. Bylo použito 82 testovacích úloh s 1000 proměnnými (TEST25 z [106]) stejných jako v tabulce 11, uvedené v oddílu 10.8.

Metoda	`NIT`	`NFV`	selhání	čas
VMB	21953	29309	2	21.23
VMC	23408	70716	3	57.65
VMP	16004	26310	1	19.15

Známe-li aproximace $\hat{B}_k$, $1 \leq k \leq m$, redukovaných Hessových matic $\hat{G}_k$, $1 \leq k \leq m$, můžeme podle (919) zkonstruovat řídkou aproximaci Hessovy matice G. Metody s proměnnou metrikou však mají jednu nevýhodu, která spočívá v tom, že je třeba uchovávat všechny matice $\hat{B}_k$, $1 \leq k \leq m$. To vyžaduje rezervaci dalších

$$\hat{m} = \sum_{k=1}^{n} \frac{1}{2} n_k(n_k + 1)$$

míst v paměti počítače (číslo $\hat{m}$ je obvykle značně větší než počet nenulových prvků řídké Hessovy matice G).

10.6 Modifikace Gaussovy–Newtonovy metody pro řídké a separovatelné úlohy

Má-li účelová funkce $F(x)$ tvar

$$F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x),$$

má podle (651) Hessova matice stejnou strukturu řídkosti jako matice $J^T(x)J(x)$. Nechť $m = O(n)$. Pak aby matice $G(x)$ měla $O(m) = O(n)$ nenulových prvků, musí mít Jacobiova matice řídké řádky (má-li řádek g_k^T n_k nenulových prvků, má matice $g_k g_k^T$ n_k^2 nenulových prvků). Strukturálně to znamená, že každá z dílčích funkcí $f_k : R^n \to R$, $1 \le k \le m$, musí záviset na $n_k = O(1)$ proměnných. Dostáváme tak speciální případ separovatelné úlohy. Tuto separovatelnou úlohu bychom mohli řešit pomocí diferenčních verzí Newtonovy metody nebo pomocí metod s proměnnou metrikou. Speciální tvar účelové funkce však dovoluje použít některé modifikace Gaussovy-Newtonovy metody, které mohou být mnohem účinnější.

Použijeme-li stejné značení jako v (913), můžeme psát

$$F(x) = \sum_{k=1}^{m} \hat{f}_k^2(\hat{x}_k), \qquad g(x) = \sum_{k=1}^{m} \hat{f}_k(\hat{x}_k) Z_k \hat{g}_k(\hat{x}_k), \tag{934}$$

$$G(x) = \sum_{k=1}^{m} Z_k \hat{g}_k(\hat{x}_k)(Z_k \hat{g}_k(\hat{x}_k))^T + \sum_{k=1}^{m} \hat{f}_k(\hat{x}_k) Z_k \hat{G}_k(\hat{x}_k) Z_k^T, \tag{935}$$

Gaussova-Newtonova metoda určuje směrový vektor $s \in R^n$ řešením normální soustavy rovnic $Bs + g = 0$, kde

$$B = J^T J = \sum_{k=1}^{m} Z_k \hat{g}_k (Z_k \hat{g}_k)^T, \qquad g = J^T f = \sum_{k=1}^{m} \hat{f}_k Z_k \hat{g}_k,$$

nebo řešením přeurčené soustavy rovnic $Js + f \approx 0$. V obou případech je výhodnější používat metody s lokálně omezeným krokem, neboť Jacobiova matice J je často velmi špatně podmíněná. Používá se buď metoda s optimálním lokálně omezeným krokem (oddíl 6.1), založená na řídkém Choleského rozkladu, nebo nepřesná metoda s lokálně omezeným krokem (oddíl 6.3), založená na iteračních metodách popsaných v oddílu 10.7. Protože Gaussova-Newtonova metoda může být neefektivní v případě úloh s velkými rezidui, je výhodné (podobně jako oddílu 8.2) kombinovat tuto metodu s jinými metodami.

Používáme-li matici $B = J^T J$, existují dvě hlavní myšlenky jak vylepšit Gaussovu-Newtonovu metodu. Je to buď její kombinace s metodou s proměnnou metrikou podle pravidel uvedených v poznámce 262 (v tomto případě má nejvýhodnější vlastnosti Marwilova metoda pro řídké úlohy či metoda hodnosti 1 pro separovatelné úlohy), nebo použití aproximace členu druhého řádu v (935) pomocí diferenční verze Newtonovy metody či metody s proměnnou metrikou (opět je nejvýhodnější volit metodu hodnosti 1).

V prvním případě se pro řídké úlohy nabízí jednoduchá kombinace Gaussovy-Newtonovy metody s Marwilovou metodou, kdy v prvním iteračním kroku nebo po restartu pokládáme $B = J^T J$ a po skončení každého iteračního kroku pokládáme

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= \mathcal{P}_S \mathcal{P}_{QG} B \quad , \quad F - F_+ \le \underline{\vartheta} F \end{aligned} \tag{936}$$

(jako v (895)), kde $J_+ = J(x_+)$ a parametr ϑ má stejný význam jako v oddílu 8.2. Globální konvergence této metody (kterou označíme GNS) plyne z věty 160 a věty 222. Superlineární konvergence této metody plyne z věty 160 a věty 223. Poznamenejme, že v (936) se nepoužívají redukované veličiny (se stříškou), ale pouze řídké matice J a B.

Pro separovatené úlohy lze kombinovat Gaussovu-Newtonovu metodu s metodou hodnosti 1. Matice B_i se v tomto případě počítá podle vzorce

$$B_i = \sum_{k=1}^{m} Z_k \hat{B}_i^k Z_k^T$$

(používáme dvojí indexování jako v oddílu 10.5), kde v prvním iteračním kroku (pro $i = 1$) nebo po restartu pokládáme $\hat{B}_i^k = \hat{g}_i^k (\hat{g}_i^k)^T$ a po skončení i-tého iteračního kroku pokládáme

$$\begin{aligned}
\hat{B}_{i+1}^k = \hat{g}_{i+1}^k (\hat{g}_{i+1}^k)^T \quad &, \quad F_i - F_{i+1} > \underline{\vartheta} F_i, \\
\hat{B}_{i+1}^k = \hat{B}_i^k \quad &, \quad F_i - F_{i+1} \le \underline{\vartheta} F_i, \quad |(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k| < c_M, \\
\hat{B}_{i+1}^k = \hat{B}_i^k + \frac{(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T}{(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k} \quad &, \quad F_i - F_{i+1} \le \underline{\vartheta} F_i, \quad |(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k| \ge c_M
\end{aligned} \tag{937}$$

(vše pro $1 \le k \le m$), kde $\hat{y}_i^k = \hat{f}_{i+1}^k \hat{g}_{i+1}^k - \hat{f}_i^k \hat{g}_i^k$ a $\hat{d}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k$. Tuto metodu označíme GNP.

Ve druhém případě, kdy se aproximuje člen druhého řádu, existuje několik možností. Předně je možné tento člen aproximovat pomocí diferenční verze Newtonovy metody, kdy v prvním iteračním kroku (pro $i = 1$) nebo po restartu pokládáme $B_i = J_i^T J_i$ a po skončení i-tého iteračního kroku pokládáme

$$\begin{aligned}
B_{i+1} = J_{i+1}^T J_{i+1} \quad &, \quad F_i - F_{i+1} > \underline{\vartheta} F_i, \\
B_{i+1} = J_{i+1}^T J_{i+1} + \sum_{k=1}^{m} \hat{f}_{i+1}^k Z_k \hat{G}_{i+1}^k Z_k^T \quad &, \quad F_i - F_{i+1} \le \underline{\vartheta} F_i,
\end{aligned} \tag{938}$$

kde $J_{i+1} = J(x_{i+1})$, $\hat{f}_{i+1}^k = \hat{f}_k(\hat{x}_{i+1}^k)$ a $\hat{G}_{i+1}^k \approx \hat{G}_k(\hat{x}_{i+1}^k)$ pro $1 \le k \le m$, ($\hat{G}_{i+1}^k$ je diferenční aproximace Hessovy matice funkce $\hat{f}_k$ v bodě $\hat{x}_{i+1}^k$ určená podle vzorce (918)). Globální a superlineární konvergence této metody (kterou označíme GNN) plyne (pro $\hat{G}_{i+1}^k = \hat{G}_k(\hat{x}_{i+1}^k)$, $1 \le k \le m$) z věty 125 a věty 160.

Další možností je aproximace členu druhého řádu pomocí metody s proměnnou metrikou hodnosti 1, kdy v prvním iteračním kroku (pro $i = 1$) nebo po restartu pokládáme $B_i = J_i^T J_i$ a $\hat{B}_i^k = \hat{I}_k$, $1 \le k \le m$ ($\hat{B}_i^k$ je aproximace Hessovy matice $\hat{G}_i^k$ a $\hat{I}_k$ je jednotková matice řádu n_k). Po skončení každého iteračního kroku pokládáme

$$\begin{aligned}
\hat{B}_{i+1}^k = \hat{B}_i^k \quad &, \quad |(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k| < c_M, \\
\hat{B}_{i+1}^k = \hat{B}_i^k + \frac{(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T}{(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k} \quad &, \quad |(\hat{y}_i^k - \hat{B}_i^k \hat{d}_i^k)^T \hat{d}_i^k| \ge c_M,
\end{aligned}$$

pro $1 \le k \le m$, a

$$\begin{aligned}
B_{i+1} = J_{i+1}^T J_{i+1} \quad &, \quad F_i - F_{i+1} > \underline{\vartheta} F_i, \\
B_{i+1} = J_{i+1}^T J_{i+1} + \sum_{k=1}^{m} \hat{f}_{i+1}^k Z_k \hat{B}_{i+1}^k Z_k^T \quad &, \quad F_i - F_{i+1} \le \underline{\vartheta} F_i.
\end{aligned} \tag{939}$$

Přitom $\hat{y}_i^k = \hat{g}_{i+1}^k - \hat{g}_i^k$, $\hat{d}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k$, $1 \le k \le m$. Tuto metodu označíme GNB.

Metoda GNB je založena na aproximaci Hessových matic $\hat{G}_k(\hat{x}_k)$ funkcí $\hat{f}_k(\hat{x}_k)$, $1 \le k \le m$. Podobným způsobem jako v oddílu 8.2 lze aproximovat přímo členy druhého řádu $\hat{f}_k(\hat{x}_k)\hat{G}_k(\hat{x}_k)$ v (935). Používají se k tomu vzorce analogické vzorcům (673) a (676). Pak v prvním iteračním kroku (pro $i = 1$) nebo po restartu pokládáme $B_i = J_i^T J_i$ a $\hat{C}_i^k = \hat{I}_k$, $1 \le k \le m$. Po skončení každého iteračního kroku pokládáme

$$\begin{aligned}
\hat{C}_{i+1}^k = \hat{C}_i^k \quad &, \quad |(\hat{z}_i^k - \hat{C}_i^k \hat{d}_i^k)^T \hat{d}_i^k| < c_M, \\
\hat{C}_{i+1}^k = \hat{C}_i^k + \frac{(\hat{z}_i^k - \hat{C}_i^k \hat{d}_i^k)(\hat{z}_i^k - \hat{C}_i^k \hat{d}_i^k)^T}{(\hat{z}_i^k - \hat{C}_i^k \hat{d}_i^k)^T \hat{d}_i^k} \quad &, \quad |(\hat{z}_i^k - \hat{C}_i^k \hat{d}_i^k)^T \hat{d}_i^k| \ge c_M,
\end{aligned}$$

pro $1 \le k \le m$, a

$$B_{i+1} = J_{i+1}^T J_{i+1} \quad , \quad F_i - F_{i+1} > \underline{\vartheta} F_i,$$
$$B_{i+1} = J_{i+1}^T J_{i+1} + \sum_{k=1}^{m} Z_k \hat{C}_{i+1}^k Z_k^T \quad , \quad F_i - F_{i+1} \le \underline{\vartheta} F_i. \tag{940}$$

Přitom $\hat{z}_i^k = \hat{f}_{i+1}^k(\hat{g}_{i+1}^k - \hat{g}_i^k)$, $\hat{d}_i^k = \hat{x}_{i+1}^k - \hat{x}_i^k$, $1 \le k \le m$. Tuto metodu označíme GNC.

Ačkoliv pro metody GNB a GNC (bez jakýchkoliv úprav) nejsou dokázány konvergenční věty, jsou jejich numerické vlastnosti velmi dobré. Jedinou nevýhodou těchto metod (podobně jako metod s proměnnou metrikou pro separovatelné úlohy) je nutnost ukládat současně všechny redukované matice $\hat{B}_i^k$ nebo $\hat{C}_i^k$, $1 \le k \le m$. Ukládá se tedy $\sum_{k=1}^m n_k(n_k+1)/2$ prvků, což může značně převýšit počet nenulových prvků matice B_i.

Normální soustavu rovnic $Bs + g = 0$ nelze použít, má-li Jacobiova matice J alespoň jeden hustý řádek (takže $n_k \sim n$ pro nějaký index $1 \le k \le m$). V tomto případě je matice $B = J^T J$ hustá (stejnou strukturu má matice G) a je tudíž nutné pracovat přímo s řídkou Jacobiovou maticí J, takže možnosti použití informací druhého řádu jsou značně omezené.

Jednou z možností je nahradit (v případě, že $F - F_+ \le \underline{\vartheta} F$) matici J_+ maticí $A_+ = J_+ + uv^T$, získanou z matice J_+ aktualizací hodnosti 1, kde vektory u a v se vybírají tak, aby matice $A_+^T A_+$ vznikla z matice $J_+^T J_+$ aktualizací BFGS. Tato možnost je vyšetřovaná v oddílu 4.2 (vzorec (350)). Dosadíme-li $A = J_+$ a $\tilde{d} = Ad = J_+ d$ do vzorce (350) (kde $\gamma = 1$ a $\rho = 1$), dostaneme

$$A_+ = A - \frac{1}{\tilde{d}^T\tilde{d}}\tilde{d}\left(A^T\tilde{d} + \sqrt{\frac{\tilde{d}^T\tilde{d}}{d^Ty}}y\right)^T = J_+ - \frac{\tilde{d}}{\|\tilde{d}\|}\left(J_+^T\frac{\tilde{d}}{\|\tilde{d}\|} + \sqrt{\frac{1}{d^Ty}}y\right)^T = J_+ + uv^T,$$

kde $u = J_+d/\|J_+d\|$ a $v = y/\sqrt{d^Ty} - J_+^T u$. Řešíme tedy přeurčenou soustavu rovnic $As + f \approx 0$, kde v prvním iteračním kroku nebo po restartu pokládáme $A = J$ a po skončení každého iteračního kroku pokládáme

$$A_+ = J_+ \quad , \quad F - F_+ > \underline{\vartheta} F,$$
$$A_+ = J_+ + \frac{J_+d}{\|J_+d\|}\left(\frac{y}{\sqrt{d^Ty}} - J_+^T\frac{J_+d}{\|J_+d\|}\right)^T \quad , \quad F - F_+ \le \underline{\vartheta} F. \tag{941}$$

Tuto metodu označíme GNV.

Jinou možností je nahradit (v případě, že $F - F_+ \le \underline{\vartheta} F$) přeurčenou soustavu rovnic $J_+s_+ + f_+ \approx 0$, rozšířenou soustavou

$$A_+s_+ + a_+ = \begin{bmatrix} J_+ \\ w^T \end{bmatrix} s_+ + \begin{bmatrix} f_+ \\ 0 \end{bmatrix} \approx 0, \tag{942}$$

kde vektor w se vybírá tak, aby matice $A_+^T A_+$ vznikla z matice $J_+^T J_+$ aktualizací R1. Jelikož podle (942) platí $A_+^T A_+ = J_+^T J_+ + ww^T$, je tato podmínka splněna tehdy, když $w = (y - J_+^T J_+ d)/\sqrt{d^T(y - J_+^T J_+ d)}$. Řešíme tedy přeurčenou soustavu rovnic $As + a \approx 0$, kde v prvním iteračním kroku nebo po restartu pokládáme $A = J$, $a = f$ a po skončení každého iteračního kroku pokládáme

$$A_+ = J_+, \quad a_+ = f_+, \quad F - F_+ > \underline{\vartheta} F,$$
$$A_+ = J_+, \quad a_+ = f_+, \quad F - F_+ \le \underline{\vartheta} F, \quad d^T(y - J_+^T J_+ d) \le 0$$
$$A_+ = \begin{bmatrix} J_+ \\ \frac{(y - J_+^T J_+ d)^T}{\sqrt{d^T(y - J_+^T J_+ d)}} \end{bmatrix}, \quad a_+ = \begin{bmatrix} f_+ \\ 0 \end{bmatrix}, \quad F - F_+ \le \underline{\vartheta} F, \quad d^T(y - J_+^T J_+ d) \le 0. \tag{943}$$

Tuto metodu označíme GNR.

10.7 Iterační řešení rozsáhlých lineárních úloh nejmenších čtverců

Řídkou reprezentaci Hessovy matice nemůžeme použít, má-li Jacobiova matice J alespoň jeden hustý řádek ($n_k \sim n$ pro nějaký index $1 \leq k \leq m$). V tomto případě je matice $J^T J$ hustá (stejnou strukturu má matice G) a je tudíž třeba pracovat s řídkou reprezentací Jacobiovy matice. Jednou z možností je nahradit normální soustavu rovnic $J^T Js + J^T f = 0$ rozšířenou soustavou rovnic

$$\begin{bmatrix} I, & J \\ J^T, & 0 \end{bmatrix} \begin{bmatrix} r \\ s \end{bmatrix} + \begin{bmatrix} f \\ 0 \end{bmatrix} = 0,$$

jejíž matice je řídká (je-li matice J řídká) a indefinitní. Tuto soustavu lze řešit řídkou verzí Bunchovy-Parlettovy eliminační metody, která je implementovaná v programu MA27 z knihovny AERE Harwell. V tomto oddílu se zaměříme na některé úpravy metody sdružených gradientů pro řešení normální soustavy rovnic $J^T Js + J^T f = 0$.

Nejjednodušší úpravou metody CG pro řešení normální soustavy rovnic je metoda CGNE.

Definice 75. *Nechť $J \in R^{m \times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \quad u_1 = f, \quad g_1 = J^T f, \quad p_1 = -g_1$$

a

$$v_i = Jp_i \quad \alpha_i = \|g_i\|^2 / \|v_i\|^2,$$

$$s_{i+1} = s_i + \alpha_i p_i, \quad u_{i+1} = u_i + \alpha_i v_i,$$

$$g_{i+1} = J^T u_{i+1}, \quad \beta_i = \|g_{i+1}\|^2 / \|g_i\|^2,$$

$$p_{i+1} = -g_{i+1} + \beta_i p_i$$

pro $1 \leq i \leq n$, kde $u_i \in R^m$, $v_i \in R^m$, $1 \leq i \leq n$, nazveme metodou CGNE určenou maticí $J \in R^{m \times n}$ a vektorem $f \in R^m$.

Snadno se přesvědčíme (položíme-li $B = J^T J$ a $q_i = J^T v_i$, $1 \leq i \leq n$), že metoda CGNE je ekvivalentní metodě CG popsané v oddílu 3.8. Vlastnosti metody CGNE se příliš neliší od vlastností metody CG. Jestliže však $m \gg n$, vyžaduje metoda CGNE větší počet operací a má větší paměťové nároky než metoda CG.

Mnohem lepší stabilitu než metoda CGNE mají metody založené na použití bidiagonalizačního Lanczosova procesu. Z praktických důvodů budeme v následující definici používat koeficienty α_i, β_i, $1 \leq i \leq n$, které nemají nic společného se stejně označenými koeficienty vystupujícími v definici 75.

Definice 76. *Nechť $J \in R^{m \times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$\beta_1 u_1 = f, \quad \alpha_1 q_1 = J^T u_1$$

a

$$\beta_{i+1} u_{i+1} = Jq_i - \alpha_i u_i,$$
$$\alpha_{i+1} q_{i+1} = J^T u_{i+1} - \beta_{i+1} q_i$$

pro $1 \leq i \leq n$, kde koeficienty α_i, β_i, $1 \leq i \leq n$ se volí tak, aby vektory $u_i \in R^m$, $q_i \in R^n$, $1 \leq i \leq n$ měly jednotkovou normu, nazveme bidiagonalizačním Lanczosovým procesem (BL) určeným maticí $J \in R^{m \times n}$ a vektorem $f \in R^m$.

Poznámka 342. Nechť $\alpha_i \neq 0$, $\beta_i \neq 0$, $1 \leq i \leq k$ pro nějaký index $1 \leq k \leq n$. Pak podle definice 76 platí $f = U_{k+1}(\beta_1 e_1)$ a

$$JQ_k = U_{k+1}B_k, \tag{944}$$

$$J^T U_{k+1} = Q_k B_k^T + \alpha_{k+1} q_{k+1} e_{k+1}^T, \tag{945}$$

kde $Q_k = [q_1, q_2, \ldots, q_k]$, $U_{k+1} = [u_1, u_2, \ldots, u_k, u_{k+1}]$, $e_1^T = [1, 0, \ldots, 0, 0]$, $e_{k+1}^T = [0, 0, \ldots, 0, 1]$ a

$$B_k = \begin{bmatrix} \alpha_1, & 0, & \ldots, & 0 \\ \beta_2, & \alpha_2, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \alpha_k \\ 0, & 0, & \ldots, & \beta_{k+1} \end{bmatrix},$$

kde $B_k \in R^{(k+1)\times k}$ je bidiagonální horní Hessenbergova matice.

Věta 230. *Uvažujme bidiagonalizační Lanczosův proces určený maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$. Nechť $\alpha_i \neq 0$, $\beta_i \neq 0$, $1 \leq i \leq k$, pro nějaký index $1 \leq k \leq n$. Pak vektory q_i, $1 \leq i \leq k$, tvoří ortonormální bázi v Krylovově podprostoru $\mathcal{K}_i = \text{span}\,(g, Bg, \ldots, B^{k-1}g)$, kde $B = J^T J$ a $g = J^T f$, a vektory u_i, $1 \leq i \leq k$, jsou vzájemně ortogonální a mají jednotkovou normu.*

Důkaz (indukcí). Pro $k = 1$ je tvrzení zřejmé, neboť $q_1 = J^T f/\|J^T f\|$ a $u_1 = f/\|f\|$. Předpokládejme, že tvrzení platí pro nějaký index $1 \leq k < n$ a že $\alpha_{k+1} \neq 0$, $\beta_{k+1} \neq 0$. Použijeme-li vztahy (944)–(945), dostaneme

$$J^T JQ_k = J^T U_{k+1}B_k = Q_k B_k^T B_k + \alpha_{k+1} q_{k+1} e_{k+1}^T B_k = Q_k T_k + \alpha_{k+1}\beta_{k+1} q_{k+1} e_k^T,$$

kde

$$T_k = B_k^T B_k = \begin{bmatrix} \alpha_1^2 + \beta_2^2, & \alpha_2\beta_2, & \ldots, & 0, & 0 \\ \alpha_2\beta_2, & \alpha_2^2 + \beta_3^2, & \ldots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \alpha_{k-1}^2 + \beta_k^2, & \alpha_k\beta_k \\ 0, & 0, & \ldots, & \alpha_k\beta_k, & \alpha_k^2 + \beta_{k+1}^2 \end{bmatrix}$$

je symetrická tridiagonální matice řádu k. Platí tedy (593), kde $B = J^T J$, $T_k = B_k^T B_k$ a $\gamma_i = \alpha_i^2 + \beta_{i+1}^2$, $\delta_i = \alpha_i\beta_i$, $1 \leq i \leq k$, a můžeme použít větu 141, podle které tvoří vektory q_i, $1 \leq i \leq k+1$ bázi v Krylovově podprostoru $\mathcal{K}_i = \text{span}\,(g, Bg, \ldots, B^{k-1}g)$, kde $B = J^T J$ a $g = J^T f$. Použijeme-li (944), dostaneme

$$U_{k+1}^T JQ_k = U_{k+1}^T U_{k+1} B_k$$

a podle (945) platí

$$U_{k+1}^T JQ_k = B_k Q_k^T Q_k + \alpha_{k+1} e_{k+1} q_{k+1}^T Q_k = B_k,$$

takže $U_{k+1}^T U_{k+1} = I$ (vektory u_i, $1 \leq i \leq k+1$, jsou vzájemně ortogonální a mají jednotkovou normu). □

Poznámka 343. Z důkazu věty 230 plyne, že symetrický Lanzcosův proces určený symetrickou pozitivně definitní maticí $B \in R^{n\times n}$ a vektorem $g \in R^n$ je ekvivalentní bidiagonalizačnímu Lanzcosovu procesu určenému maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$, pokud $B = J^T J$ a $g = J^T f$. Ekvivalence spočívá v tom, že oba dva procesy generují stejné vektory q_i, $1 \leq i \leq k$, a platí $\gamma_i = \alpha_i^2 + \beta_{i+1}^2$, $\delta_i = \alpha_i\beta_i$, $1 \leq i \leq k$, kde k je index takový, že $\alpha_i \neq 0$, $\beta_i \neq 0$, $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \leq i \leq k$.

Poznámka 344. Bidiagonalizační Lanczosův proces lze použít k řešení soustavy rovnic $J^T Js + J^T g = 0$. Pokládáme $s_1 = 0$ a vektory s_{i+1}, $1 \leq i \leq k$, hledáme tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \|Js + f\|,$$

Jelikož $s \in \mathcal{K}_i$ právě tehdy, když $s = Q_i z$, kde $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i} \|B_i z + \beta_1 e_1\|$$

(plyne to ze vztahů $f = U_{i+1}(\beta_1 e_1)$, $JQ_i = U_{i+1} B_i$ a $U_{i+1}^T U_{i+1} = I$). Pokud $\alpha_{k+1}\beta_{k+1} = 0$ je vektor $s_{k+1} \in \mathcal{K}_k$, řešením soustavy rovnic $J^T Js + J^T f = 0$ (plyne to z poznámky 229 a poznámky 343).

Poznámka 345. Směrové vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 344 jsou shodné se směrovými vektory s_{i+1}, $1 \leq i \leq k$, generovanými metodou CGNE určenou maticí $J \in R^{m \times n}$ a vektorem $f \in R^m$ (plyne to z věty 142 a poznámky 343).

Výhodou bidiagonalizačního Lanczosova procesu je skutečnost, že vektory s_{i+1}, $1 \leq i \leq k$, mohou být určeny pomocí stabilních operací. To tvoří základ metody LSQR. Princip metody LSQR spočívá v tom, že se rekurentně určují rozklady

$$P_i B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \qquad P_i(\beta_1 e_1) = \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix},$$

kde

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \ldots, & 0 \\ 0, & \rho_2, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \rho_i \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ -- \\ \eta_i \end{bmatrix}.$$

Přitom P_i, $1 \leq i \leq k$, jsou ortogonální matice dimenze $i+1$ a R_i, $1 \leq i \leq k$, jsou bidiagonální horní trojúhelníkové matice dimenze i. Matice P_i, $1 \leq i \leq k$, se počítají rekurentně pomocí Givensových matic elementárních rotací $\tilde{P}_i \in R^{2\times 2}$, $1 \leq i \leq k$, studovaných v oddílu 8.4.

Poznámka 346. Předpokládejme, že

$$\begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} [B_i, b_{i+1}, \delta_1 e_1] = \begin{bmatrix} \rho_1, & \sigma_2, & \ldots, & 0, & 0, & 0, & \eta_1 \\ 0, & \rho_2, & \ldots, & 0, & 0, & 0, & \eta_2 \\ -- & -- & -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \rho_{i-1}, & \sigma_i, & 0, & \eta_{i-1} \\ 0, & 0, & \ldots, & 0, & \tilde{\rho}_i, & 0, & \tilde{\eta}_i \\ 0, & 0, & \ldots, & 0, & \beta_{i+1}, & \alpha_{i+1}, & 0 \end{bmatrix}, \tag{946}$$

kde b_{i+1} je vektor, který obsahuje prvních $i+1$ prvků posledního sloupce matice B_{i+1}. Abychom vynulovali prvek β_{i+1}, sestrojíme Givensovu ortogonální matici

$$\tilde{P}_i = \frac{1}{\sqrt{\tilde{\rho}_i^2 + \beta_{i+1}^2}} \begin{bmatrix} \tilde{\rho}_i, & \beta_{i+1} \\ -\beta_{i+1}, & \tilde{\rho}_i \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i, & \beta_{i+1} \\ -\beta_{i+1}, & \tilde{\rho}_i \end{bmatrix}, \qquad \rho_i = \sqrt{\tilde{\rho}_i^2 + \beta_{i+1}^2}.$$

Pak podle věty 144 platí

$$\tilde{P}_i \begin{bmatrix} \tilde{\rho}_i \\ \beta_{i+1} \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i^2 + \beta_{i+1}^2 \\ 0 \end{bmatrix} = \begin{bmatrix} \rho_i \\ 0 \end{bmatrix}, \qquad \tilde{P}_i \begin{bmatrix} \tilde{\eta}_i \\ 0 \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \tilde{\rho}_i \tilde{\eta}_i \\ -\beta_{i+1}\tilde{\eta}_i \end{bmatrix} = \begin{bmatrix} \eta_i \\ \tilde{\eta}_{i+1} \end{bmatrix}, \tag{947}$$

$$\tilde{P}_i \begin{bmatrix} 0 \\ \alpha_{i+1} \end{bmatrix} = \frac{1}{\rho_i} \begin{bmatrix} \beta_{i+1}\alpha_{i+1} \\ \tilde{\rho}_i \alpha_{i+1} \end{bmatrix} = \begin{bmatrix} \sigma_{i+1} \\ \tilde{\rho}_{i+1} \end{bmatrix}. \tag{948}$$

Ortogonální matice P_i, $1 \le i \le k$, budeme hledat ve tvaru $P_1 = \tilde{P}_1$ a

$$P_i = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix},$$

kde I je jednotková matice řádu $i-2$. Pak podle (946) a (947) pro $1 < i \le k$ platí

$$P_i B_i = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \quad P_i(\delta_1 e_1) = \begin{bmatrix} I, & 0 \\ 0, & \tilde{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1}, & 0 \\ 0, & 1 \end{bmatrix} (\delta_1 e_1) = \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix}.$$

Použijeme-li vzorce (947) a (948), dostaneme následující tvrzení.

Lemma 97. *Prvky matic R_i a vektorů h_i, $1 \le i \le k$, lze počítat podle rekurentních vztahů $\tilde{\rho}_1 = \alpha_1$, $\tilde{\eta}_1 = \beta_1$ a*

$$\begin{aligned} \rho_i = \sqrt{\tilde{\rho}_i^2 + \beta_{i+1}^2}, \qquad & \lambda_i = \frac{\tilde{\rho}_i}{\rho_i}, \qquad \mu_i = \frac{\beta_{i+1}}{\rho_i}, \\ \tilde{\rho}_{i+1} = \lambda_i \alpha_{i+1}, \qquad & \sigma_{i+1} = \mu_i \alpha_{i+1}, \\ \eta_i = \lambda_i \tilde{\eta}_i, \qquad & \tilde{\eta}_{i+1} = -\mu_i \tilde{\eta}_i. \end{aligned}$$

Nyní odvodíme rekurentní vztahy pro vektory s_{i+1}, $1 \le i \le k$. Jelikož

$$P_i(B_i z + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix}$$

a $P_i^T P_i = I$, můžeme položit $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg\min_{z \in R^i} \left\| \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \tilde{\eta}_{i+1} \end{bmatrix} \right\|.$$

Řešení této úlohy je řešením soustavy rovnic $R_i z_i + h_i = 0$ (věta 167). Vzhledem k jednoduché struktuře matic R_i, $1 \le i \le k$, můžeme vektory z_i, $1 \le i \le k$, a tudíž i vektory s_{i+1}, $1 \le i \le k$, určovat rekurentně.

Lemma 98. *Vektory $s_{i+1} = Q_i z_i$, $1 \le i \le k$, kde $R_i z_i + h_i = 0$, lze určit pomocí rekurentních vztahů $s_1 = 0$, $p_1 = q_1$ a*

$$\begin{aligned} s_{i+1} &= s_i - \frac{\eta_i}{\rho_i} p_i, \\ p_{i+1} &= q_{i+1} - \frac{\sigma_{i+1}}{\rho_i} p_i \end{aligned}$$

pro $1 \le i \le k$.

Důkaz Platí $R_1 = [\rho_1]$, $R_1^{-1} = [1/\rho_1]$ a

$$R_i = \begin{bmatrix} R_{i-1}, & r_{i-1} \\ 0, & \rho_i \end{bmatrix}, \quad R_i^{-1} = \begin{bmatrix} R_{i-1}^{-1}, & -R_{i-1}^{-1} r_{i-1}/\rho_i \\ 0, & 1/\rho_i \end{bmatrix}$$

pro $1 < i \le k$, kde $r_{i-1} = \sigma_i e_{i-1}$ a e_{i-1} je poslední sloupec jednotkové matice řádu $i-1$ (vztah pro R_i^{-1} můžeme ověřit dosazením do rovnosti $R_i R_i^{-1} = I$). Položíme-li $z_i = -R_i^{-1} h_i$, dostaneme z předchozích rovností rekurentní vztahy $r_1 = [1/\rho_1]$, $z_1 = -[\eta_1/\rho_1]$ a

$$r_i = \begin{bmatrix} -r_{i-1}/\rho_i \\ 1/\rho_i \end{bmatrix}, \quad z_i = \begin{bmatrix} z_{i-1} + (\eta_i/\rho_i)\sigma_i r_{i-1} \\ -\eta_i/\rho_i \end{bmatrix}.$$

pro $1 < i \le k$, takže

$$\begin{aligned} p_i &\stackrel{\Delta}{=} \rho_i Q_i r_i = q_i - \frac{\sigma_i}{\rho_{i-1}}\rho_{i-1}Q_{i-1}r_{i-1} = q_i - \frac{\sigma_i}{\rho_{i-1}}p_{i-1}, \\ s_{i+1} &= Q_i z_i = Q_{i-1}z_{i-1} + \frac{\eta_i}{\rho_i}\sigma_i Q_{i-1}r_{i-1} - \frac{\eta_i}{\rho_i}q_i \\ &= s_i - \frac{\eta_i}{\rho_i}(q_i - \frac{\sigma_i}{\rho_{i-1}}\rho_{i-1}Q_{i-1}r_{i-1}) = s_i - \frac{\eta_i}{\rho_i}p_i. \end{aligned}$$

□

Rekurentní vztahy uvedené v předchozích dvou lemmatech tvoří základ metody LSQR.

Definice 77. *Nechť $J \in R^{m\times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \quad \beta_1 u_1 = f, \quad \alpha_1 q_1 = J^T u_1, \quad \tilde\eta_1 = \beta_1, \quad \tilde\rho_1 = \alpha_1, \quad p_1 = q_1$$

a

$$\begin{gathered} \beta_{i+1}u_{i+1} = Jq_i - \alpha_i u_i, \qquad \alpha_{i+1}q_{i+1} = J^T u_{i+1} - \beta_{i+1}q_i, \\ \rho_i = \sqrt{\tilde\rho_i^2 + \beta_{i+1}^2}, \quad \lambda_i = \frac{\tilde\rho_i}{\rho_i}, \quad \tau_i = \frac{\beta_{i+1}}{\rho_i}, \quad \eta_i = \lambda_i\tilde\eta_i, \\ \tilde\rho_{i+1} = \lambda_i\alpha_{i+1}, \qquad \sigma_{i+1} = \tau_i\alpha_{i+1}, \qquad \tilde\eta_{i+1} = -\tau_i\tilde\eta_i, \\ s_{i+1} = s_i - \frac{\eta_i}{\rho_i}p_i, \qquad p_{i+1} = q_{i+1} - \frac{\sigma_{i+1}}{\rho_i}p_i \end{gathered}$$

pro $1 \le i \le n$, kde koeficienty α_i, β_i, $1 \le i \le n$, se volí tak, aby vektory $u_i \in R^m$, $q_i \in R^n$, $1 \le i \le n$, měly jednotkovou normu, nazveme metodou LSQR určenou maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$.

Metodu LSQR můžeme použít k realizaci nepřesné metody s lokálně omezeným krokem úplně stejně jako metodu CGNE (nebo CG), neboť podle poznámky 345 generují obě metody stejné vektory s_{i+1}, $1 \le i \le k$, kde $k \le n$ a $J^T J s_{k+1} + J^T f = 0$. Ukážeme ještě, jak je možné odhadovat přesnost řešení.

Věta 231. *Nechť $s_{i+1} \in R_n$, α_{i+1}, β_{i+1}, $\rho_i > 0$, η_i, $1 \le i \le k$, jsou veličiny generované metodou LSQR. Pak pro $1 \le i \le k$ platí*

$$\|J^T(Js_{i+1} + f)\| = \alpha_{i+1}\beta_{i+1}\frac{|\eta_i|}{\rho_i}.$$

Důkaz Nechť $\alpha_{i+1} \neq 0$, $\beta_{i+1} \neq 0$. Pak použitím vztahů (944)–(945) a poznámky 344 dostaneme

$$\begin{aligned} J^T(Js_{i+1} + f) &= J^T(JQ_i z_i + f) = J^T U_{i+1}(B_i z_i + \beta_1 e_1) = \\ &= (Q_i B_i^T + \alpha_{i+1}q_{i+1}e_{i+1}^T)(B_i z_i + \beta_1 e_1) = \alpha_{i+1}q_{i+1}e_{i+1}^T B_i z_i = \\ &= \alpha_{i+1}\beta_{i+1}q_{i+1}e_i^T z_i, \end{aligned}$$

neboť $B_i^T(B_i z_i + \beta_1 e_1) = 0$ podle definice vektoru z_i, $e_{i+1}^T e_1 = 0$ a $e_{i+1}^T B_i = \beta_{i+1}e_i^T$. Ale $Q_i^T Q_i = I$ a tudíž $Q_i^T s_{i+1} = Q_i^T Q_i z_i = z_i$, takže $e_i^T z_i = e_i^T Q_i^T s_{i+1} = q_i^T s_{i+1}$, což spolu s $\|q_{i+1}\| = 1$ dává

$$\|J^T(Js_{i+1} + f)\| = \alpha_{i+1}\beta_{i+1}|q_i^T s_{i+1}|.$$

Ale

$$q_i^T s_{i+1} = q_i^T s_i + \frac{\eta_i}{\rho_i}q_i^T p_i = q_i^T Q_{i-1}z_{i-1} - \frac{\eta_i}{\rho_i}q_i^T q_i + \frac{\eta_i\sigma_i}{\rho_i\rho_{i-1}}q_i^T p_{i-1} = -\frac{\eta_i}{\rho_i},$$

neboť $q_i^T Q_{i-1} = 0$, $q_i^T q_i = 1$ a vektor p_{i-1} je lineární kombinací sloupců matice Q_{i-1}, tudíž $q_i^T p_{i-1} = 0$. Jestliže $\alpha_{i+1} = 0$, $\beta_{i+1} = 0$, platí $\|J^T(Js_{i+1} + f)\| = 0$ (poznámka 344). □

Větu 231 můžeme využít k zastavení iteračního procesu (není třeba počítat reziduum $\|J^T(Js_{i+1} + f)\|$).

10.8 Numerické porovnání

Nejprve porovnáme účinnost metod pro rozsáhlé řídké a separovatelné úlohy pomocí 71 testovacích úloh s 1000 proměnnými ze sbírky TEST25. Jsou to tytéž úlohy, které byly použity v oddílu 9.8 (tabulka 9), implementované buď jako řídké (počítá se gradient $g(x)$) nebo jako separovatelné (počítají se redukované gradienty $\hat{g}_k(\hat{x})$, $1 \le k \le m$). V tabulce 10 jsou uvedeny výsledky získané těmito metodami:

`TRNMS-xx` - Diferenční verze Newtonovy metody pro řídké úlohy s aproximací Hessovy matice účelové funkce podle vzorce (887),

`TRVMS-xx` - metoda s proměnnou metrikou pro řídké úlohy s Marwilovou projekcí (895),

`TRNMP-xx` - Diferenční verze Newtonovy metody pro separovatelné úlohy s aproximací Hessových matic dílčích funkcí podle vzorce (918),

`LSVMP-xx` - kombinovaná metoda s proměnnou metrikou pro separovatelné úlohy (metoda VMP popsaná v oddílu 10.5),

Metody, jejichž označení začíná písmeny TR jsou realizovány jako metody s lokálně omezeným krokem a písmena LS označují metody spádových směrů. Přitom číslice xx udávají číslo použitého algoritmu. Pro srovnání jsou též uvedeny výsledky získané pomocí metod `LMVM-18`, `LMVM-21` a `LSTND-T` testovaných v oddílu 9.8.

Řídké úlohy					
Metoda	NIT	NFV	NFG	NCG	čas
TRNMS-10	6379	6615	32333	–	5.13
TRNMS-11	8029	8775	45153	–	5.53
TRNMS-12	8988	9589	51106	58202	6.30
TRVMS-11	58836	69103	51106	–	34.75
LSTND-T	7614	11997	128785	99928	9.56
LMVM-18	107007	109673	109673	–	14.46
LMVM-21	98065	101277	101277	–	13.13
Separovatelné úlohy					
TRNMS-10	6138	6403	31994	–	8.29
TRNMS-11	7861	8575	46422	–	10.03
TRNMS-12	8510	9105	50985	53574	9.93
TRNMP-10	6303	6577	21485	–	14.53
TRNMP-11	7834	8509	28848	–	15.97
TRNMP-12	8777	9409	33599	54152	17.51
LSVMP-3	13136	20685	20685	–	10.99
LSVMP-6	12782	17206	17206	237518	15.88
LMVM-18	13136	20685	206851	–	10.99
LMVM-21	12782	172060	172060	53574	15.88

Tabulka 10: TEST25 – 71 úloh

Tabulka 10 obsahuje celkový počet iterací `NIT`, celkový počet použitých funkčních hodnot `NFV`, celkový počet použitých gradientů `NFG`, celkový počet iterací metody sdružených gradientů při použití algoritmu 12 `NCG` a celkový čas výpočtu. Pro lepší srovnání robustnějších metod bylo použito 76 úloh ze sbírky TEST25. Výsledky jsou uvedeny v tabulce 11, která obsahuje navíc počet úloh (z celkového počtu 82), které daná metoda vyřešila.

Metoda	NIT	NFV	NFG	NCG	čas	počet
TRNMS-10	7249	7538	38450	–	7.86	81
TRNMS-12	10550	11255	64531	676171	30.93	77
TRNMP-10	6803	7133	23930	–	16.89	81
TRNMP-12	10019	10775	40477	722522	44.64	77
LSVMP-3	14189	22106	22106	–	12.43	81
LSVMP-6	15089	20863	20863	1392612	70.60	78

Tabulka 11: TEST25 – 76 úloh

Z výsledků uvedených v těchto tabulkách lze vyvodit několik závěrů:

- Diferenční verze Newtonovy metody jsou velmi efektivní pro řešení úloh s řídkými Hessovými maticemi, pokud nedochází k přílišnému nárůstu nenulových prvků při provádění Choleského rozkladu.
- Metody s proměnnou metrikou využívající řídkou strukturu Hessovy matice jsou neúčinné. V tabulce 10 jsou uvedeny pouze výsledky pro metodu používající Marwilovy projekce (895). Výsledky získané ostatními metodami byly ještě horší.
- Pro separovatelné úlohy je výhodnější určit nejprve řídkou strukturu Hessovy matice a pak použít deferenční vztah (887). Je to obvykle efektivnější než počítat redukované Hessovy matice podle vzorce (918).
- Metoda s proměnnou metrikou pro separovatelné úlohy `LSVMP-3` je velmi robustní i účinná a může konkurovat diferenčním verzím Newtonovy metody.
- Z tabulky 11 je patrné, že pro špatně podmíněné úlohy není účelné používat iterační algoritmy pro výpočet lokálně omezeného kroku. Tyto algoritmy jsou vhodné pro dobře podmíněné úlohy, zejména tehdy, dochází-li k velkému nárůstu nenulových prvků při provádění Choleského rozkladu.

K testování metod pro rozsáhlé úlohy nejmenších čtverců byly použity úlohy ze sbírky TEST26 zmíněné v oddílu 1.5. Nejprve porovnáme účinnost metod popsaných v oddílu 10.6 pomocí 56 úloh s 1000 proměnnými (4 úlohy ze sbírky TEST26 byly vynechány, protože je některá z testovaných metod nevyřešila). V tabulce 12 jsou uvedeny výsledky získané těmito metodami:

`TRGN-xx` - Gaussova-Newtonova metoda,
`TRGNS-xx` - modifikovaná Gaussova-Newtonova metoda GNS podle (936),
`TRGNP-xx` - modifikovaná Gaussova-Newtonova metoda GNP podle (937),
`TRGNN-xx` - modifikovaná Gaussova-Newtonova metoda GNN podle (938),
`TRGNB-xx` - modifikovaná Gaussova-Newtonova metoda GNB podle (939),
`TRGNC-xx` - modifikovaná Gaussova-Newtonova metoda GNC podle (940),
`TRGNJ-xx` - Gaussova-Newtonova metoda, která používá Jacobiovu matici J místo matice normální soustavy rovnic B.
`TRGNR-xx` - modifikovaná Gaussova-Newtonova metoda GNR podle (943),
`TRGNV-xx` - modifikovaná Gaussova-Newtonova metoda GNV podle (941),

Metody, jejichž označení začíná písmeny TR jsou realizovány jako metody s lokálně omezeným krokem a písmena LS označují metody spádových směrů. Přitom číslice xx udávají číslo použitého algoritmu (u metod používajících Jacobiovu matici značí CG metodu CGNE a LS metodu LSQR). Pro srovnání jsou též uvedeny výsledky získané pomocí metod `TRNMS-xx`, `LSVMP-xx` použitých v tabulce 10 a metod `LMVM-18`, `LMVM-21` testovaných v oddílu 9.8.

Metoda	NIT	NFV	NFG	NCG	čas	počet
TRGN-10	5472	5842	5527	–	6.53	59
TRGN-11	9645	10057	9699	–	6.08	60
TRGN-12	7255	7782	7308	1180653	64.33	56
TRGNS-10	4561	4761	4617	–	5.72	60
TRGNS-11	6939	7148	6995	–	4.21	60
TRGNS-12	9018	9248	9073	981738	65.95	57
TRGNP-10	4597	4869	4651	–	6.09	60
TRGNP-11	32612	40311	32668	–	21.85	54
TRGNP-12	7407	7982	7463	1175320	65.00	57
TRGNN-10	4775	4936	5217	–	5.75	60
TRGNN-11	6344	6496	6810	–	3.88	60
TRGNN-12	6774	7258	7313	1087690	61.23	57
TRGNB-10	4648	4827	4704	–	6.17	60
TRGNB-11	9922	10296	9977	–	7.20	60
TRGNB-12	6685	7108	6740	1173049	65.06	57
TRGNC-10	5914	6322	5970	–	7.33	60
TRGNC-11	8213	8977	8269	–	6.69	60
TRGNC-12	7468	8012	7524	1187550	65.97	57
TRGNJ-CG	13923	14464	13977	1232187	58.68	56
TRGNJ-LS	7470	8008	7523	1226757	79.77	56
TRGNR-CG	9931	10488	9985	1224959	57.73	56
TRGNR-LS	7089	7629	7143	1225386	79.15	56
TRGNV-CG	13640	14163	13727	1120004	54.88	56
TRMNS-10	5668	38705	38567	–	21.21	58
TRMNS-11	9396	87105	86245	–	24.46	58
TRMNS-12	9267	78893	78465	1097322	71.52	57
LSVMP-3	15097	11480	11480	–	22.15	59
LSVMP-6	27359	34260	34260	10141395	534.05	57
LMVMP-18	205821	233012	233012	–	56.90	55
LMVMP-21	200173	215871	215871	–	49.98	56

Tabulka 12: TEST26 – 56 úloh

Tabulka 12 obsahuje celkový počet iterací NIT, celkový počet použitých funkčních hodnot NFV, celkový počet použitých gradientů NFG, celkový počet iterací metody sdružených gradientů při použití algoritmu 12 NCG, celkový čas výpočtu a počet úloh (z celkového počtu 60), které daná metoda vyřešila. Pro lepší srovnání robustnějších metod bylo použito všech 60 úloh ze sbírky TEST26. Výsledky jsou uvedeny v tabulce 13, která obsahuje navíc počet selhání. K selhání došlo, když nestačilo 4000 iterací nebo 5000 vyčíslení součtu čtverců pro vyřešení dané úlohy.

Metoda	NIT	NFV	NFG	selhání	čas
TRGN-10	13678	14085	13737	1	18.53
TRGN-11	9780	10200	9838	–	6.31
TRGNS-10	6358	6730	6418	–	9.15
TRGNS-11	7226	7456	7286	–	4.41
TRGNN-10	5515	5819	7281	–	7.55
TRGNN-11	7403	7785	8234	–	4.88
TRMNS-10	6492	43809	43643	2	24.16
TRMNS-11	11329	99518	98491	2	29.82
LSVMP-3	16217	30393	30393	1	23.64
LMVMP-18	324609	353012	353012	5	97.46
LMVMP-21	319616	335871	335871	4	91.23

Tabulka 13: TEST26 – 60 úloh

Z výsledků uvedených v těchto tabulkách lze vyvodit několik závěrů:

- Specializované metody pro minimalizaci součtu čtverců (Gaussova-Newtonova metoda a její modifikace) jsou robustnější a efektivnější než metody určené k minimalizaci obecné účelové funkce. Diferenční verze Newtonovy metody, která se velmi osvědčila pro řešení úloh ze sbírky TEST25 (tabulka 11) při řešení úloh ze sbírky TEST26 dvakrát selhala a spotřebovala mnohem víc strojového času (tabulka 13).
- Modifikace aproximující člen druhého řádu v Hessově matici (zejména GNS a GNN) velmi zvyšují efektivitu Gaussovy-Newtonovy metody.
- Některé z testovacích úloh mají velmi špatně podmíněné Jacobiovy matice. Proto je použití iteračních metod (algoritmus 12, CGNE, LSQR) celkově méně výhodné než použití řídkého Choleského rozkladu.
- Kvazinewtonovské aktualizace Jacobiovy matice (metody GNR a GNV) jen mírně zlepšují efektivitu Gaussovy-Newtonovy metody. Použití Jacobiovy matice (metoda CGNE) je často výhodnější než použití matice normální soustavy rovnic (algoritmus 12). Metoda LSQR je sice pomalejší, ale dává přesnější řešení (spotřebuje se menší počet iterací a hodnot účelové funkce).
- Pro řešení úloh ze sbírky TEST26 není vhodné používat metody s proměnnou metrikou s omezenou pamětí `LMVMP-xx`. Celkem dobře si vede metoda s proměnnou metrikou pro separovatelné úlohy `LSVMP-3`.

11 Metody pro řešení soustav nelineárních rovnic

11.1 Základní vlastnosti metod pro řešení soustav nelineárních rovnic

Nechť $f : \mathcal{D}_F \to R^n$ je zobrazení definované na množině $\mathcal{D}_F \subset R^n$ (používáme stejné značení jako v oddílu 8). Naším úkolem bude nalézt bod $x^* \in R^n$ takový, že $f(x^*) = 0$. K řešení této úlohy bylo vyvinuto mnoho metod založených na různých přístupech. Zde se omezíme pouze na metody příbuzné optimalizačním metodám, které jsou obvykle jednoduché a účinné. Pomineme například homotopické a simpliciální metody a metody založené na řešení soustav diferenciálních rovnic. Většinou budeme předpokládat, že zobrazení $f : \mathcal{D}_F \to R^n$ je spojitě diferencovatelné na nějaké otevřené množině $\mathcal{D}$, $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$. V tomto případě budeme psát $f \in \mathcal{C}^1$ nebo $f \in \mathcal{C}^1 : \mathcal{D} \to R^n$. Příbuznost metod pro řešení soustav nelineárních rovnic s optimalizačními metodami plyne z toho, že:

(1) Optimalizační metody můžeme chápat jako metody pro řešení soustavy rovnic $g(x) = 0$, kde $g : \mathcal{D} \to R^n$ je gradient minimalizované funkce $F : \mathcal{D} \to R$. V tomto případě jde o speciální soustavu rovnic, neboť Jacobiova matice zobrazení g je Hessovou maticí funkce F a je tedy symetrická (neboť z $g \in \mathcal{C}^1$ na $\mathcal{D}$, plyne $F \in \mathcal{C}^2$ na $\mathcal{D}$). Řešením soustavy rovnic $g(x) = 0$ však můžeme získat nejen lokální minimum, ale i sedlový bod nebo dokonce lokální maximum funkce F.

(2) Řešení soustavy rovnic $f(x) = 0$ můžeme převést na minimalizaci funkce

$$F(x) = \frac{1}{2}\|f(x)\|^2 = \frac{1}{2}\sum_{k=1}^{n} f_k^2(x) \tag{949}$$

(součet čtverců). V tomto případě však můžeme získat lokální minimum funkce $F(x)$, které není řešením soustavy rovnic $f(x) = 0$.

Vztah mezi lokálními extrémy funkce $F(x) = (1/2)\|f(x)\|^2$ a řešením soustavy rovnic $f(x) = 0$ udává tato věta.

Věta 232. *Nechť $f \in \mathcal{C}^1 : \mathcal{D} \to R^n$ a nechť bod $x^* \in \mathcal{D}$ je lokálním minimem funkce $F(x) = (1/2)\|f(x)\|^2$, přičemž Jacobiova matice $J(x^*)$ zobrazení f v bodě x^* je regulární. Pak platí $f(x^*) = 0$.*

Důkaz Gradient funkce $F(x) = (1/2)\|f(x)\|^2$ v bodě $x^* \in \mathcal{D}$ lze vyjádřit ve tvaru

$$g(x^*) = J^T(x^*)f(x^*)$$

(vzorec (650)). Jelikož matice $J(x^*)$ je regulární, můžeme psát

$$f(x^*) = (J^T(x^*))^{-1}g(x^*),$$

takže $f(x^*) = 0$ právě tehdy, když $g(x^*) = 0$, což je nutná podmínka pro lokální extrém funkce $F(x)$. □

Podobně jako jsme v kapitole 1 definovali základní optimalizační metodu, můžeme definovat základní metodu pro řešení soustav nelineárních rovnic.

Definice 78. *základní metoda pro řešení soustav nelineárních rovnic je iterační proces, jehož výsledkem je posloupnost $x_i \in R^n$, $i \in N$, taková, že*

$$x_{i+1} = x_i + \alpha_i s_i,$$

kde směrový vektor $s_i \in R_n$ se určuje na základě hodnot x_j, f_j, J_j, $1 \le j \le i$, a délka kroku $\alpha_i > 0$ se určuje na základě chování funkce $F(x) = (1/2)\|f(x)\|^2$ v okolí bodu $x_i \in R^n$.

Definice 79. *Řekneme, že základní metoda pro řešení soustav nelineárních rovnic je globálně konvergentní, jestliže pro libovolný počáteční vektor $x_1 \in R^n$ platí*

$$\lim_{i\to\infty} \|f(x_i)\| = 0.$$

Mezi nejjednodušší a nejznámější metody pro řešení soustav nelineárních rovnic patří Newtonova metoda. Tato metoda je definována vztahy

$$\begin{aligned} s_i &= -J^{-1}(x_i)f(x_i), \\ \alpha_i &= 1 \end{aligned}$$

(předpokládáme, že matice $J(x_i)$, $i \in N$, jsou regulární). Z tohoto vyjádření je zřejmé, že směrový vektor Newtonovy metody pro řešení soustav nelineárních rovnic shodný se směrovým vektorem Gaussovy-Newtonovy metody pro minimalizaci součtu čtverců $F(x) = (1/2)\|f(x)\|^2$, neboť platí

$$(J^T(x_i)J(x_i))^{-1}J^T(x_i) = J^{-1}(x_i).$$

Matice $B_i = J^T(x_i)J(x_i)$ je v tomto případě pozitivně definitní, takže Newtonovu metodu pro řešení soustav nelineárních rovnic můžeme realizovat jako metodu spádových směrů (na rozdíl od Newtonovy metody pro nepodmíněnou minimalizaci popsané v oddílu 5.3, kterou je třeba realizovat jako metodu s lokálně omezeným krokem). Výhoda Newtonovy metody pro řešení soustav nelineárních rovnic spočívá v tom, že řešíme soustavu lineárních rovnic $J(x_i)s_i + f(x_i) = 0$, kde matice soustavy ná číslo podmíněnosti $\kappa(J(x_i))$, zatímco v případě Gaussovy-Newtonovy metody řešíme soustavu rovnic $J^T(x_i)J(x_i)s_i + g(x_i) = 0$ (kde $g(x_i) = J^T(x_i)f_(x_i)$), s maticí, jejíž číslo podmíněnosti je $\kappa(J^T(x_i)J(x_i)) = \kappa(J(x_i))^2$. Je-li matice $J(x_i)$ špatně podmíněná, projevují se u Gaussovy-Newtonovy metody výrazněji zaokrouhlovací chyby, které mohou snížit rychlost konvergence.

Při vyšetřování konvergence metod pro řešení soustav nelineárních rovnic budeme používat předpoklady J1–J6 uvedené v oddílu 8. Na rozdíl od optimalizačních metod popsaných v kapitolách 2 a 5, kde se v důkazech globální konvergence nepoužívá předpoklad F5, budeme nyní (v případě Newtonovy metody) potřebovat nějakou analogii předpokladu J5. Předpoklad J5 je velmi silný, jak ukazuje tato věta.

Věta 233. *Nechť je splněn předpoklad J5, kde množina $\mathcal{D}$ je konvexní. Pak soustava nelineárních rovnic $f(x) = 0$ má na $\mathcal{D}$ nanejvýš jedno řešení.*

Důkaz Nechť $x^* \in \mathcal{D}$ a $f(x^*) = 0$. Nechť je splněn předpoklad J5. Pak podle (665) pro $x \in \mathcal{D}$, $x \neq x^*$, dostaneme $\|f(x)\| = \|f(x) - f(x^*)\| \geq \underline{J}\|x - x^*\| > 0$. □

Poznámka 347. Protože je obtížné zajisitit platnost předpokladů J4–J5 na příliš velké otevřené množině $\mathcal{D}$, budeme v důkazech globální konvergence Newtonovy metody popužívat slabší předpoklady

Předpoklad J4a. *Existuje konstanta $\overline{J} > 0$ taková, že*

$$\|J(x_i)s\| \leq \overline{J}\|s\| \qquad \forall i \in N \qquad \forall s \in R^n. \tag{950}$$

Předpoklad J5a. *Existuje konstanta $\underline{J} > 0$ taková, že*

$$\|J(x_i)s\| \geq \underline{J}\|s\| \qquad \forall i \in N \qquad \forall s \in R^n. \tag{951}$$

Platnost nerovností (950) a (951) předpokládáme pouze v bodech x_i, $i \in N$, generovaných iterační metodou. Je zřejmé, že z J4–J5 plyne J4a–J5a.

V této kapitole se budeme většinou zabývat metodami, které místo Jacobiových matic $J_i = J(x_i)$, $i \in N$, používají jejich aproximace A_i, $i \in N$, splňující tyto předpoklady.

Předpoklad A3a. *Existuje číslo* $\overline{\vartheta} > 0$ *takové, že*

$$\|A_i - J_i\| \leq \overline{\vartheta} \qquad \forall i \in N. \tag{952}$$

Předpoklad A4a. *Existuje číslo* $\overline{A} > 0$ *takové, že*

$$\|A_i s\| \leq \overline{A} s \qquad \forall i \in N \qquad \forall s \in R^n. \tag{953}$$

Předpoklad A5a. *Existuje číslo* $\underline{A} > 0$ *takové, že*

$$\|A_i s\| \geq \underline{A} s \qquad \forall i \in N \qquad \forall s \in R^n. \tag{954}$$

Podmínka (953) je ekvivalentní podmínce $\|A_i\| \leq \overline{A}$ a podmínka (954) je ekvivalentní podmínce $\|A_i^{-1}\| \leq 1/\underline{A}$.

Poznámka 348. Poznamenejme, že z (950) a (952) plyne

$$\|A_i s\| \leq \|J_i s\| + \|(A_i - J_i)s\| \leq (\overline{J} + \overline{\vartheta})\|s\|,$$

takže platí (953) s $\overline{A} = \overline{J} + \overline{\vartheta}$. Jestliže $\overline{\vartheta} < \underline{J}$, pak z (951) a (952) plyne

$$\|A_i s\| \geq \|J_i s\| - \|(A_i - J_i)s\| \geq (\underline{J} - \overline{\vartheta})\|s\|,$$

takže platí (954) s $\underline{A} = \underline{J} - \overline{\vartheta}$. Jsou-li splněny předpoklady J4a, J5a a A3a, budeme předpokládat, že $\overline{\vartheta} < \underline{J}$ a že čísla $\overline{A}$, $\underline{A}$ použitá v (953), (954) jsou určena vztahy $\overline{A} = \overline{J} + \overline{\vartheta}$, $\underline{A} = \underline{J} - \overline{\vartheta}$.

Podmínka $\overline{\vartheta} < \underline{J}$ je sice postačující pro to, aby platila nerovnost (954), ale v důkazech globální konvergence metod pro řešení soustav nelineárních rovnic budeme potřebovat silnější podmínku.

$$\overline{\vartheta} < \frac{1 - \overline{\omega}}{2} \underline{J}, \tag{955}$$

kde číslo $\overline{\omega}$ udává přesnost výpočtu směrového vektoru (definice 80 a definice 81). Pokud $\underline{A} = \underline{J} - \overline{\vartheta}$, můžeme podmínku (955) nahradit nerovností

$$\overline{\vartheta} < \frac{1 - \overline{\omega}}{1 + \overline{\omega}} \underline{A}, \quad \text{neboli} \quad \overline{\vartheta} = \lambda \frac{1 - \overline{\omega}}{1 + \overline{\omega}} \underline{A}, \quad 0 \leq \lambda < 1, \tag{956}$$

která je vhodnější pro vyšetřování globální konvergence.

Předpoklady A3a s (956) a A4a–A5a mohou být nahraženy slabšími předpoklady, ve kterých A_i je regulární matice, f_i je hodnota zobrazení f v bodě x_i a s_i je použitý směrový vektor.

Předpoklad A3b. *Existuje číslo* $\overline{\vartheta} > 0$ *takové, že*

$$\|(A_i - J_i)^T f_i\| \leq \overline{\vartheta}\|f_i\| \qquad \forall i \in N. \tag{957}$$

Předpoklad A4b. *Existuje číslo* $\overline{A} > 0$ *takové, že*

$$\|A_i s_i\| \leq \overline{A}\|s_i\| \qquad \forall i \in N. \tag{958}$$

Předpoklad A5b. *Existuje číslo* $\underline{A} > 0$ *takové, že*

$$\|A_i s_i\| \geq \underline{A}\|s_i\| \qquad \forall i \in N. \tag{959}$$

Poznamenejme, že z předpokladů J4a, J5a a A3b s (956) neplyne platnost předpokladů A4b–A5b (nerovnosti (958)–(959) musí být splněny nezávisle na (956)–(957)). Význam předpokladů A3b s (956) a A4b–A5b spočívá v tom, že některé metody splňují předpoklad A3b s (956) automaticky (sdružená kvazinewtonovská metoda uvedená v poznámce 368 splňuje (957) s $\overline{\vartheta} = 0$) a platnost předpokladů A4b–A5b lze zajistit algoritmicky (přerušováním iteračního procesu). Předpoklady A3b s (956) a A4b–A5b stačí k tomu, abychom dokázali globální konvergenci metod spádových směrů i metod s lokálně omezeným krokem.

11.2 Metody spádových směrů

Při vyšetřování metod spádových směrů pro řešení soustav nelineárních rovnic budeme používat označení $h_i = A_i^T f_i$ pro aproximaci gradientu $g_i = J_i^T f_i$. Poznamenejme, že podmínka ($\overline{\text{S1}}$), použitá v definici 80, implikuje nerovnost

$$h_i^T s_i = f_i^T A_i s_i = f_i^T (A_i s_i + f_i) - f_i^T f_i \leq \overline{\omega}\|f_i\|^2 - \|f_i\|^2 = -(1-\overline{\omega})\|f_i\|^2 < 0. \tag{960}$$

Definice 80. *Řekneme, že základní metoda pro řešení soustav nelineárních rovnic* $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, *je metodou spádových směrů, jestliže směrové vektory* $s_i \in R^n$, $i \in N$, *se určují tak, že*

$$\|A_i s_i + f_i\| \leq \overline{\omega}_i \|f_i\|, \tag{$\overline{\text{S1}}$}$$

kde $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$, *a délky kroku* $\alpha_i > 0$, $i \in N$, *se vybírají tak, že* α_i *je první člen posloupnosti* α_i^j, $j \in N$ *(kde* $\alpha_i^1 = 1$ *a* $\underline{\beta}\alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta}\alpha_i^j$ $\forall j \in N$*) takový, že buď*

$$F_{i+1} - F_i \leq \underline{\rho}\alpha_i h_i^T s_i, \tag{$\overline{\text{S2a}}$}$$

nebo

$$F_{i+1} - F_i \leq -2\underline{\rho}(1-\overline{\omega})\alpha_i F_i = -\underline{\rho}(1-\overline{\omega})\alpha_i \|f_i\|^2, \tag{$\overline{\text{S2b}}$}$$

nebo

$$\|f_{i+1}\| - \|f_i\| \leq -\underline{\rho}(1-\overline{\omega})\alpha_i \|f_i\|, \tag{$\overline{\text{S2c}}$}$$

kde $0 < \underline{\beta} \leq \overline{\beta} < 1$ *a* $0 < \underline{\rho} < 1 - \lambda$ *(*λ *je číslo vystupující v (956)).*

Lemma 99. *(Konzistence) Nechť zobrazení* $f : \mathcal{D} \to R^n$ *vyhovuje předpokladům J1, J4, J6. Nechť matice* A_i, $i \in N$, *jsou regulární a splňují předpoklady A3b s (956) a A5b. Pak lze v každém iteračním kroku metody spádových směrů (definice 80) nalézt směrový vektor* $s_i \in R^n$ *vyhovující podmínce (*$\overline{\text{S1}}$*) a délku kroku* $\alpha_i > 0$ *vyhovující libovolné z podmínek (*$\overline{\text{S2}}$*), kde* $0 < \underline{\rho} < 1 - \lambda$. *Pokud* $\overline{\vartheta} = 0$, *platí* $\lambda = 0$ *a nepotřebujeme, aby byl splněn předpoklad A5b.*

Důkaz Existence směrového vektoru $s_i \in R^n$ vyhovujícího podmínce ($\overline{\text{S1}}$) plyne bezprostředně z regularity matice A_i (vektor s_i můžeme zvolit tak, že $\|A_i s_i + f_i\| = 0$). Z podmínky ($\overline{\text{S1}}$), z A3b a z definice vektorů $g_i = J_i^T f_i$, $h_i = A_i^T f_i$ lze jednoduše odvodit nerovnosti

$$(1-\overline{\omega})\|f_i\| \leq \|A_i s_i\| \leq (1+\overline{\omega})\|f_i\|, \tag{961}$$

$$(1-\overline{\omega})\|f_i\|^2 \leq -h_i^T s_i \leq (1+\overline{\omega})\|f_i\|^2, \tag{962}$$

$$|h_i^T s_i - g_i^T s_i| \leq \overline{\vartheta}\|f_i\|\|s_i\|. \tag{963}$$

Nerovnost (961) spolu s předpokladem A5b dává

$$\|s_i\| \leq \frac{1+\overline{\omega}}{\underline{A}}\|f_i\|, \tag{964}$$

Použijeme-li nerovnosti (962)–(964), dostaneme

$$\begin{aligned} -g_i^T s_i &\geq -h_i^T s_i - \overline{\vartheta}\|f_i\|\|s_i\| \geq -h_i^T s_i - \overline{\vartheta}\frac{1+\overline{\omega}}{\underline{A}}\|f_i\|^2 \\ &\geq -h_i^T s_i + \frac{\overline{\vartheta}}{\underline{A}}\frac{1+\overline{\omega}}{1-\overline{\omega}}h_i^T s_i = -(1-\lambda)h_i^T s_i \geq (1-\lambda)(1-\overline{\omega})\|f_i\|^2 > 0, \end{aligned} \tag{965}$$

takže podle lemmatu 4 existuje pro libovolné číslo $0 < \varepsilon_1 < 1$ délka kroku $\alpha_i > 0$ určená Armijovým výběrem (poznámka 23) taková, že

$$F_{i+1} - F_i \leq \varepsilon_1\alpha_i g_i^T s_i \leq \varepsilon_1\alpha_i(1-\lambda)h_i^T s_i \leq -\varepsilon_1\alpha_i(1-\overline{\omega})(1-\lambda)\|f_i\|^2$$

(předpoklady lemmatu 4 jsou splněny, neboť podle věty 157 z předpokladů J1, J4, J6 plyne, že funkce F vyhovuje podmínce F3). Položme $\underline{\rho} = \varepsilon_1(1-\lambda)$, takže $0 < \underline{\rho} < 1-\lambda \leq 1$. Pak

$$F_{i+1} - F_i \leq \underline{\rho}\alpha_i h_i^T s_i \leq -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|^2,$$

takže podmínky ($\overline{\text{S2a}}$) a ($\overline{\text{S2b}}$) jsou konzistentní, pokud $0 < \underline{\rho} < 1-\lambda$. Podmínka ($\overline{\text{S2c}}$) je také konzistentní, neboť z

$$2\|f_i\|(\|f_{i+1}\| - \|f_i\|) \leq (\|f_{i+1}\| + \|f_i\|)(\|f_{i+1}\| - \|f_i\|) = 2(F_{i+1} - F_i)$$

a z ($\overline{\text{S2b}}$) plyne, že

$$\|f_{i+1}\| - \|f_i\| \leq \frac{F_{i+1} - F_i}{\|f_i\|} \leq -2\underline{\rho}(1-\overline{\omega})\alpha_i\frac{F_i}{\|f_i\|} = -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|.$$

Pokud $\overline{\vartheta} = 0$, platí $h_i = g_i$, takže podmínka $-g_i^T s_i > 0$ plyne bezprostředně z (960) a není třeba používat nerovnost (965) (tedy ani předpoklad A5b). □

Poznámka 349. Poznamenejme, že kromě konzistence podmínek ($\overline{\text{S2a}}$)–($\overline{\text{S2c}}$) jsme dokázali implikace ($\overline{\text{S2a}}$) ⇒ ($\overline{\text{S2b}}$) ⇒ ($\overline{\text{S2c}}$) (pro stejnou hodnotu parametru $\underline{\rho}$). Ukážeme, že platí také opačné implikace ($\overline{\text{S2c}}$) ⇒ ($\overline{\text{S2b}}$) ⇒ ($\overline{\text{S2a}}$), kde každá následující podmínka je splněna s poněkud menší (ale nenulovou) hodnotou parametru $\underline{\rho}$. Z nerovnosti

$$\frac{F_{i+1} - F_i}{F_i} = \frac{(\|f_{i+1}\| + \|f_i\|)(\|f_{i+1}\| - \|f_i\|)}{\|f_i\|^2} \leq \frac{\|f_{i+1}\| - \|f_i\|}{\|f_i\|}$$

plyne, že platí-li ($\overline{\text{S2c}}$) pro nějakou hodnotu parametru $\underline{\rho}$, je splněna i podmínka ($\overline{\text{S2b}}$) s poloviční hodnotou tohoto parametru. Použijeme-li nerovnost (962), dostaneme

$$-2\underline{\rho}(1-\overline{\omega})\alpha_i F_i = -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|^2 \leq \underline{\rho}\frac{1-\overline{\omega}}{1+\overline{\omega}}\alpha_i h_i^T s_i,$$

takže platí-li ($\overline{\text{S2b}}$) pro nějakou hodnotu parametru $\underline{\rho}$, je splněna i podmínka ($\overline{\text{S2a}}$) s hodnotou tohoto parametru vynásobenou číslem $(1-\overline{\omega})/(1+\overline{\omega})$.

Lemma 100. *Nechť jsou splněny předpoklady lemmatu 99. Pak existuje konstanta $0 < \underline{\alpha} \leq 1$ taková, že délky kroku určené metodou spádových směrů (definice 80) splňují podmínku $0 < \underline{\alpha} \leq \alpha_i \leq 1$, $i \in N$.*

Důkaz Podle poznámky 349 se můžeme omezit na metody spádových směrů používající podmínku ($\overline{\text{S2b}}$). Při výběru délky kroku podle ($\overline{\text{S2b}}$) platí buď $\alpha_i = \alpha_i^1 = 1$ nebo $\alpha_i = \alpha_i^k = \beta\alpha_i^{k-1}$, kde $0 < \underline{\beta} \leq \beta \leq \overline{\beta} < 1$ a $F(x_i + \alpha_i^{k-1}s_i) - F(x_i) \geq -\underline{\rho}(1-\overline{\omega})\alpha_i^{k-1}\|f_i\|^2$. Pokud $\alpha_i < 1$, můžeme psát

$$F(x_i + \frac{\alpha_i}{\beta}s_i) - F(x_i) \geq -\underline{\rho}(1-\overline{\omega})\frac{\alpha_i}{\beta}\|f_i\|^2.$$

Z druhé strany, použijeme-li tvrzení 1 o střední hodnotě (pokládáme $d_i = \mu(\alpha_i/\beta)s_i$, kde $0 \leq \mu \leq 1$) a předpoklady J1, J4, J6, můžeme psát

$$\begin{aligned} F(x_i + \frac{\alpha_i}{\beta}s_i) - F(x_i) &= \frac{\alpha_i}{\beta}g^T(x_i + d_i)s_i \\ &\leq \frac{\alpha_i}{\beta}\left(g_i^T s_i + \|g(x_i + d_i) - g(x_i)\|\|s_i\|\right) \\ &\leq \frac{\alpha_i}{\beta}\left(g_i^T s_i + \frac{\alpha_i}{\beta}(\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2\right), \end{aligned}$$

neboť podle (667) platí

$$\|g(x_i + d_i) - g(x_i)\| \leq \left(\overline{J}^2 + \overline{G}\,\overline{f}\right)\|d_i\| \leq \frac{\alpha_i}{\beta}\left(\overline{J}^2 + \overline{G}\,\overline{f}\right)\|s_i\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$-\underline{\rho}(1-\overline{\omega})\|f_i\|^2 \leq g_i^T s_i + \frac{\alpha_i}{\beta}(\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2$$

a použijeme-li (964) a (965), můžeme psát

$$\frac{\alpha_i}{\beta}\left(\overline{J}^2 + \overline{G}\,\overline{f}\right)\frac{(1+\overline{\omega})^2}{\underline{A}^2}\|f_i\|^2 \geq -\underline{\rho}(1-\overline{\omega})\|f_i\|^2 - g_i^T s_i \geq (1-\overline{\omega})(1-\underline{\rho}-\lambda)\|f_i\|^2$$

což spolu s $\beta \geq \underline{\beta}$ dává

$$\alpha_i \geq \frac{\underline{\beta}(1-\overline{\omega})(1-\underline{\rho}-\lambda)\underline{A}^2}{(\overline{J}^2 + \overline{G}\,\overline{f})(1+\overline{\omega})^2}.$$

Položíme-li

$$\underline{\alpha} = \min\left(1, \frac{\underline{\beta}(1-\overline{\omega})(1-\underline{\rho}-\lambda)\underline{A}^2}{(\overline{J}^2 + \overline{G}\,\overline{f})(1+\overline{\omega})^2}\right), \tag{966}$$

platí $0 < \underline{\alpha} \leq \alpha_i \leq 1\ \forall i \in N$. □

Poznámka 350. Předpoklad A5b potřebujeme pouze k tomu, abychom mohli použít nerovnost (964). Proto můžeme tento předpoklad nahradit předpokladem, že $\|s_i\| \leq \overline{c}\|f_i\|$, $i \in N$, kde $\overline{c} > 0$ je konstanta, která nezávisí na indexu $i \in N$.

Lemma 101. *Uvažujme základní metodu pro řešení soustav nelineárních rovnic (definice 78) takovou, že $\|s_i\| \leq \overline{c}\|f_i\|$, $i \in N$, a délky kroku splňují některou z podmínek ($\overline{\text{S2}}$) s $0 < \underline{\alpha} \leq \alpha_i \leq 1$, $i \in N$. Pak $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Důkaz provedeme pro ($\overline{\text{S2c}}$), neboť tato podmínka vyplývá z podmínek ($\overline{\text{S2a}}$) a ($\overline{\text{S2b}}$) (poznámka 349). Podle ($\overline{\text{S2c}}$) platí

$$\|f_{i+1}\| \leq (1 - \underline{\rho}(1-\overline{\omega})\alpha_i)\|f_i\| \leq (1 - \underline{\rho}(1-\overline{\omega})\underline{\alpha})\|f_i\| \stackrel{\Delta}{=} q\|f_i\|,$$

kde $0 < q < 1$, neboť $0 < \underline{\rho} < 1$, $0 < 1 - \overline{\omega} \leq 1$ a $0 < \underline{\alpha} \leq 1$. Porovnáním s geometrickou řadou dostaneme

$$\sum_{i=1}^{\infty} \|f_i\| \le \frac{1}{1-q}\|f_1\| < \infty,$$

což implikuje $\|f_i\| \to 0$. Použijeme-li nerovnosti $\|s_i\| \le \overline{c}\|f_i\|$, $i \in N$, můžeme psát

$$\sum_{i=1}^{\infty} \|s_i\| \le \overline{c} \sum_{i=1}^{\infty} \|f_i\| < \infty,$$

takže posloupnost x_i, $i \in N$, splňuje Bolzanovu-Cauchyovu podmínku. Proto $x_i \to x^\star$, což dohromady s $f_i \to 0$ dává $f(x^\star) = 0$. □

Věta 234. *(globální konvergence). Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům J1, J4, J6. Nechť matice A_i, $i \in N$, jsou regulární a splňují předpoklady A3b s (956) a A5b. Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů (definice 80). Pak $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 100 a lemmatu 101. Předpoklady lemmatu 101 jsou splněny, neboť podle (964) pro $i \in N$ platí $\|s_i\| \le \overline{c}\|f_i\|$, kde $\overline{c} = (1+\overline{\omega})/\underline{A}$. □

Poznámka 351. Z odhadu $F_{i+1} \le qF_i$, $i \in N$, kde $0 < q < 1$, plyne, že $x_i \to x^\star$ alespoň R-lineárně.

Nyní se budeme zabývat superlineární konvergencí metod spádových směrů. Budeme přitom používat podmínku ($\overline{\text{S2c}}$) s $0 < \underline{\rho} < 1$.

Věta 235. *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná metodou spádových směrů taková, že $x_i \to x^*$, kde $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínce ($\overline{S2c}$). Nechť platí*

$$\lim_{i\to\infty} \frac{\|A_i s_i + f_i\|}{\|f_i\|} = 0 \tag{967}$$

a

$$\lim_{i\to\infty} \frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0. \tag{968}$$

Pak existuje index $k \in N$ takový, že $\alpha_i = 1$, pokud $i \ge k$ a posloupnost x_i, $i \in N$, konverguje superlineárně k bodu $x^* \in R^n$.

Důkaz Důkaz povedeme poněkud obecněji, neboť získané výsledky použijeme v důkazu vět 267 a 270. To znamená, že v částech (a)–(b) budeme místo (967) předpokládat pouze platnost podmínky ($\overline{\text{S1}}$), neboli

$$\limsup_{i\to\infty} \frac{\|A_i s_i + f_i\|}{\|f_i\|} \le \overline{\omega} < 1.$$

Platí-li (967), můžeme ve všech vzorcích položit $\overline{\omega} = 0$.

(a) Nechť $0 < \underline{J} < \|J^{-1}(x^*)\|^{-1} \le \|J(x^*)\| < \overline{J}$. Ukážeme, že existuje index $k_2 \in N$ takový, že

$$\frac{1-\overline{\omega}}{\overline{J}}\|f_i\| \le \|s_i\| \le \frac{1+\overline{\omega}}{\underline{J}}\|f_i\|,$$

pokud $i \ge k_2$. Označme $\omega_i = (A_i s_i + f_i)/\|f_i\|$ a $\vartheta_i = (A_i - J_i)s_i/\|s_i\|$. Pak platí

$$J_i s_i = (A_i s_i + f_i) - (A_i - J_i)s_i - f_i = \omega_i\|f_i\| - \vartheta_i\|s_i\| - f_i,$$

takže

$$\|s_i\| \ge \frac{1 - \|\omega_i\|}{\|J_i\| + \|\vartheta_i\|}\|f_i\|$$

a jelikož $\|\omega_i\| \le \overline{\omega}$ a $\|\vartheta_i\| \to 0$ (podle ($\overline{\text{S1}}$) a (968)) a $\|J_i\| \to \|J^*\| < \overline{J}$, existuje index $k_1 \in N$ takový , že $\|s_i\| \ge \|f_i\|(1-\overline{\omega})/\overline{J}$, pokud $i \ge k_1$. Podobně platí

$$s_i = J_i^{-1}(\omega_i\|f_i\| - \vartheta_i\|s_i\| - f_i),$$

takže

$$\|s_i\| \le \frac{\|J_i^{-1}\|(1+\|\omega_i\|)}{1-\|J_i^{-1}\|\|\vartheta_i\|}\|f_i\|$$

a jelikož $\|\omega_i\| \le \overline{\omega}$ a $\|\vartheta_i\| \to 0$ (podle ($\overline{\text{S1}}$) a (968)) a $\|J_i^{-1}\| \to \|(J^*)^{-1}\| < 1/\underline{J}$, existuje index $k_2 \ge k_1$ takový, že $\|s_i\| \le \|f_i\|(1+\overline{\omega})/\underline{J}$, pokud $i \ge k_2$.

(b) Ukážeme, že existuje index $k \ge k_2$ takový, že hodnota $\alpha_i = 1$ vyhovuje podmínce ($\overline{\text{S2c}}$) s $0 < \underline{\rho} < 1$, pokud $i \ge k$. Použijeme-li dva členy Taylorova rozvoje, dostaneme

$$f(x_i + s_i) = f_i + J_i s_i + o(\|s_i\|) = (A_i s_i + f_i) - (A_i - J_i)s_i + o(\|s_i\|)$$

neboli

$$\frac{\|f(x_i+s_i)\|}{\|f_i\|} \le \|\omega_i\| + \|\vartheta_i\|(1+\overline{\omega})/\underline{J} + o(\|f_i\|)/\|f_i\|, \tag{969}$$

takže $\limsup_{i\to\infty}(\|f(x_i+s_i)\| - \|f_i\|)/\|f_i\| \le -(1-\overline{\omega})$ (podle ($\overline{\text{S1}}$) a (968)), a jelikož $0 < \underline{\rho} < 1$, existuje index $k \ge k_2$ takový, že podmínka ($\overline{\text{S2c}}$) s $\alpha_i = 1$ je splněna, pokud $i \ge k$.

(c) Předpokládejme nyní že platí (967)–(968). Pomocí věty 6 o střední hodnotě dostaneme

$$\frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \le \frac{\overline{J}}{\underline{J}}\frac{\|f_{i+1}\|}{\|f_i\|},$$

takže podle (967)–(968) a (969) platí

$$\lim_{i\to\infty}\frac{\|x_{i+1}-x^*\|}{\|x_i-x^*\|} = \lim_{i\to\infty}\frac{\overline{J}}{\underline{J}}(\|\omega_i\| + \|\vartheta_i\|(1+\overline{\omega})/\underline{J} + o(\|f_i\|)/\|f_i\|) = 0$$

a $x^* \to x$ Q-superlineárně. □

Poznámka 352. Věta 235 zůstane v platnosti, i tehdy používáme-li k výběru délky kroku podmínku ($\overline{\text{S2b}}$) (nebo ($\overline{\text{S2a}}$)). Abychom mohli používat kroky jednotkové délky, což se předpokládá v důkazu superlineární konvergence, musíme v tomto případě snížit hodnotu parametru $\underline{\rho}$ podle poznámky 349. Například z ($\overline{\text{S2b}}$) plyne $2\underline{\rho}\alpha_i \le 1 - F_{i+1}/F_i < 1$ (předpokládáme, že $F_i > 0$ $\forall i \in N$), takže $\alpha_i = 1$ lze volit pouze tehdy, když $2\underline{\rho} < 1$.

Poznámka 353. Položíme-li $A_i = J(x_i)$, $i \in N$, dostaneme Newtonovu metodu. V tomto případě předpoklady J4a–J5a implikují A4a–A5a a předpoklad A3a platí s $\overline{\vartheta} = 0$, takže lze položit $\lambda = 0$ ve všech vzorcích uvedených v předchozím textu. Z těchto úvah plyne, že Newtonova metoda realizovaná jako metoda spádových směrů je globálně konvergentní (jsou-li splněny předpoklady J1, J4, J5a a J6).

Následující věta ukazuje, jak lze vlastnosti libovolné metody spádových směrů odvodit z vlastností Newtonovy metody.

Věta 236. *Nechť matice $J(x_i)$, $i \in N$, splňují předpoklad J5a a matice A_i, $i \in N$, splňují předpoklad A3a s (955). Nechť s_i je směrový vektor vyhovující podmínce ($\overline{\text{S1}}$). Pak platí*

$$\|J_i s_i + f_i\| \le \tilde{\omega}\|f_i\|,$$

kde $\tilde{\omega} = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$. Jinými slovy, platí-li ($\overline{\text{S1}}$) a A3a s (955), můžeme vektor s_i považovat za směrový vektor získaný Newtonovou metodou, kde příslušná soustava lineárních rovnic je řešena s přesností $\tilde{\omega} = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$.

Důkaz Použijeme-li (961) a A3a, dostaneme

$$(1+\overline{\omega})\|f_i\| \geq \|A_i s_i\| \geq \|J_i s_i\| - \|(A_i - J_i)s_i\| \geq (\underline{J} - \overline{\vartheta})\|s_i\|,$$

neboli

$$\|s_i\| \leq \frac{1+\overline{\omega}}{\underline{J} - \overline{\vartheta}}\|f_i\|.$$

Můžeme tedy psát

$$\|J_i s_i + f_i\| \leq \|A_i s_i + f_i\| + \|(J_i - A_i)s_i\| \leq \overline{\omega}\|f_i\| + \overline{\vartheta}\|s_i\| \leq \frac{J\overline{\omega} + \overline{\vartheta}}{\underline{J} - \overline{\vartheta}}\|f_i\| \triangleq \tilde{\omega}\|f_i\|.$$

Přitom $\tilde{\omega} = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$, pokud $\overline{\vartheta} < (1-\overline{\omega})\underline{J}/2$. □

Teoretické výsledky shrnuté v lemmatu 99 a větě 236 vyžadují splnění předpokladu A3a (s vhodnou hodnotou $\overline{\vartheta} > 0$, která může vycházet velmi malá). Tento předpoklad má teoretický význam, ale v praxi ho není možno ověřit (používáme-li matici A, neznáme obvykle matici J, neboť v opačném případě by bylo vhodné použít Newtonovu metodu, která je superlineárné konvergentní). Proto je třeba globální konvergenci zajistit jiným způsobem (jde o to aby byla splněna některá z podmínek ($\overline{\text{S2}}$)). V případě, že neplatí ($\overline{\text{S2}}$)) pro α_i větší než zadaná dolní mez, přeruší se iterační proces, což znamená, že se spočte matice J a použije se krok Newtonovy metody. Tyto úvahy jsou shrnuty ve formě algoritmu.

Algoritmus 27. Data $0 \leq \overline{\omega} < 1$, $0 < \underline{\rho} < 1$, $0 < \underline{\beta} \leq \overline{\beta} < 1$, $\overline{\varepsilon} > 0$, $\overline{c} > 0$, $0 < \underline{k} \leq \overline{k}$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$, vypočteme $f_1 = f(x_1)$ a položíme $i = 1$ a $l = 1$.

Krok 2 Pokud $\|f_i\| \leq \overline{\varepsilon}$, ukončíme výpočet.

Krok 3 Pokud $l = 1$, vypočteme Jacobiovu matici $J_i = J(x_i)$ a položíme $A_i = J_i$ (restart). Zvolíme přesnost $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$ a vypočteme směrový vektor $s_i \in R^n$ vyhovující podmínce ($\overline{\text{S1}}$).

Krok 4 Pokud $l > 1$ a $\|s_i\| > \overline{c}\,\|f_i\|$, položíme $l = 1$ a přejdeme na krok 3.

Krok 5a Položíme $\alpha_i^1 = 1$ a $k = 1$.

Krok 5b Položíme $x_{i+1} = x_i + \alpha_i^k s_i$ a vypočteme $f_i = f(x_i)$. Je-li splněna některá (vybraná) podmínka z ($\overline{\text{S2}}$), přejdeme na krok 6.

Krok 5c Pokud $l = 1$ a $k > \overline{k}$, ukončíme výpočet (předčasné ukončení způsobené selháním Newtonovy metody). Pokud $l > 1$ a $k > \underline{k}$, položíme $l = 1$ a přejdeme na krok 3. V ostatních případech určíme délku kroku α_i^{k+1} tak aby platilo $\underline{\beta}\alpha_i^k \leq \alpha_i^{k+1} \leq \overline{\beta}\alpha_i^k$, položíme $k := k+1$ a přejdene na krok 5b.

Krok 6 Určíme novou matici A_{i+1} (například pomocí vhodné kvazinewtonovské aktualizace popsané v oddílu 11.5), položíme $i := i+1$, $l := l+1$ a přejdeme na krok 2.

Věta 237. *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům J1, J4, J5a a J6. Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná algoritmem 27, kde $\overline{\varepsilon} = 0$ a číslo $\overline{k}$ je dostatečně velké. Pak buď existuje index $i \in N$ takový, že $f(x_i) = 0$, nebo platí $x_i \to x^*$, kde $f(x^*) = 0$.*

Důkaz Nechť číslo $\overline{k}$ je zvoleno tak, že $\overline{k} > \log\underline{\alpha}/\log\overline{\beta}$, kde $\underline{\alpha} > 0$ je číslo určerné vztahem (966), ve kterém $\underline{A} = \underline{J}$ a $\lambda = 0$. Pak nutně libovolná vybraná podmínka z ($\overline{\text{S2}}$) je splněna pro $k \leq \overline{k}$, takže algoritmus 27 nemůže skončit v kroku 5c. Může skončit v kroku 2, pokud $f(x_i) = 0$. V opačném případě podle lemmatu 100 a lemmatu 101 platí $x_i \to x^*$, kde $f(x^*) = 0$. □

11.3 Metody s lokálně omezeným krokem

Při výkladu metod s lokálně omezeným krokem budeme používat označení

$$L_i(s) = \|A_i s + f_i\| - \|f_i\|$$

pro lineární funkci, která lokálně aproximuje rozdíl $\|f(x_i + s)\| - \|f(x_i)\|$ a označení

$$\omega_i(s) = (A_i s + f_i)/\|f_i\|$$

pro přesnost určení směrového vektoru (předpokládáme, že $\|f_i\| \neq 0$, neboť v opačném případě je bod x_i řešením soustavy rovnic $f(x) = 0$). Dále budeme používat označení

$$\rho_i(s) = (\|f(x_i + s)\| - \|f_i(x_i)\|)/L_i(s)$$

pro podíl skutečného a předpověděného poklesu normy zobrazení $f : \mathcal{D}_F \to R^n$.

Definice 81. *Řekněme, že základní metoda pro řešení soustav nelineárních rovnic $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou s lokálně omezeným krokem, jestliže:*

(1) Směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že

$$\|s_i\| \leq \Delta_i, \tag{$\overline{\text{T1a}}$}$$

$$\|s_i\| < \Delta_i \quad \Rightarrow \quad \|A_i s_i + f_i\| \leq \overline{\omega}_i \|f_i\|, \tag{$\overline{\text{T1b}}$}$$

$$-L_i(s_i) \geq \underline{\sigma} \|A_i s_i\|, \tag{$\overline{\text{T1c}}$}$$

kde $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$ a $0 < \underline{\sigma} < 1$.

(2) Délky kroku $\alpha_i \geq 0$, $i \in N$, se určují tak, že

$$\rho_i(s_i) \leq 0 \quad \Rightarrow \quad \alpha_i = 0, \tag{$\overline{\text{T2a}}$}$$

$$\rho_i(s_i) > 0 \quad \Rightarrow \quad \alpha_i = 1. \tag{$\overline{\text{T2b}}$}$$

(3) Meze $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se určují tak, že

$$\rho_i(s_i) < \underline{\rho} \quad \Rightarrow \quad \underline{\beta}\|s_i\| \leq \Delta_{i+1} \leq \overline{\beta}\|s_i\|, \tag{$\overline{\text{T3a}}$}$$

$$\rho_i(s_i) \geq \underline{\rho} \quad \Rightarrow \quad \Delta_i \leq \Delta_{i+1} \leq \overline{\Delta}, \tag{$\overline{\text{T3b}}$}$$

kde $0 < \underline{\beta} < \overline{\beta} < 1$ a $0 < \underline{\rho} < 1/2$.

Poznámka 354. Při vyšetřování metod s lokálně omezeným krokem budeme používat označení

$$N_1 = \{i \in N : \|s_i\| < \Delta_i\},$$

$$N_2 = \{i \in N : \rho_i(s_i) \geq \underline{\rho}\}.$$

Jelikož $\underline{\rho} > 0$, platí $x_{i+1} = x_i + s_i$, pokud $i \in N_2$.

Lemma 102. *Nechť zobrazení $f : \mathcal{D}_F \to R^n$ vyhovuje předpokladům J1, J4, J6. Nechť matice A_i, $i \in N$, jsou regulární a splňují předpoklady A3b–A5b s (956). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem ($\overline{\text{T1}}$)–($\overline{\text{T3}}$) (s $0 < 2\underline{\rho} < 1 - \lambda$). Pak existuje konstanta $\underline{c} > 0$ taková, že*

$$\|s_i\| \geq \underline{c}\|f_i\| \quad \forall i \in N.$$

Důkaz (a) Nechť $i \in N_1$. Potom z ($\overline{\text{T1b}}$) plyne

$$\big|\|A_i s_i\| - \|f_i\|\big| \le \|A_i s_i + f_i\| \le \overline{\omega}\|f_i\|,$$

takže $(1-\overline{\omega})\|f_i\| \le \|A_i s_i\| \le \|A_i\|\|s_i\|$. Platí tedy

$$\|s_i\| \ge \frac{1-\overline{\omega}}{\overline{A}}\|f_i\|.$$

(b) Nechť $i \notin N_1$ a $i \notin N_2$. Z ($\overline{\text{T1c}}$) plyne, že $L_i(s_i) \le 0$, takže

$$\begin{aligned} L_i(s_i)\|f_i\| &= \left(\|A_i s_i + f_i\| - \|f_i\|\right)\|f_i\| \ge \left(\|A_i s_i + f_i\|^2 - \|f_i\|^2\right) \\ &= 2\left(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i\right) \triangleq 2Q_i(s_i). \end{aligned} \tag{970}$$

Jestliže $\|f(x_i+s_i)\| \le \|f(x_i)\|$, pak nerovnost $\rho_i(s_i) < \underline{\rho}$ spolu s (970) dává

$$\begin{aligned} F(x_i+s_i) - F(x_i) &= \frac{1}{2}\left(\|f(x_i+s_i)\|^2 - \|f(x_i)\|^2\right) \\ &\ge \left(\|f(x_i+s_i)\| - \|f(x_i)\|\right)\|f(x_i)\| \\ &\ge \underline{\rho} L_i(s_i)\|f_i\| \ge 2\underline{\rho} Q_i(s_i). \end{aligned}$$

Jestliže $\|f(x_i+s_i)\| \ge \|f(x_i)\|$, platí tato nerovnost triviálně. Můžeme tedy psát

$$F(x_i+s_i) - F(x_i) \ge 2\underline{\rho} Q_i(s_i).$$

Z druhé strany, použijeme-li tvrzení 1 o střední hodnotě (pokládáme $d_i = \mu s_i$, kde $0 \le \mu \le 1$), předpoklady J1, J4, J6 a (965), můžeme psát

$$\begin{aligned} F(x_i+s_i) - F(x_i) &\le g_i^T s_i + \|g(x_i+d_i) - g(x_i)\|\|s_i\| \\ &\le g_i^T s_i + (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2 \\ &\le (1-\lambda) h_i^T s_i + (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2 \\ &\le (1-\lambda) Q_i(s_i) + (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2, \end{aligned}$$

neboť $h_i^T s_i = f_i^T A_i s_i \le Q_i(s_i)$ a podle (667) platí

$$\|g(x_i+d_i) - g(x_i)\| \le (\overline{J}^2 + \overline{G}\,\overline{f})\|d_i\| \le (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$2\underline{\rho} Q_i(s_i) \le (1-\lambda) Q_i(s_i) + (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2,$$

neboli

$$-(1-\lambda-2\underline{\rho}) Q_i(s_i) \le (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2.$$

Podmínky ($\overline{\text{T1c}}$) a A5b spolu s nerovností (970) dávají

$$-Q_i(s_i) \ge -\frac{1}{2} L_i(s_i)\|f_i\| \ge \frac{\underline{\sigma}}{2}\|A_i s_i\|\|f_i\| \ge \frac{\underline{\sigma}}{2}\underline{A}\|s_i\|\|f_i\|.$$

Dosadíme-li tento vztah do předchozí nerovnosti, dostaneme

$$\frac{\underline{\sigma}\,\underline{A}}{2}(1-\lambda-2\underline{\rho})\|s_i\|\|f_i\| \le -(1-\lambda-2\underline{\rho}) Q_i(s_i) \le (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2,$$

neboli

$$\|s_i\| \geq \frac{\underline{\sigma}\,\underline{A}(1-\lambda-2\underline{\rho})}{2(\overline{J}^2+\overline{G}\,\overline{f})}\,\|f_i\|,$$

(c) Nechť $i=1$. Jestliže $\|f_1\|=0$, pak jistě $\|s_1\| \geq \underline{c}\|f_1\|$ pro libovolnou konstantu $\underline{c}>0$. Jestliže $\|f_1\| \neq 0$, dostaneme

$$\|s_1\| \geq \frac{\|s_1\|}{\|f_1\|}\|f_1\|.$$

(d) Nechť $i \notin N_1$, $i \in N_2$ a $i \neq 1$. Nechť $k<i$ je maximální index, pro který současně neplatí $k \notin N_1$, $k \in N_2$ a $k \neq 1$. Použijeme-li ($\overline{\text{T3a}}$)–($\overline{\text{T3b}}$) a ($\overline{\text{T1a}}$), můžeme psát

$$\|s_i\| = \Delta_i \geq \Delta_{k+1} \geq \min(\Delta_k, \underline{\beta}\|s_k\|) \geq \min\left(\|s_k\|, \underline{\beta}\|s_k\|\right) = \underline{\beta}\|s_k\|,$$

takže podle ($\overline{\text{T2a}}$)–($\overline{\text{T2b}}$) a (a)–(c) platí

$$\|s_i\| \geq \underline{\beta}\|s_k\| \geq \underline{c}\|f_k\| \geq \underline{c}\|f_i\|,$$

kde

$$\underline{c} = \underline{\beta}\min\left(\frac{1-\overline{\omega}}{\overline{A}}, \frac{\underline{\sigma}\,\underline{A}(1-\lambda-2\underline{\rho})}{2(\overline{J}^2+\overline{G}\,\overline{f})}, \frac{\|s_1\|}{\|f_1\|}\right).$$

□

Věta 238. *(globální konvergence). Nechť jsou splněny předpoklady lemmatu 102. Pak $x_i \to x^\star$, přičemž $f(x^\star)=0$.*

Důkaz (a) Nejprve ukážeme, že $f_i \to 0$. Předpokládejme, že toto tvrzení neplatí. Protože posloupnost $\|f_i\|$, $i \in N$, je podle ($\overline{\text{T2a}}$)–($\overline{\text{T2b}}$) nerostoucí, existuje číslo $\underline{\varepsilon}>0$ takové, že $\|f_i\| \geq \underline{\varepsilon}$, $\forall i \in N$ a podle lemmatu 102 platí

$$\|s_i\| \geq \underline{c}\,\underline{\varepsilon}, \quad \forall i \in N.$$

Předpokládejme nejprve, že množina N_2 je nekonečná. Protože $x_{i+1}=x_i+s_i$, pokud $i \in N_2$, můžeme psát

$$\begin{aligned}\|f_i\|-\|f_{i+1}\| &= \|f(x_i)\|-\|f(x_i+s_i)\| \geq -\underline{\rho}L_i(s_i)\\ &\geq \underline{\rho}\,\underline{\sigma}\|A_is_i\| \geq \underline{\rho}\,\underline{\sigma}\,\underline{A}\,\underline{c}\,\underline{\varepsilon}, \quad \forall i \in N_2.\end{aligned}$$

Odtud plyne

$$\begin{aligned}\|f_1\| &\geq \lim_{i\to\infty}(\|f_1\|-\|f_{i+1}\|) = \sum_{i=1}^{\infty}(\|f_i\|-\|f_{i+1}\|)\\ &\geq \sum_{i\in N_2}(\|f_i\|-\|f_{i+1}\|) \geq \sum_{i\in N_2}\underline{\rho}\,\underline{\sigma}\,\underline{A}\,\underline{c}\,\underline{\varepsilon} = \infty,\end{aligned}$$

což dává spor. Předpokládejme nyní, že množina N_2 je konečná. Potom ($\overline{\text{T3a}}$) implikuje $\Delta_i \to 0$, což dohromady s ($\overline{\text{T1a}}$) dává $\|s_i\| \to 0$. Ale to je ve sporu s nerovností $\|s_i\| \geq \underline{c}\underline{\varepsilon}$ $\forall i \in N$.

(b) Použitím ($\overline{\text{T1c}}$) dostaneme $L_i(s_i)=\|A_is_i+f_i\|-\|f_i\| \leq 0$, takže

$$\|f_i\| \geq \|A_is_i+f_i\| \geq \|A_is_i\|-\|f_i\|.$$

Tato nerovnost implikuje $\|A_is_i\| \leq 2\|f_i\|$, takže

$$\underline{A}\|s_i\| \leq \|A_is_i\| \leq 2\|f_i\|. \tag{971}$$

Nyní ukážeme, že $\sum_{i=1}^{\infty}\|s_i\| < \infty$. Je-li množina N_2 konečná, existuje index $l \notin N_2$ takový, že $i \notin N_2$ $\forall i \geq l$. Platí tedy

$$\sum_{i=1}^{\infty}\|s_i\| \leq \sum_{i=1}^{l-1}\|s_i\| + \|s_l\|\sum_{i=l}^{\infty}\overline{\beta}^{i-l} \leq (l-1)\overline{\Delta} + \|s_l\|/(1-\overline{\beta}) < \infty.$$

podle ($\overline{\text{T3a}}$). Je-li množina N_2 nekonečná, můžeme tak jako v (a) psát

$$\begin{aligned}\|f_1\| &\geq \sum_{i=1}^{\infty}(\|f_i\| - \|f_{i+1}\|) \geq \sum_{i\in N_2}(\|f_i\| - \|f_{i+1}\|)\\ &\geq \underline{\rho}\,\underline{\sigma}\sum_{i\in N_2}\|A_i s_i\| \geq \underline{\rho}\,\underline{\sigma}\,\underline{A}\sum_{i\in N_2}\|s_i\|.\end{aligned}$$

Označme $N_2 = \{l_1, l_2, l_3, \dots\}$. Použijeme-li (971) a lemma 102, dostaneme

$$\|s_{l_j+1}\| \leq \frac{2}{\underline{A}}\|f_{l_j+1}\| \leq \frac{2}{\underline{A}}\|f_{l_j}\| \leq \frac{2}{\underline{c}\,\underline{A}}\|s_{l_j}\|$$

a ($\overline{\text{T3a}}$) implikuje $\|s_{l_j+k}\| \leq \overline{\beta}\|s_{l_j+k-1}\|$ pro $2 \leq k \leq l_{j+1} - l_j - 1$. Platí tedy

$$\begin{aligned}\sum_{i=1}^{\infty}\|s_i\| &= \sum_{i=1}^{l_1-1}\|s_i\| + \sum_{j=1}^{\infty}\left[\|s_{l_j}\| + \sum_{k=1}^{l_{j+1}-l_j-1}\|s_{l_j+k}\|\right]\\ &\leq (l_1-1)\overline{\Delta} + \sum_{j=1}^{\infty}\|s_{l_j}\|\left[1 + \frac{2}{\underline{c}\,\underline{A}}\sum_{k=1}^{l_{j+1}-l_j-1}\overline{\beta}^{k-1}\right]\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{c}\,\underline{A}}\frac{1}{1-\overline{\beta}}\right]\sum_{i\in N_2}\|s_i\|\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{c}\,\underline{A}}\frac{1}{1-\overline{\beta}}\right]\frac{\|f_1\|}{\underline{\rho}\,\underline{\sigma}\,\underline{A}} < \infty.\end{aligned}$$

Z nerovnosti $\sum_{i=1}^{\infty}\|x_{i+1} - x_i\| \leq \sum_{i=1}^{\infty}\|s_i\| < \infty$ plyne, že posloupnost x_i, $i \in N$, splňuje Bolzanovu-Cauchyovu podmínku, takže $x_i \to x^\star$, což spolu s $f_i \to 0$ dává $f(x^\star) = 0$. □

Věta 239. *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem ($\overline{\text{T1}}$)–($\overline{\text{T3}}$) taková, že $x_i \to x^*$, kde $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Nechť*

$$\lim_{i\to\infty}\overline{\omega}_i = 0 \tag{972}$$

a

$$\lim_{i\to\infty}\frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0. \tag{973}$$

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz (a) Nechť $0 < \underline{J} < \|J^{-1}(x^*)\|^{-1} \leq \|J(x^*)\| < \overline{J}$. Ukážeme, že existuje index $k_1 \in N$ takový, že

$$-L_i(s_i) \geq \underline{\sigma}\,\underline{J}\|s_i\|$$

a

$$\|f_i\| \geq \frac{1}{2}\underline{J}\|s_i\|,$$

pokud $i \geq k_1$. Označme $\vartheta_i = (A_i - J_i)s_i/\|s_i\|$. Pak platí

$$\|A_i s_i\| = \|J_i s_i + \vartheta_i \|s_i\|\| \geq \|J_i s_i\| - \|\vartheta_i\|\|s_i\|$$

a jelikož $\|\vartheta_i\| \to 0$, $J_i \to J(\underline{x}^*)$ a $\underline{J} < \|J^{-1}(x^*)\|^{-1}$, existuje index $k_1 \in N$ takový, že $\|A_i s_i\| \geq \underline{J}\|s_i\|$, pokud $i \geq k_1$. Použijeme-li ($\overline{\text{T1c}}$), můžeme psát

$$-L_i(s_i) \geq \underline{\sigma}\|A_i s_i\| \geq \underline{\sigma}\,\underline{J}\|s_i\|.$$

Z definice $L_i(s_i)$ a z ($\overline{\text{T1c}}$) plyne

$$0 \geq L_i(s_i) = \|A_i s_i + f_i\| - \|f_i\|,$$

neboli

$$|\|A_i s_i\| - \|f_i\|| \leq \|A_i s_i + f_i\| \leq \|f_i\|,$$

takže $\|A_i s_i\| \leq 2\|f_i\|$, což spolu s nerovností $\|A_i s_i\| \geq \underline{J}\|s_i\|$ dává $\|f_i\| \geq (\underline{J}/2)\|s_i\|$, pokud $i \geq k_1$.

(b) Ukážeme, že existuje index $k_2 \geq k_1$ takový, že $i \in N_2$, pokud $i \geq k_2$. Použijeme-li dva členy Taylorova rozvoje, dostaneme

$$f(x_i + s_i) = f(x_i) + J_i s_i + o(\|s_i\|) = f(x_i) + A_i s_i - (A_i - J_i)\, s_i + o(\|s_i\|)$$

takže

$$\begin{aligned} \rho_i(s_i) &= \frac{\|f(x_i)\| - \|f(x_i + s_i)\|}{-L_i(s_i)} \geq \frac{-L_i(s_i) - \|\vartheta_i\|\|s_i\| + o(\|s_i\|)}{-L_i(s_i)} \geq \\ &\geq 1 - \frac{\|\vartheta_i\|\|s_i\| + o(\|s_i\|)}{\underline{\sigma}\,\underline{J}\|s_i\|} \to 1, \end{aligned}$$

neboť $\|\vartheta_i\| \to 0$. Jelikož $\underline{\rho} < 1$, existuje index $k_2 \geq k_1$ takový, že $\rho_i(s_i) \geq \underline{\rho}$, pokud $i \geq k_2$.

(c) Ukážeme, že existuje index $k \geq k_2$ takový, že $i \in N_1$, pokud $i \geq k$. Poznamenejme nejprve, že množina $N_1 \subset N$ je nekonečná. Kdyby tomu tak nebylo, muselo by platit $\|s_i\| \geq \Delta_i \geq \Delta_{k_2}$ $\forall i \geq k_2$, neboť z (b) plyne $i \in N_2$ $\forall i \geq k_2$. To je však spor, neboť podle (a) platí $\|s_i\| \leq 2\|f_i\|/\underline{J}$, takže $\|f_i\| \to 0$ implikuje $\|s_i\| \to 0$. Omezme se nyní pouze na indexy $i \geq k_2$, $i \in N_1$ a označme $\omega_i = (A_i s_i + f_i)/\|s_i\|$. Podle (972), (973) a ($\overline{\text{T1b}}$) platí $\|\omega_i\| \to 0$ a $\|\vartheta_i\| \to 0$, takže stejným způsobem jako v důkazu věty 235 (s $\overline{\omega} = 0$) se dá ukázat, že existuje index $k_3 \geq k_2$, $k_3 \in N_1$ takový, že

$$\|f_i\|/\overline{J} \leq \|s_i\| \leq \|f_i\|/\underline{J}$$

$\forall i \geq k_3$, $i \in N_1$. Použijeme-li dva členy Taylorova rozvoje, můžeme psát

$$f_{i+1} = f(x_i + s_i) = f_i + J_i s_i + o(\|s_i\|),$$

neboť $i \in N_2$. Označme

$$\lambda_i = \frac{f_{i+1} - f_i - A_i s_i}{\|f_i\|} = \frac{f_{i+1} - f_i - J_i s_i}{\|f_i\|} - \frac{(A_i - J_i)s_i}{\|f_i\|},$$

takže

$$\|\lambda_i\| = \|\vartheta_i\|\frac{\|s_i\|}{\|f_i\|} + o(1) \leq \frac{1}{\underline{J}}\|\vartheta_i\| + o(1).$$

Pak z $\|\vartheta_i\| \to 0$ plyne $\|\lambda_i\| \to 0$ a jelikož zároveň $\|\omega_i\| \to 0$, existuje index $k \geq k_3$, $k \in N_1$ takový, že $\|\lambda_i\| < (\underline{J}/\overline{J})/2$ a $\|\omega_i\| < (\underline{J}/\overline{J})/2$ $\forall i \geq k$, $i \in N_1$. Můžeme tedy psát

$$\begin{aligned}\|s_{i+1}\| &\leq \frac{1}{\underline{J}}\|f_{i+1}\| \leq \frac{1}{\underline{J}}\left(\|f_{i+1}-f_i-A_is_i\|+\|A_is_i+f_i\|\right) \leq \\ &\leq \frac{\overline{J}}{\underline{J}}\left(\|\lambda_i\|+\|\omega_i\|\right)\|s_i\| < \left(\frac{1}{2}+\frac{1}{2}\right)\|s_i\| = \|s_i\|.\end{aligned}$$

Jelikož $i \in N_2$ podle (b), platí $\Delta_{i+1} \geq \Delta_i$, což dává $\|s_{i+1}\| < \|s_i\| \leq \Delta_i \leq \Delta_{i+1}$, takže $i+1 \in N_1$. Indukcí dostaneme $i \in N_1$ $\forall i \geq k$.

(d) Superlineární konvergence. Platí

$$\frac{\|f_{i+1}\|}{\|f_i\|} \leq \frac{\|f_{i+1}-f_i-A_is_i\|+\|A_is_i+f_i\|}{\|f_i\|} \leq \|\lambda_i\|+\|\omega_i\|,$$

což spolu s $\|\lambda_i\| \to 0$ a $\|\omega_i\| \to 0$ dává

$$\lim_{i\to\infty}\frac{\|x_{i+1}-x^*\|}{\|x_i-x^*\|} \leq \frac{\overline{J}}{\underline{J}}\frac{\|f_{i+1}\|}{\|f_i\|} = 0.$$

□

11.4 Newtonova metoda

Newtonova metoda používá matice $A_i = J(x_i)$, $i \in N$, takže $\vartheta_i = (A_i - J_i)s_i/\|s_i\| = 0$, $i \in N$, a z předpokladů J4a–J5a plyne platnost předpokladů A4a–A5a.

Věta 240. *Nechť jsou splněny předpoklady J1, J4, J5a a J6. Pak Newtonova metoda realizovaná buď jako metoda spádových směrů nebo jako metoda s lokálně omezeným krokem je globálně konvergentní. Platí-li $x_i \to x^*$ a $\|\omega_i\| \to 0$, je rychlost konvergence Q-superlineární.*

Důkaz Globální konvergence plyne bezprostředně z věty 234 a věty 238. Superlineární konvergence plyne bezprostředně z věty 235 a věty 239, neboť $\vartheta_i = 0$, $i \in N$. □

Poznámka 355. Newtonova metoda pro řešení soustav nelineárních rovnic může být realizována jako globálně konvergentní metoda spádových směrů, což není možné v případě Newtonovy metody pro minimalizaci bez omezujících podmínek.

Nejsou-li Jacobiovy matice zadány analyticky, můžeme používat diferenční verze Newtonovy metody. V tom případě je však třeba odhadnout nepřesnosti, které vznikají při diferenční aproximaci Jacobiových matic.

Lemma 103. *Nechť je splněn předpoklad J6 a nechť*

$$Ae_j = \frac{f(x+\delta e_j)-f(x)}{\delta} \tag{974}$$

pro $1 \leq j \leq n$, kde e_j, $1 \leq j \leq n$, jsou sloupce jednotkové matice řádu n. Pak platí

$$\|A - J(x)\| \leq \frac{1}{2}\overline{G}\sqrt{n}\delta.$$

Důkaz Použijeme-li větu o střední hodnotě, dostaneme

$$f(x+\delta e_j) = f(x) + J(x)\delta e_j + \int_0^1 (J(x+\tau\delta e_j) - J(x))\delta e_j \mathrm{d}\tau,$$

takže

$$\begin{aligned}\|(A-J(x))e_j\| &= \left\|\frac{f(x+\delta e_j)-f(x)}{\delta}-J(x)e_j\right\| \le \frac{1}{\delta}\left\|\int_0^1 (J(x+\tau\delta e_j)-J(x))\delta e_j \mathrm{d}\tau\right\| \\ &\le \frac{1}{2\delta}\overline{G}\delta^2\|e_j\|^2 = \frac{1}{2}\overline{G}\delta.\end{aligned}$$

Nechť $s \in R^n$ je libovolný vektor s jednotkovou normou. Pak platí

$$\begin{aligned}\|(A-J(x))s\| &= \left\|\sum_{j=1}^n (A-J(x))e_j e_j^T s\right\| \le \sum_{j=1}^n |e_j^T s|\|(A-J(x))e_j\| \le \frac{1}{2}\overline{G}\delta\sum_{j=1}^n |e_j^T s| \\ &\le \frac{1}{2}\overline{G}\sqrt{n}\delta\|s\| = \frac{1}{2}\overline{G}\sqrt{n}\delta\end{aligned}$$

a jelikož

$$\|A-J(x)\| = \max_{\|s\|=1}\|(A-J(x))s\|,$$

dostaneme tvrzení lemmatu. □

Věta 241. *Nechť jsou splněny předpoklady J5a a J6. Je-li matice A určena podle vzorce (974), kde*

$$\delta < \frac{(1-\overline{\omega})\underline{J}}{\overline{G}\sqrt{n}}$$

a $0 \le \overline{\omega} < 1$, platí $\|A-J(x)\| \le \overline{\vartheta}$, kde $\overline{\vartheta} < (1/2)(1-\overline{\omega})\underline{J}$. Navíc $\|As+f\| \le \overline{\omega}$ implikuje $\|Js+f\| \le \tilde{\omega}$, kde $\tilde{\omega} = (\underline{J}\overline{\omega}+\overline{\vartheta})/(\underline{J}-\overline{\vartheta}) < 1$.

Důkaz Podle lemmatu 103 lze položit $\overline{\vartheta} = \overline{G}\sqrt{n}\delta/2$, takže nerovnost $\overline{\vartheta} < (1/2)(1-\overline{\omega})\underline{J}$ je splněna, platí-li $\delta \le (1-\overline{\omega})\underline{J}/(\overline{G}\sqrt{n})$. Zbytek tvrzení plyne z věty 236 □

Poznámka 356. Věta 241 ukazuje, že lze zvolit diferenci $\delta > 0$ tak, aby matice určená podle vztahu (974) splňovala podmínku pro globální konvergenci metody spádových směrů i metody s lokálně omezeným krokem. Je vidět, že diferenci δ je třeba zvolit tím menší, čím menší je číslo $\underline{J}$ použité v předpokladu J5a a čím větší je číslo $\overline{G}$ použité v předpokladu J6.

11.5 Kvazinewtonovské metody

Definice 82. *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 78) je kvazinewtonovskou metodou, jestliže*

$$A_i s_i + f_i = 0, \tag{975}$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i + u_i v_i^T, \tag{976}$$

kde $u_i \in R^n$, $v_i \in R^n$, a vyhovující podmínce

$$A_{i+1} d_i = y_i, \tag{977}$$

kde $y_i = f_{i+1} - f_i$, $d_i = x_{i+1} - x_i$.

Poznámka 357. V tomto oddílu se budeme zabývat pouze přesnými kvazinewtonovskými metodami (podmínka (975)), takže $(A_i s_i + f_i)/\|f_i\| = 0$, $i \in N$. Neplatí však $(A_i - J_i)s_i/\|s_i\| = 0$, $i \in N$ (matice A_i se mohou od matic J_i dosti lišit).

Věta 242. *Nechť $A_+ = A + uv^T$ a $Ad \neq y$. Pak $A_+d = y$ právě tehdy, když $v^Td \neq 0$ a $u = (y - Ad)/v^Td$, takže*

$$A_+ = A + \frac{(y - Ad)v^T}{v^Td}. \tag{978}$$

Jestliže $Ad = y$, platí $u = 0$, takže $A_+ = A$.

Důkaz Z podmínky $A_+d = y$ dostaneme $A_+d = Ad + uv^Td = y$. Jestliže $Ad = y$, stačí položit $u = v = 0$, takže $A_+ = A$. Jestliže $Ad \neq y$, musí platit $v^Td \neq 0$ a $u = (y - Ad)/v^Td$. □

Poznámka 358. Položíme-li $v = d$ dostaneme Broydenovu dobrou metodu

$$A_+ = A + \frac{(y - Ad)d^T}{d^Td}. \tag{979}$$

Položíme-li $v = A^Ty$, dostaneme Broydenovu špatnou metodu

$$A_+ = A + \frac{(y - Ad)y^TA}{y^TAd}. \tag{980}$$

Nechť

$$e_k^Td = \max_{1 \leq i \leq n} e_i^Td.$$

Položíme-li $v = e_k$, dostaneme přímou metodu aktualizace sloupců

$$A_+ = A + \frac{(y - Ad)e_k^T}{e_k^Td}, \tag{981}$$

která aktualizuje vždy pouze jeden sloupec matice A.

Věta 243. *Nechť A je regulární matice a nechť platí (978). Pak matice A_+ je regulární právě tehdy, když $v^TA^{-1}y \neq 0$.*

Důkaz Nechť $A_+ = A + uv^T$. Pak podle Shermanova-Morrisonova vzorce (poznámka 106) platí

$$A_+^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}, \tag{982}$$

takže A_+ je regulární právě tehdy, když $1 + v^TA^{-1}u \neq 0$. Dosadíme-li do této nerovnosti $u = (y - Ad)/v^Td$, dostaneme

$$1 + v^TA^{-1}u = 1 + \frac{v^TA^{-1}y - v^Td}{v^Td} = \frac{v^TA^{-1}y}{v^Td},$$

takže A_+ je regulární právě tehdy, když $v^TA^{-1}y \neq 0$. □

Poznámka 359. Věta 243 opodstatňuje použití Broydenovy špatné metody. Jestliže $y \neq 0$ a matice A je regulární, pak volba $v = A^Ty$ dává $v^TA^{-1}y = y^TAA^{-1}y = y^Ty = \|y\|^2 \neq 0$.

Zatím jsme uvedli tři možnosti, jak volit vektor v. K výběru vektoru v lze (podobně jako u metod s proměnnou metrikou v oddílu 4.4) použít minimalizační principy. Minimalizuje se obvykle číslo podmíněnosti $\kappa(M) = \|M\|\|M^{-1}\|$ nebo funkce $\|I - M\|\|I - M^{-1}\|$, kde

$$M = A^{-1}A_+ = I - \frac{(d - A^{-1}y)v^T}{v^Td} = I - \frac{(d - w)v^T}{v^Td} \tag{983}$$

$$M^{-1} = A_+^{-1}A = I + \frac{(d - A^{-1}y)v^T}{v^TA^{-1}y} = I + \frac{(d - w)v^T}{v^Tw} \tag{984}$$

a $w = A^{-1}y$. Vztah (984) plyne ze Shermanova-Morrisonova vzorce (poznámka 106). Jeho správnost lze též ověřit vynásobením matic M a M^{-1}.

Lemma 104. *Nechť κ je spektrální číslo podmíněnosti matice $M = I - uv^T$, $\delta = 1 - v^T u$ je její determinant a*

$$\sigma = \delta + \frac{\|u\|^2\|v\|^2}{2}, \qquad \tau = \sqrt{1 + 4\frac{1 - v^T u}{\|u\|^2\|v\|^2}}. \tag{985}$$

Pak platí

$$\kappa = \frac{\sigma + \sqrt{\sigma^2 - \delta^2}}{|\delta|} \tag{986}$$

a

$$\frac{\kappa - 1}{\kappa + 1} = \tau, \quad \delta \le 0,$$
$$\frac{\kappa - 1}{\kappa + 1} = \frac{1}{\tau}, \quad \delta > 0.$$

Důkaz Platí $M^T M = I - uv^T - vu^T + \|u\|^2 vv^T$, takže

$$\det(M^T M) = \delta^2 = (1 - v^T u)^2, \qquad \mathrm{Tr}(M^T M) = n - 2v^T u + \|u\|^2\|v\|^2.$$

Matice $M^T M$ má $n - 2$ jednotkových vlastních čísel a pro její zbylá dvě vlastní čísla $0 \le \lambda_1 \le \lambda_2$ platí

$$\lambda_1 + \lambda_2 = \mathrm{Tr}(M^T M) - (n - 2) = 2 - 2v^T u + \|u\|^2\|v\|^2 = 2\sigma, \qquad \lambda_1\lambda_2 = \delta^2.$$

Z $|\delta| = \sqrt{\lambda_1\lambda_2} \le (\lambda_1 + \lambda_2)/2 = \sigma$ plyne nerovnost $\sigma^2 - \delta^2 \ge 0$, takže lze psát

$$\lambda_1 = \sigma - \sqrt{\sigma^2 - \delta^2}, \qquad \lambda_2 = \sigma + \sqrt{\sigma^2 - \delta^2} \tag{987}$$

(neboť čísla λ_1, λ_2 jsou řešením kvadratické rovnice $\lambda^2 - 2\sigma\lambda + \delta = 0$). Ukážeme, že $0 \le \lambda_1 \le 1 \le \lambda_2$. Podle (987) jsou tyto nerovnosti splněny pokud $|\sigma - 1| \le \sqrt{\sigma^2 - \delta^2}$, neboli po umocnění $\delta^2 - 2\sigma + 1 \le 0$. To však platí, neboť použijeme-li vztahy $\delta = 1 - v^T u$, $2\sigma = 2 - 2v^T u + \|u\|^2\|v\|^2$ a Schwarzovu nerovnost $\|u\|^2\|v\|^2 - (v^T u)^2 \ge 0$, dostaneme

$$\delta^2 - 2\sigma + 1 = 1 - 2v^T u + (v^T u)^2 - 2 + 2v^T u - \|u\|^2\|v\|^2 + 1 = (v^T u)^2 - \|u\|^2\|v\|^2 \le 0.$$

Jelikož $0 \le \lambda_1 < 1 < \lambda_2$, můžeme psát

$$\kappa^2 = \frac{\lambda_2}{\lambda_1} = \frac{\sigma + \sqrt{\sigma^2 - \delta^2}}{\sigma - \sqrt{\sigma^2 - \delta^2}} = \frac{(\sigma + \sqrt{\sigma^2 - \delta^2})^2}{\delta^2} \quad \Rightarrow \quad \kappa = \frac{\sigma + \sqrt{\sigma^2 - \delta^2}}{|\delta|}.$$

Zřejmě

$$\frac{\kappa - 1}{\kappa + 1} = \frac{\sigma - |\delta| + \sqrt{\sigma^2 - \delta^2}}{\sigma + |\delta| + \sqrt{\sigma^2 - \delta^2}} = \frac{\sqrt{\sigma - |\delta|}(\sqrt{\sigma - |\delta|} + \sqrt{\sigma + |\delta|})}{\sqrt{\sigma + |\delta|}(\sqrt{\sigma + |\delta|} + \sqrt{\sigma - |\delta|})} = \sqrt{\frac{\sigma - |\delta|}{\sigma + |\delta|}}.$$

Pokud $\delta \le 0$, platí $\sigma + |\delta| = \|u\|^2\|v\|^2/2$, $\sigma - |\delta| = \sigma + \delta$, takže

$$\frac{\kappa - 1}{\kappa + 1} = \sqrt{\frac{\|u\|^2\|v\|^2 + 4\delta}{\|u\|^2\|v\|^2}} = \sqrt{1 + 4\frac{1 - v^T u}{\|u\|^2\|v\|^2}} = \tau.$$

Pokud $\delta > 0$, platí $\sigma - |\delta| = \|u\|^2\|v\|^2/2$, $\sigma + |\delta| = \sigma + \delta$, takže

$$\frac{\kappa - 1}{\kappa + 1} = \sqrt{\frac{\|u\|^2\|v\|^2}{\|u\|^2\|v\|^2 + 4\delta}} = \frac{1}{\tau}.$$

□

Lemma 105. *Nechť*

$$M_1 = I - \frac{(d-w)v_1^T}{v_1^T d}, \qquad M_2 = I - \frac{(d-w)(v_1+v_2)^T}{(v_1+v_2)^T d},$$

kde $v_1 \in \mathcal{L}(d,w)$, $v_2 \in \mathcal{L}(d,w)^\perp$. *Pak platí* $\kappa(M_1) \le \kappa(M_2)$.

Důkaz Označme $u_1 = (d-w)/v_1^T d$ a $u_2 = (d-w)/(v_1+v_2)^T d$. Jelikož $v_1 \in \mathcal{L}(d,w)$, $v_2 \in \mathcal{L}(d,w)^\perp$, platí

$$\|v_1+v_2\|^2 = \|v_1\|^2 + \|v_2\|^2 \ge \|v_1\|^2, \qquad (v_1+v_2)^T d = v_1^T d, \qquad (v_1+v_2)^T (d-w) = v_1^T (d-w), \quad (988)$$

takže $u_1 = u_2$ a $\delta_1 = 1 - v_1^T u_1 = 1 - (v_1+v_2)^T u_2 = \delta_2$. Pokud $\delta_1 = \delta_2 \le 0$, můžeme podle lemmatu 104 psát

$$\begin{aligned} \frac{\kappa(M_2)-1}{\kappa(M_2)+1} &= \sqrt{1 + 4\frac{\delta_2}{\|u_2\|^2\|v_1+v_2\|^2}} = \sqrt{1 - 4\frac{|\delta_2|}{\|u_2\|^2\|v_1+v_2\|^2}} \\ &\ge \sqrt{1 - 4\frac{|\delta_1|}{\|u_1\|^2\|v_1\|^2}} = \sqrt{1 + 4\frac{\delta_1}{\|u_1\|^2\|v_1\|^2}} = \frac{\kappa(M_1)-1}{\kappa(M_1)+1}. \end{aligned}$$

Podobným způsobem, tentokrát s použitím výrazu $1/\tau$, se dokáže, že tato nerovnost platí i pro $\delta_1 = \delta_2 > 0$. Jelikož funkce $(t-1)/(t+1)$ je pro $t \ge 0$ rostoucí, dostaneme $\kappa(M_2) \ge \kappa(M_1)$. □

Podle lemmatu 105 je účelné volit $v \in \mathcal{L}(d,w)$, což lze realizovat tak, že $v = \vartheta d - w = \vartheta d - A^{-1}y$ (Broydenovu dobrou metodu dostaneme pro $\vartheta = \infty$). Pak lze psát

$$A_+ = A + \frac{(y - Ad)(\vartheta d - A^{-1}y)^T}{(\vartheta d - A^{-1}y)^T d}. \quad (989)$$

Věta 244. *Nechť* A_+ *je matice určená podle vzorce (989), takže platí (983), kde* $v = \vartheta d - w$ *a* $w = A^{-1}y$. *Předpokládejme, že vektory* d *a* w *jsou lineárně nezávislé a označme* $a = d^T d$, $b = d^T w$, $c = w^T w$, *takže* $a > 0$, $b > 0$ *a* $ac > b^2$. *Pak* $\kappa(M)$ *je minimální právě tehdy, když*

(a) $\vartheta = \sqrt{c/a}$, *pokud* $a \ge b$ *a* $c \ge b$,
(b) $\vartheta = -\sqrt{c/a}$, *pokud* $a > b > c$ *nebo* $c > b > a$.

Důkaz Označme $u = (d-w)/v^T d$, $v = \vartheta d - w$. Pak platí

$$\begin{aligned} v^T u &= \frac{(d-w)^T(\vartheta d - w)}{(\vartheta d - w)^T d} = \frac{\vartheta(a-b) + (c-b)}{\vartheta a - b}, \\ \|u\|^2 &= \frac{(d-w)^T(d-w)}{((\vartheta d - w)^T d)^2} = \frac{a - 2b + c}{(\vartheta a - b)^2}, \\ \|v\|^2 &= (\vartheta d - w)^T(\vartheta d - w) = \vartheta^2 a - 2\vartheta b + c, \end{aligned}$$

takže $a - 2b + c > 0$, $\vartheta^2 a - 2\vartheta b + c > 0$ a

$$\delta = 1 - v^T u = \frac{\vartheta b - c}{\vartheta a - b}, \quad (990)$$

$$\varphi = \frac{1}{4}(\tau^2 - 1) = \frac{1 - v^T u}{\|u\|^2\|v\|^2} = \frac{(\vartheta a - b)(\vartheta b - c)}{(a - 2b + c)(\vartheta^2 a - 2\vartheta b + c)}. \quad (991)$$

Zřejmě τ má lokální extrém právě tehdy, když φ má lokální extrém. Derivujeme-li funkci $\varphi(\vartheta)$ podle ϑ, dostaneme po přímých ale formálně komplikovaných úpravách (kdy se téměř vše vyruší) vyjádření

$$\varphi'(\vartheta) = \frac{(\vartheta^2 a - c)(ac - b^2)}{(a - 2b + c)(\vartheta^2 a - 2\vartheta b + c)^2},$$

takže $\varphi'(\vartheta) = 0$ právě tehdy, když $\vartheta^2 a = c$, neboli když $\vartheta = \pm\sqrt{c/a}$. Rozšířením zlomku (990) číslem $\vartheta a + b$ a použitím vztahu $\vartheta^2 a = c$, dostaneme

$$\delta(\vartheta) = \frac{\vartheta b - c}{\vartheta a - b} = \frac{(\vartheta b - c)(\vartheta a + b)}{\vartheta^2 a^2 - b^2} = \frac{\vartheta(b^2 - ac)}{ac - b^2} = -\vartheta$$

a podobné úpravy výrazu (991) vedou k vyjádření

$$\begin{aligned} \varphi(\vartheta) &= \frac{(\vartheta a - b)(\vartheta b - c)}{(a - 2b + c)(\vartheta^2 a - 2\vartheta b + c)} = \frac{2bc - \vartheta(ac + b^2)}{2(a - 2b + c)(c - \vartheta b)} \, \frac{c + \vartheta b}{c + \vartheta b} \\ &= \frac{bc(ac - b^2) - \vartheta ac(ac - b^2)}{2c(a - 2b + c)(ac - b^2)} = \frac{b - \vartheta a}{2(a - 2b + c)}, \end{aligned} \tag{992}$$

což dává

$$\tau^2(\vartheta) = 1 + 4\varphi(\vartheta) = \frac{2(a - 2b + c) + 4(b - \vartheta a)}{2(a - 2b + c)} = \frac{a + c - 2\vartheta a}{a - 2b + c} \tag{993}$$

(tyto vztahy platí pouze pro $\vartheta = \pm\sqrt{c/a}$). Pokud $\vartheta = \sqrt{c/a}$, platí $\delta \leq 0$, takže $(\kappa - 1)/(\kappa + 1) = \tau$. Pokud $\vartheta = -\sqrt{c/a}$, platí $\delta > 0$, takže $(\kappa - 1)/(\kappa + 1) = 1/\tau$. Hodnotu $\vartheta = \sqrt{c/a}$ tedy volíme, pokud $\tau(\sqrt{c/a}) \leq 1/\tau(-\sqrt{c/a})$, neboli

$$\tau^2(\sqrt{c/a}) = \frac{1 + c - 2\sqrt{ac}}{a - 2b + c} \leq \frac{a - 2b + c}{1 + c + 2\sqrt{ac}} = \frac{1}{\tau^2(-\sqrt{c/a})}.$$

Jelikož $a - 2b + c > 0$ a $1 + c + 2\sqrt{ac} > 0$, je tato nerovnost ekvivalentní nerovnosti $(a+c)^2 - 4ac \leq (a - 2b + c)^2$, kterou lze upravit na $(a - b)(c - b) \geq 0$. Jelikož $a > 0$, $b > 0$ a $ac > b^2$, nemůže současně platit $a \leq b$ a $c \leq b$, takže nutně $a \geq b$ a $c \geq b$. Hodnotu $\vartheta = -\sqrt{c/a}$ pak volíme ve zbylých případech, tedy když $a > b > c$ nebo $c > b > a$. □

Lemma 106. *Nechť jsou splněny předpoklady lemmatu 105. Pak platí* $\|I - M_1\|\|I - M_1^{-1}\| \leq \|I - M_2\|\|I - M_2^{-1}\|$.

Důkaz Použijeme-li vzorce (983), (984) a vztah $\|uv^T\| = \sqrt{\|uv^T vu^T\|} = \|u\|\|v\|$, dostaneme

$$\|I - M\|\|I - M^{-1}\| = \frac{\|d - w\|\|v\|}{|v^T d|} \frac{\|d - w\|\|v\|}{|v^T w|} = \frac{\|d - w\|^2 \|v\|^2}{|v^T d||v^T w|}. \tag{994}$$

Jelikož jsou splněny předpoklady lemmatu 105, platí (988), takže

$$\|I - M_2\|\|I - M_2^{-1}\| = \frac{\|d - w\|^2 \|v_1 + v_2\|^2}{|(v_1 + v_2)^T d||(v_1 + v_2)^T w|} \geq \frac{\|d - w\|^2 \|v_1\|^2}{|v_1^T d||v_1^T w|} = \|I - M_1\|\|I - M_1^{-1}\|.$$

□

Věta 245. *Nechť jsou splněny předpoklady věty 244. Pak* $\|I - M\|\|I - M^{-1}\|$ *je minimální právě tehdy, když*

(a) $\vartheta = \sqrt{c/a}$, *pokud* $b \leq 0$,
(b) $\vartheta = -\sqrt{c/a}$, *pokud* $b > 0$.

Důkaz Podle vzorce (994) a vztahů uvedených na začátku důkazu věty 244 je výraz $\|I - M\|\|I - M^{-1}\|$ minimální, je-li zlomek

$$\frac{|v^T d||v^T u|}{\|v\|^2} = \frac{|\vartheta a - b||\vartheta b - c|}{\vartheta^2 a - 2\vartheta b + c}$$

maximální. Jelikož

$$\frac{|\vartheta a - b||\vartheta b - c|}{\vartheta^2 a - 2\vartheta b + c} = (a - 2b + c)|\varphi(\vartheta)|$$

a $a - 2b + c > 0$, jsou maxima tohoto výrazu extrémy funkce $\varphi(\vartheta)$, neboli $\vartheta = \pm\sqrt{c/a}$. Použijeme-li vztah (992), dostaneme

$$\begin{aligned} \varphi(\sqrt{c/a}) &= \frac{b + \sqrt{ac}}{2(a - 2b + c)} > 0, \\ \varphi(-\sqrt{c/a}) &= \frac{b - \sqrt{ac}}{2(a - 2b + c)} < 0 \end{aligned}$$

($a > 0$, $b > 0$ a neboť $ac > b^2$). Hodnotu $\vartheta = \sqrt{c/a}$ tedy volíme, pokud

$$\varphi(\sqrt{c/a}) = \frac{\sqrt{ac} + b}{2(a - 2b + c)} \leq \frac{\sqrt{ac} - b}{2(a - 2b + c)} = -\varphi(\sqrt{-c/a}),$$

neboli když $b \leq 0$. Hodnotu $\vartheta = -\sqrt{c/a}$ pak volíme ve zbylém případě, tedy když $b > 0$ □

Kvazinewtonovské metody lze (podobě jako metody s proměnnou metrikou vyšetřované v oddílu 4.1) realizovat tak, že místo matice $A \approx J$ aktualizujeme matici $S = A^{-1} \approx J^{-1}$.

Věta 246. *(Aktualizace matice $S = A^{-1}$). Nechť jsou splněny předpoklady věty 243. Nechť $S = A^{-1}$ a $S_+ = A_+^{-1}$, kde A_+ je matice určená podle aktualizace (978) s $v^T A^{-1} y \neq 0$. Pak platí*

$$S_+ = S + \frac{(d - Sy)v^T S}{v^T S y} = S + \frac{(d - Sy)z^T}{z^T y} \tag{995}$$

kde $z = S^T v$.

Důkaz Podle (982), kde $u = (y - Ad)/v^T d$, platí

$$S_+ = S - \frac{Suv^T S}{\delta} = S + \frac{(d - Sy)v^T S}{\delta v^T d},$$

kde δ je zatím neznámé číslo. Z rovnice $S_+ y = d$ však plyne

$$S_+ y = Sy + \frac{v^T S y}{\delta v^T d}(d - Sy) = d.$$

takže nutně $\delta = v^T S y / v^T d$. □

Poznámka 360. Položíme-li $v = d$, neboli $z = S^T d$, dostaneme Broydenovu dobrou metodu

$$S_+ = S + \frac{(d - Sy)d^T S}{d^T S y}. \tag{996}$$

Položíme-li $v = (S^{-1})^T y$, neboli $z = y$, dostaneme Broydenovu špatnou metodu

$$S_+ = S + \frac{(d - Sy)y^T}{y^T y}. \tag{997}$$

Nechť

$$e_k^T y = \max_{1 \leq i \leq n} e_i^T y.$$

Položíme-li $S^T v = e_k$, neboli $z = e_k$, dostaneme inverzní metodu aktualizace sloupců

$$S_+ = S + \frac{(d - Sy)e_k^T}{e_k^T y}. \tag{998}$$

Poznámka 361. (Dualita). Vztah (995) dostaneme ze vztahu (978) záměnou $d \to y$, $y \to d$, $A \to S$. Dobrá a špatná Broydenova metoda jsou vzájemně duální. Podobně přímá a inverzní metoda aktualizace sloupců jsou vzájemně duální.

Poznámka 362. Položíme-li $v = \vartheta d - Sy$, neboli $z = S^T v = \vartheta S^T d - S^T Sy$, platí

$$S_+ = S + \frac{(d - Sy)(\vartheta d - Sy)^T S}{(\vartheta d - Sy)^T Sy},$$

což je inverzní tvar vzorce (989). Zvolíme-li parametr ϑ podle věty 244, minimalizuje tato metoda číslo podmíněnosti matice $A^{-1}A_+ = SS_+^{-1}$. Duální metoda

$$S_+ = S + \frac{(d - Sy)(\vartheta y - S^{-1}d)^T}{(\vartheta y - S^{-1}d)^T y}$$

minimalizuje číslo podmíněnosti matice $S^{-1}S_+ = AA_+^{-1}$, volíme-li parametr ϑ podle věty 244, kde nyní $a = \|y\|^2$, $b = y^T S^{-1}d$, $c = \|S^{-1}d\|^2$.

Poznámka 363. Dobrá Broydenova metoda a optimálně podmíněná metoda (989) dávají velmi dobré výsledky. Metody k nim duální jsou méně efektivní.

Kvazinewtonovské metody pro řešení soustav lineárních rovnic lze též odvodit pomocí minimalizačního principu.

Věta 247. *Nechť W je čtvercová regulární matice řádu n. Pak matici A_+, která je řešením úlohy*

$$\|(A_+ - A)W^{-1}\|_F = \min_{\tilde{A}d=y} \|(\tilde{A} - A)W^{-1}\|_F \tag{999}$$

lze vyjádřit ve tvaru (978), kde $v = W^T Wd$. Pokud $W = I$ dostaneme Broydenovu dobrou metodu.

Důkaz Položme $B = AW^{-1}$, $\tilde{B} = \tilde{A}W^{-1}$, $B_+ = A_+W^{-1}$. Pak úloha (999) je ekvivalentní úloze

$$\|B_+ - B\|_F = \min_{\tilde{B}\tilde{d}=y} \|\tilde{B} - B\|_F, \tag{1000}$$

kde $\tilde{d} = Wd$ (neboť $\tilde{B}\tilde{d} = \tilde{A}W^{-1}Wd = \tilde{A}d$). Řešení úlohy (1000) lze podle věty 219 zapsat ve tvaru

$$B_+ = B + \frac{(y - B\tilde{d})\tilde{d}^T}{\tilde{d}^T\tilde{d}},$$

neboli

$$A_+W^{-1} = AW^{-1} + \frac{(y - AW^{-1}Wd)d^T W^T}{d^T W^T Wd},$$

což po vynásobení zprava maticí W dává (978), kde $v = W^T Wd$. □

Poznámka 364. Analogický postup lze použít pro aktualizaci matice S. Nechť W je čtvercová regulární matice. Pak matici S_+ minimalizující Frobeniovu normu $\|(\tilde{S} - S)W^{-1}\|_F$ na množině čtvercových matic $\tilde{S}$ řádu n splňujících kvazinewtonovskou podmínku $\tilde{S}y = d$ lze vyjádřit ve tvaru (995), kde $z = W^T Wy$. Pokud $W = I$ dostaneme Broydenovu špatnou metodu.

Kvazinewtonovské metody splňují kvazinewtonovskou podmínku podobně jako metody s proměnnou metrikou (stačí porovnat (977) a (272)). Metody s proměnnou metrikou s přesným výběrem délky kroku nalezenou minimum ryze konvezní kvadratické funkce $Q(x)$ po konečném počtu kroků. Ukážeme, že kvazinewtonovské metody s jednotkovým výběrem délky kroku ($\alpha_i = 1$, $i \in N$) naleznou řešení soustavy lineárních rovnic

$$J^*(x - x^*) = 0 \tag{1001}$$

s regulární maticí J^* také po konečném počtu kroků. Při důkazu tohoto tvrzení budeme používat vyjádření

$$x_{i+1} = x_i - S_i f_i \tag{1002}$$

a

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} \tag{1003}$$

pro $i \in N$, kde S_i jsou regulární matice, $f_i \neq 0$ a $z_i^T y_i \neq 0$ (zde $z_i = S_i^T v_i$).

Lemma 107. *Uvažujme iterační proces (1002), (1003) aplikovaný na soustavu lineárních rovnic (1001) s regulární maticí. Pak pro libovolný index $i \in N$ a pro libovolný exponent $k \geq 0$ je vektor $(J^* S_{i+1})^k f_{i+1}$ lineární kombinací vektorů $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$.*

Důkaz (indukcí). Předpokládejme, že pro nějaký exponent $k \geq 0$ je vektor $(J^* S_{i+1})^k f_{i+1}$ lineární kombinací vektorů $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$. Platí to zcela jistě pro $k = 0$, neboť z (1001) a (1002) plyne

$$y_i = f_{i+1} - f_i = J^* d_i = -J^* S_i f_i, \tag{1004}$$

takže

$$(J^* S_{i+1})^0 f_{i+1} = f_{i+1} = f_i + y_i = f_i - J^* S_i f_i = (I - J^* S_i)(J^* S_i)^0 f_i.$$

Použijeme-li (1003) a (1004), dostaneme

$$J^* S_{i+1} = J^* S_i + (J^* d_i - J^* S_i y_i) \frac{z_i^T}{z_i^T y_i} = J^* S_i - (I - J^* S_i) J^* S_i f_i \frac{z_i^T}{z_i^T y_i}.$$

Jelikož vektor $(J^* S_{i+1})^k f_{i+1}$ je lineární kombinací vektorů $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k$ a jelikož matice $J^* S_i$ a $(I - J^* S_i)$ komutují, je vektor $(J^* S_{i+1})^{k+1} f_{i+1} = J^* S_{i+1} (J^* S_{i+1})^k f_{i+1}$ lineární kombinací vektorů $(I - J^* S_i)(J^* S_i)^j f_i$, $0 \leq j \leq k + 1$. □

Lemma 108. *Nechť jsou splněny předpoklady lemmatu 107 a nechť $i \in N$ je index takový, že vektor f_{i+1} není násobkem vektoru f_i. Pak vektory $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, jsou lineárně nezávislé pro každé číslo $l \in N$ takové, že $2l \leq i + 1$.*

Důkaz (indukcí). Předpokládejme, že vektory $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, jsou lineárně nezávislé pro nějaké číslo $l \in N$ takové, že $2l \leq i - 1$. Platí to zcela jistě pro $l = 1$, neboť podle (1004) dostaneme

$$\begin{aligned} (J^* S_i)^0 f_i &= f_i, \\ (J^* S_i)^1 f_i &= -y_i = f_i - f_{i+1} \end{aligned}$$

a tyto vektory jsou lineárně nezávislé, neboť vektor f_{i+1} není násobkem vektoru f_i.
(a) Podle lemmatu 107 je vektor $(J^* S_{i-2l+2})^k f_{i-2l+2}$ lineární kombinací vektorů $(I - J^* S_{i-2l+1})(J^* S_{i-2l+1})^j$ f_{i-2l+1}, $0 \leq j \leq k$. Jelikož $l + 1$ lineárně nezávislých vektorů $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, vyjadřujeme pomocí $l + 1$ vektorů $(I - J^* S_{i-2l+1})(J^* S_{i-2l+1})^k f_{i-2l+1}$, $0 \leq k \leq l$, musí být tyto vektory také lineárně nezávislé. Odtud bezprostředně plyne, že i vektory $(J^* S_{i-2l+1})^k f_{i-2l+1}$, $0 \leq k \leq l$, jsou lineárně nezávislé.
(b) Použijeme-li (1004), dostaneme

$$y_{i-2l} = -J^* S_{i-2l} f_{i-2l} \neq 0.$$

Ukážeme, že vektor y_{i-2l} není lineární kombinací vektorů $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$. Použijeme-li kvazinewtonovskou podmínku

$$S_{i-2l+1}y_{i-2l} = d_{i-2l} = (J^*)^{-1}y_{i-2l},$$

můžeme psát

$$(I - J^*S_{i-2l+1})y_{i-2l} = 0. \tag{1005}$$

Předpokládejme, že vektor y_{i-2l} je lineární kombinací vektorů $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$. Pak odpovídající lineární kombinace vektorů $(I - J^*S_{i-2l+1})(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, by musela být nulová (viz (1005), což je spor s lineární nezávislostí těchto vektorů (viz (a)).
(c) Podle lemmatu 107 je vektor $(J^*S_{i-2l+1})^k f_{i-2l+1}$, lineární kombinací vektorů $(I - J^*S_{i-2l})(J^*S_{i-2l})^j$ f_{i-2l}, $0 \le j \le k$, a tedy i lineární kombinací vektorů $(J^*S_{i-2l})^j f_{i-2l}$, $0 \le j \le k+1$. Navíc vektor y_{i-2l} lze vyjádřit ve tvaru $y_{i-2l} = -J^*S_{i-2l}f_{i-2l}$, (viz (γ)). Jelikož $l+2$ lineárně nezávislých vektorů y_{i-2l} a $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$ (viz (b)) vyjadřujeme pomocí $l+2$ vektorů $(J^*S_{i-2l})^k f_{i-2l}$, $0 \le k \le l+1$, musí být tyto vektory také lineárně nezávislé. □

Věta 248. *Nechť jsou splněny předpoklady lemmatu 107. Pak existuje index $1 \le i \le 2n-1$ takový, že $f_{i+2} = 0$, takže bod $x_{i+2} \in R^n$ je řešením soustavy lineárních rovnic (1001).*

Důkaz Předpokládejme, že pro $i = 2n-1$ není vektor f_{i+1} násobkem vektoru f_i. Pak podle lemmatu 108 jsou vektory $(J^*S_{2n-2l+1})^k f_{2n-2l+1}$, $0 \le k \le l$, lineárně nezávislé pro každé číslo $l \in N$ takové, že $l \le n$. Pro $l = n$ je těchto vektorů $n+1$, což je ve sporu s tím, že mají dimenzi n. Existuje tedy index $1 \le i \le 2n-1$ takový, že vektor f_{i+1} je násobkem vektoru f_i, neboli

$$f_{i+1} = \lambda_i(f_{i+1} - f_i) = \lambda_i y_i.$$

Podle (1003) a (1004) pak platí

$$f_{i+2} = f_{i+1} + y_{i+1} = f_{i+1} - J^*S_{i+1}f_{i+1} = \lambda_i(y_i - J^*S_{i+1}y_i) = \lambda_i(y_i - J^*d_i) = \lambda_i(y_i - y_i) = 0,$$

takže bod $x_{i+2} \in R^n$ je řešením soustavy lineárních rovnic (1001). □

Nevýhodou kvazinewtonovských metod realizovaných standardním způsobem (kdy se v každém iteračním kroku provádí aktualizace (978)) je skutečnost, že není zaručena jejich globální konvergence (matice A_i, $i \in N$, obvykle nesplňují předpoklady A3b s (956) a A5b). Proto je třeba tyto metody kombinovat s Newtonovou metodou nebo její diferenční verzí. Kvazinewtonovské metody spádových směrů se obvykle realizují tak, že se v algoritmu 27 pokládá $A_1 = J_1$ (nebo $S_1 = J_1^{-1}$) a kdykoliv nelze splnit podmínku (S2a) (nebo (S2b), nebo (S2c)), iterační proces se přeruší a položí se $A_{i+1} = J_{i+1}$ (nebo $S_{i+1} = J_{i+1}^{-1}$). Jelikož se v některých iteračních krocích používá Jacobiova matice je vhodné používat trojúhelníkový nebo ortogonální rozklad matice A. Použijeme-li ortogonální rozklad popsaný v oddílu 11.9, platí $A = QR$ a $S = R^{-1}Q^T$, takže je prakticky jedno, pracujeme-li s maticí A nebo s maticí S (implementačně vhodnější je pracovat s maticí $A = QR$). Kvazinewtonovské metody s lokálně omezeným krokem se obvykle realizují tak, že se pokládá $A_1 = J_1$ a v případě (T3a), se položí $A_{i+1} = J_{i+1}$ zatímco v případě (T3b) se matice A_{i+1} aktualizuje podle (978). Tyto úpravy mají své opodstatnění, neboť platí tato věta.

Věta 249. *Nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ taková, že pokud $\|x_1 - x^*\| \le \overline{\delta}$ a $\|A_1 - J_1\| \le \overline{\vartheta}$, posloupnost x_i, $i \in N$, určená dobrou Broydenovou metodou (979) s jednotkovým výběrem délky kroku ($\alpha_i = 1$, $i \in N$), konverguje Q-superlineárně k bodu $x^* \in R^n$.*

Důkaz Věta 249 je speciálním důsledkem obecnějších vět 267 a 268 dokázaných v oddílu 12.4. □

11.6 Nemonotonní kvazinewtonovské metody

Matice A_i, $i \in N$, generované kvazinewtonovkými metodami obecně nesplňují předpoklady A3b–A5b s (956). V důsledku toho nemusí být směrové vektory s_i, $i \in N$, spádové pro funkci $\|f\|$ (v bodech x_i, $i \in N$). S druhé strany funkce $\|f\|$ není optimální účelovou funkcí pro řešení soustav nelineárních rovnic (při nepodmíněné minimalizaci, která je v jistém smyslu ekvivalentní řešení soustavy nelineárních rovnic $g(x) = 0$, nevyžadujeme monotonní snižování normy gradientu). Proto je logické nahradit monotonní kritéria ($\overline{\text{S2a}}$)–($\overline{\text{S2c}}$) slabšími nemonotonními kriterii. Například podmínku ($\overline{\text{S2c}}$) lze nahradit podmínkou

$$\|f_{i+1}\| - \|f_i\| \leq -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\| + \eta_i, \qquad (\overline{\text{S2d}})$$

$i \in N$, kde

$$\sum_{i=1}^{\infty} \eta_i = \overline{\eta} < \infty. \qquad (1006)$$

Lemma 109. *(Konzistence) Vyhovuje-li zobrazení $f : R^n \to R^n$ předpokladu J3, je podmínka ($\overline{\text{S2d}}$) splněna, pokud $\alpha_i \leq \eta_i/(\overline{J}\|s_i\| + \underline{\rho}(1-\overline{\omega})\|f_i\|)$.*

Důkaz Podle předpokladu J3 platí

$$\|f_{i+1}\| - \|f_i\| \leq |\|f_{i+1}\| - \|f_i\|| \leq \|f_{i+1} - f_i\| = \|f(x_i + \alpha_i s_i) - f(x_i)\| \leq \overline{J}\alpha_i\|s_i\|,$$

takže podmínka ($\overline{\text{S2d}}$) je splněna pokud $\overline{J}\alpha_i\|s_i\| \leq -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\| + \eta_i$, což dává tvrzení lemmatu. □

Definice 83. *Nemonotonní metodou spádových směrů nazveme metodu spádových směrů (definice 80), kde podmínka ($\overline{\text{S2c}}$) je nahražena podmínkou ($\overline{\text{S2d}}$), přičemž $A_i = J_i$ a $\eta_i = 0$ pro $i \in M$. Přitom $M \subset N$, matice A_i, $i \in N$, jsou regulární a čísla $\eta_i \geq 0$, $i \in N$, splňují podmínku (1006).*

Věta 250. *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům J1, J4, J5a a J6. Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná nemonotonní metodou spádových směrů (definice 83), kde množina $M \subset N$ je nekonečná. Pak $f(x_i) \to 0$.*

Důkaz (a) Ukážeme nejprve, že $\liminf_{i\to\infty} \|f(x_i)\| = 0$. Předpokládejme naopak, že existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|f(x_i)\| \geq \underline{\varepsilon}\ \forall i \in N$. Nechť $N_1 = \{i \in N\backslash M : \|f_{i+1}\| > \|f_i\|\}$, $N_2 = \{i \in N\backslash M : \|f_{i+1}\| \leq \|f_i\|\}$ a $K = \{1, \ldots, k\}$. Pak podle ($\overline{\text{S2c}}$) a (1006) dostaneme

$$\begin{aligned} \|f_{k+1}\| &= \|f_1\| + \sum_{i=1}^{k}(\|f_{i+1}\| - \|f_i\|) \leq \|f_1\| - \sum_{i\in M\cap K} \underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\| + \sum_{i\in N_1\cap K} \eta_i \\ &\leq \|f_1\| + \overline{\eta} - \sum_{i\in M\cap K} \underline{\rho}(1-\overline{\omega})\underline{\alpha}\|f_i\| \leq \|f_1\| + \overline{\eta} - k_1\underline{\rho}(1-\overline{\omega})\underline{\alpha}\,\underline{\varepsilon}, \end{aligned}$$

kde k_1 je počet prvků množiny $M \cap K$, neboť podle lemmatu 100 platí $\alpha_i \geq \underline{\alpha}$, $i \in M$, a podle předpokladu je $\|f_i\| \geq \underline{\varepsilon}$, $i \in N$. Jelikož množina M je nekonečná, můžeme $k \in N$ volit tak, že

$$k_1 > \frac{\|f_1\| + \overline{\eta}}{\underline{\rho}(1-\overline{\omega})\underline{\alpha}\,\underline{\varepsilon}}.$$

Pak ale $\|f_{k+1}\| \leq \|f_1\| + \overline{\eta} - k_1\underline{\rho}(1-\overline{\omega})\underline{\alpha}\,\underline{\varepsilon} < \|f_1\| + \overline{\eta} - (\|f_1\| + \overline{\eta}) = 0$, což je spor, neboť norma je vždy nezáporná.

(b) Ukážeme nyní, že $\limsup_{i\to\infty} \|f(x_i)\| = 0$. Nechť naopak $0 < \underline{\varepsilon} < \overline{\varepsilon} < \limsup_{i\to\infty} \|f_i\|$. Podle (1006) existuje index $k \in N$ takový že

$$\sum_{i=k}^{\infty} \eta_i < \overline{\varepsilon} - \underline{\varepsilon}.$$

Jelikož podle (a) platí $0 = \liminf_{i\to\infty} \|f_i\| < \underline{\varepsilon}$ a jelikož předpokládáme, že $\overline{\varepsilon} < \limsup_{i\to\infty} \|f_i\|$, existují indexy $k_2 > k_1 \geq k$ takové, že $\|f_{k_1}\| \leq \underline{\varepsilon}$ a $\|f_{k_2}\| \geq \overline{\varepsilon}$. Podle ($\overline{\text{S2c}}$) však platí

$$\|f_{k_2}\| = \|f_{k_1}\| + \sum_{i=k_1}^{k_2-1} (\|f_{i+1}\| - \|f_i\|) \leq \|f_{k_1}\| + \sum_{i=k_1}^{k_2-1} \eta_i \leq \underline{\varepsilon} + \sum_{i=k}^{\infty} \eta_i < \underline{\varepsilon} + (\overline{\varepsilon} - \underline{\varepsilon}) = \overline{\varepsilon},$$

což je ve sporu s předpokladem, že $\|f_{k_2}\| \geq \overline{\varepsilon}$. □

Poznámka 365. V předpokladech věty 250 je podstatné, že množina M je nekonečná. Tuto podmínku splňují například metody popsané v oddílech 12.1 a 12.6. Na matice A_i, $i \in N\backslash M$, nejsou kladeny žádné požadavky (kromě regularity, která zaručuje splnění podmínky ($\overline{\text{S1}}$)).

Poznámka 366. Věta 250 zaručuje, že $\|f_i\| \to 0$. Posloupnost x_i, $i \in N$, však nemusí konvergovat. Pokud má tato posloupnost hromadný bod $x^* \in R^n$, platí $f(x^*) = 0$.

Nyní se budeme zabývat nemonotonními metodami spádových směrů, kde množina M je určena předpisem

$$M = \{l \in N : \ l = (j-1)m + 1, \ j \in N\}, \tag{1007}$$

a kde matice A_i, $i \notin M$, splňují slabý princip omezeného znehodnocení

$$\|A_i - J_i\| \leq c_1\|A_{i-1} - J_{i-1}\| + c_2\|x_i - x_{i-1}\| \tag{1008}$$

($c_1 > 0$ a $c_2 > 0$ jsou konstanty nezávislé na indexu $i \notin M$).

Lemma 110. *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům J1, J4, J5a a J6. Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná nemonotonní metodou spádových směrů (definice 83), pro kterou platí (1007) a (1008). Pak*

$$\lim_{i\to\infty} \|A_i - J_i\| = 0, \qquad \lim_{i\to\infty} \|x_{i+1} - x_i\| = 0.$$

Důkaz Nechť $i = l + k - 1$, kde $l \in M$ a $1 \leq k \leq m$. Dokážeme indukcí, že pro libovolný index $1 \leq k \leq m$ platí

$$\lim_{l\to\infty} \|A_{l+k-1} - J_{l+k-1}\| = 0, \qquad \lim_{l\to\infty} \|x_{l+k} - x_{l+k-1}\| = 0. \tag{1009}$$

Pro $k = 1$ je $\|A_l - J_l\| = 0$, takže platí A3a s $\vartheta = 0$ a A5a s $\underline{A} = \underline{J}$. Můžeme tedy použít nerovnost (964), podle které

$$\|x_{l+1} - x_l\| = \|\alpha_l s_l\| \leq \|s_l\| \leq \frac{1+\overline{\omega}}{\underline{J}}\|f_l\|$$

(neboť $0 < \alpha_l \leq 1$), což spolu s $\|f_l\| \to 0$ (věta 250) dává $\|x_{l+1} - x_l\| \to 0$. Předpokládejme nyní, že (1009) platí pro nějaký index $1 \leq k < m$ (dokázali jsme to pro $k = 1$). Použijeme-li (1008) dostaneme

$$\lim_{l\to\infty} \|A_{l+k} - J_{l+k}\| \leq c_1 \lim_{l\to\infty} \|A_{l+k-1} - J_{l+k-1}\| + c_2 \lim_{l\to\infty} \|x_{l+k} - x_{l+k-1}\| = 0.$$

To znamená, že pro dostatečně velký index $l \in M$ splňuje matice A_{l+k} předpoklady A3a s (955) a A5a. Můžeme tedy použít (964), takže z $\|f_{l+k}\| \to 0$ plyne $\|x_{l+k+1} - x_{l+k}\| \to 0$. Tím je indukční krok dokončen. Zvolme libovolně číslo $\varepsilon > 0$. Z (1009) plyne existence čísel $n_k \in N$, $1 \leq k \leq m$, takových, že pro $l \geq n_k$ platí $\|A_{l+k-1} - J_{l+k-1}\| < \varepsilon$ a $\|x_{l+k} - x_{l+k-1}\| < \varepsilon$. Položme $n = \max(n_1, \ldots, n_m)$. Pak pro $i \geq n$ platí $\|A_i - J_i\| < \varepsilon$ a $\|x_{i+1} - x_i\| < \varepsilon$ a jelikož číslo ε bylo zvoleno libovolně, dostaneme tvrzení lemmatu. □

Věta 251. *Nechť jsou splněny předpoklady lemmatu 110. Nechť posloupnost x_i, $i \in N$, má hromadný bod $x^* \in \mathcal{D}$, kde Jacobiova matice $J(x^*)$ je regulární. Pak $x_i \to x^*$ a $f(x^*) = 0$. Jestliže navíc $(A_i s_i + f_i)/\|f_i\| \to 0$ a $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínce ($\overline{\text{S2d}}$), pak $x_i \to x^*$ Q-superlineárně.*

Důkaz (a) Nechť $\underline{J} \leq (1/2)\|J^{-1}(x^*)\|^{-1}$. Jelikož předpokládáme, že $x^* \in \mathcal{D}$ a $f \in C^1$ na $\mathcal{D}$, závisejí v okolí bodu x^* koeficienty a tudíž i singulární čísla Jacobiovy matice spojitě na x, takže existuje číslo $\delta > 0$ takové, že $J(x)d \geq \underline{J}\|d\|$ $\forall x \in \mathcal{B}(x^*, \delta)$. Nechť $0 < \varepsilon < \delta$ a $P(x^*, \varepsilon, \delta) = \{x \in R^n : \varepsilon < \|x - x^*\| < \delta\}$. Pak podle věty 233 $P(x^*, \varepsilon, \delta)$ neobsahuje žádné řešení soustavy nelineárních rovnic $f(x) = 0$ a tudíž ani žádný hromadný bod posloupnosti x_i, $i \in N$. Existuje tedy index $k_1 \in N$ takový, že $x_i \notin P(x^*, \varepsilon, \delta)$, pokud $i \geq k_1$. Jelikož $x_i \to x^*$, existuje index $k_2 \geq k_1$ takový, že $\|x_i - x^*\| < \varepsilon$, pokud $i \geq k_2$ a jelikož $\|x_{i+1} - x_i\| \to 0$, existuje index $k \geq k_2$ takový, že $\|x_{i+1} - x_i\| < \delta - \varepsilon$, pokud $i \geq k$. Pak ale $\|x_{i+1} - x^*\| \leq \|x_i - x^*\| + \|x_{i+1} - x_i\| < \delta$, pokud $i \geq k$, a jelikož $x_{i+1} \notin P(x^*, \varepsilon, \delta)$, musí být $\|x_{i+1} - x^*\| < \varepsilon$, pokud $i \geq k$. Postupujeme-li takto dále, dostaneme $\|x_i - x^*\| < \varepsilon$ $\forall i \geq k$ a jelikož číslo $\varepsilon > 0$ bylo vybráno libovolně, platí $x_i \to x^*$.

(b) Podle lemmatu 110 platí $(A_i - J_i)s_i/\|s_i\| \to 0$. Jestliže navíc $(A_i s_i + f_i)/\|f_i\| \to 0$, jsou splněny předpoklady věty 235, takže $x_i \to x^*$ Q-superlineárně. □

Na závěr ukážeme, že kvazinewtonovské metody většinou splňují slabý princip omezeného znehodnocení, takže pro ně platí věta 251.

Věta 252. *Nechť zobrazení $f : R^n \to R^n$ splňuje předpoklad J6 a nechť*

$$A_{i+1} = A_i + \frac{(y_i - A_i d_i)v_i^T}{v_i^T d_i}, \tag{1010}$$

kde $d_i = x_{i+1} - x_i$, $y_i = f_{i+1} - f_i$ a $|v_i^T d_i| \geq \underline{\gamma}\|v_i\|\|d_i\|$. Pak

$$\|A_{i+1} - J_{i+1}\| \leq c_1\|A_i - J_i\| + c_2\|d_i\|.$$

kde $c_1 = 1 + 1/\underline{\gamma}$ a $c_2 = \overline{G}(1 + 1/\underline{\gamma})$.

Důkaz Použijeme-li (1010), dostaneme

$$\begin{aligned} A_{i+1} - J_{i+1} &= A_i - J_i + J_i - J_{i+1} - \frac{(A_i - J_i)d_i v_i^T}{v_i^T d_i} + \frac{(y_i - J_i d_i)v_i^T}{v_i^T d_i} \\ &= J_i - J_{i+1} + (A_i - J_i)\left(I - \frac{d_i v_i^T}{v_i^T d_i}\right) + \frac{(y_i - J_i d_i)v_i^T}{v_i^T d_i} \end{aligned}$$

Podle věty o střední hodnotě (tvrzení 6) a předpokladu J6 platí

$$\begin{aligned} \|y_i - J_i d_i\| &= \left\| \int_0^1 (J(x_i + \lambda d_i) - J(x_i))d_i \mathrm{d}\lambda \right\| \leq \int_0^1 \|J(x_i + \lambda d_i) - J(x_i)\|\|d_i\|\mathrm{d}\lambda \\ &\leq \int_0^1 \overline{G}\|d_i\|^2 \lambda \mathrm{d}\lambda = \frac{1}{2}\overline{G}\|d_i\|^2 \leq \overline{G}\|d_i\|^2. \end{aligned}$$

Můžeme tedy psát

$$\begin{aligned} \|A_{i+1} - J_{i+1}\| &\leq \overline{G}\|d_i\| + \|A_i - J_i\|\left(1 + \frac{\|d_i\|\|v_i\|}{|v_i^T d_i|}\right) + \overline{G}\|d_i\|\frac{\|d_i\|\|v_i\|}{|v_i^T d_i|} \\ &\leq (1 + \frac{1}{\underline{\gamma}})\|A_i - J_i\| + \overline{G}(1 + \frac{1}{\underline{\gamma}})\|d_i\|, \end{aligned}$$

což dokazuje tvrzení věty. □

11.7 Sdružené kvazinewtonovské metody

Newtonova metoda, která používá první derivace zobrazení $f : R^n \to R^n$, konverguje velmi rychle, ale spotřebuje v každém iteračním kroku $O(n^3)$ aritmetických operací (na řešení soustavy rovnic $Js + f = 0$). Kvazinewtonovské metody, které nepoužívají první derivace, konvergují pomaleji, ale spotřebují v každém iteračním kroku pouze $O(n^2)$ aritmetických operací (používáme-li aktualizace popsané v oddílu 11.9). Proto je rozumné vyvíjet metody, které pro urychlení konvergence používají první derivace zobrazení f, ale spotřebují v každém iteračním kroku pouze $O(n^2)$ aritmetických operací. Tyto metody pracují s vektory J_+d a J_+^Tq, které lze určit buď ze znalosti Jacobiovy matice J_+, nebo pomocí automatického derivování popsaného v oddílu 14 (vektor J_+d lze také určit pomocí numerického derivování).

Poznámka 367. Nahradíme-li v aktualizaci (978) vektor y vektorem J_+d, dostaneme

$$A_+ = A - \frac{(A - J_+)dv^T}{v^Td}.$$

Pak platí $A_+d = J_+d$. Tento přístup není příliš významný, neboť vektor y je obvykle dobrou aproximací vektoru J_+d a není proto nutné počítat první derivace.

Poznámka 368. Položíme-li

$$A_+ = A - \frac{p\, f_+^T(A - J_+)}{f_+^Tp}, \tag{1011}$$

kde $f_+^Tp \neq 0$, platí $f_+^TA_+ = f_+^TJ_+$, takže vektor $A_+^Tf_+$ se rovná gradientu funkce $F = (1/2)\|f\|^2$ v bodě x_+. To má velký význam, neboť řešíme-li soustavu rovnic $A_+s_+ + f_+ = 0$ přesně, platí

$$g_+^Ts_+ = f_+^TJ_+s_+ = f_+^TA_+s_+ = -f_+^Tf_+ < 0,$$

takže směrový vektor s_+ je spádový pro funkci $F = (1/2)\|f\|^2$ v bodě x_+.

Metody tohoto typu nazýváme sdruženými kvazinewtonovskými metodami. Obecný tvar sdružených kvazinewtonovských metod je definován takto.

Definice 84. *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 78) je sdruženou kvazinewtonovskou metodou, jestliže*

$$A_is_i + f_i = 0,$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i - \frac{p_iq_i^T(A_i - J_{i+1})}{q_i^Tp_i}, \tag{1012}$$

kde $p_i \in R^n$, $q_i \in R^n$ jsou vektory takové, že $q_i^Tp_i \neq 0$. Pak platí $A_{i+1}^Tq_i = J_{i+1}^Tq_i$

Poznámka 369. Položíme-li v (1011) $p = (A - J_+)d$, dostaneme

$$A_+ = A - \frac{(A - J_+)d\, f_+^T(A - J_+)}{f_+^T(A - J_+)d}. \tag{1013}$$

V tomto případě platí současně $A_+d = J_+d$ a $f_+^TA_+ = f_+^TJ_+$.

Metody tohoto typu nazýváme oboustrannými kvazinewtonovskými metodami. Obecný tvar oboustranných kvazinewtonovských metod je definován takto.

Definice 85. *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 78) je oboustrannou kvazinewtonovskou metodou, jestliže*

$$A_is_i + f_i = 0,$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i - \frac{(A_i - J_{i+1})d_i q_i^T (A_i - J_{i+1})}{q_i^T (A_i - J_{i+1})d_i}, \tag{1014}$$

kde $q_i \in R^n$ je vektor takový, že $q_i^T (A_i - J_{i+1})d_i \neq 0$. Pak platí $A_{i+1}d_i = J_{i+1}d_i$ a $q_i^T A_{i+1} = q_i^T J_{i+1}$.

Oboustranné kvazinewtonovské metody mají důležitou vlastnost lineárního ukončení (naleznou řešení soustavy lineárních rovnic po nejvýše $n+1$ krocích).

Věta 253. *(Lineární ukončení) Nechť x_i, $i \in N$, je posloupnost generovaná oboustrannou kvazinewtonovskou metodou s jednotkovým výběrem délky kroku ($d_i = s_i$, $i \in N$) aplikovanou na soustavu lineárních rovnic $J(x - x^*) = 0$ s regulární maticí J. Nechť $f_i = J(x_i - x^*) \neq 0$, $1 \leq i \leq n+1$. Pak $f_{n+2} = J(x_{n+2} - x^*) = 0$ a $x_{n+2} = x^*$.*

Důkaz Předpokládejme, že $f_i \neq 0$, $1 \leq i \leq n+1$. Dokážeme indukcí, že pro $1 \leq i \leq n$ není vektor $d_i \neq 0$ lineární kombinací vektorů d_j, $1 \leq j < i$, a že pro $1 \leq j < i \leq n+1$ platí

$$(A_i - J)d_j = 0, \tag{1015}$$

$$q_j^T (A_i - J) = 0. \tag{1016}$$

Nechť $i = 1$. Jelikož $A_1 d_1 = A_1 s_1 = -f_1$, $f_1 \neq 0$ a matice A_1 je regulární, platí $d_1 \neq 0$ a dále není co dokazovat. Indukční krok:

(a) Nechť $1 < i \leq n$. Jelikož $A_i d_i = A_i s_i = -f_i$, $f_i \neq 0$ a matice A_i je regulární, platí $d_i \neq 0$. Jelikož

$$f_{i+1} = J(x_i + d_i - x^*) = f_i + Jd_i \neq 0,$$

musí platit

$$(A_i - J)d_i = A_i s_i + f_i - Jd_i - f_i = -(f_i + Jd_i) \neq 0,$$

takže vektor d_i nemůže být lineární kombinací vektorů d_j, $1 \leq j < i$, pro které platí (1015).

(b) Použijeme-li (1014), můžeme psát

$$A_{i+1} - J = A_i - J - \frac{(A_i - J)d_i q_i^T (A_i - J)}{q_i^T (A_i - J)d_i}. \tag{1017}$$

Z (1015) a (1017) plyne, že $(A_{i+1} - J)d_j = 0$ pro $1 \leq j < i$. Dále platí

$$(A_{i+1} - J)d_i = (A_i - J)d_i - (A_i - J)d_i = 0,$$

takže $(A_{i+1} - J)d_j = 0$ pro $1 \leq j \leq i$.

(c) Z (1016) a (1017) plyne, že $q_j^T (A_{i+1} - J) = 0$ pro $1 \leq j < i$. Dále platí

$$q_i^T (A_{i+1} - J)) = q_i^T (A_i - J) - q_i^T (A_i - J) = 0,$$

takže $(A_{i+1} - J)d_j = 0$ pro $1 \leq j \leq i$.

Tím je indukční krok dokončen. Jelikož vektory d_i, $1 \leq i \leq n$, jsou lineárně nezávislé a podle (1015) platí $(A_{i+1} - J)d_i = 0$, $1 \leq i \leq n$, můžeme psát $A_{i+1} = J$ a tudíž

$$f(x_{i+2}) = J(x_{i+2} - x^*) = J(x_{i+1} + d_{i+1} - x^*) = f_{i+1} + Jd_{i+1} = f_{i+1} + A_{i+1}s_{i+1} = 0.$$

□

Poznámka 370. Vlastnost (1015) (nebo (1016)) se nazývá dědičností. Kvazinewtonovské metody vyšetřované v oddílu 11.5 ani obecné sdružené kvazinewtonovské metody tuto vlastnost nemají.

Sdružené kvazinewtonovské metody vyhovují sdruženému minimalizačnímu principu.

Věta 254. *Nechť W je čtvercová regulární matice řádu n a $z = J_+^T q$. Pak matici A_+, která je řešením úlohy*

$$\|(A_+ - A)^T W^{-1}\|_F = \min_{\tilde{A}^T q = z} \|(\tilde{A} - A)^T W^{-1}\|_F \tag{1018}$$

lze vyjádřit ve tvaru (1012), kde $p = W^T W z$.

Důkaz Nahradíme-li v (999) matice A, $\tilde{A}$, A_+ jejich transposicemi a vektory $\tilde{A}d$, y výrazy $\tilde{A}^T q$, z, můžeme podle věty 247 (po dosazení do (978)) psát

$$A_+^T = A^T - \frac{(A - J_+)^T q\, p^T}{p^T q},$$

kde $p = W^T W z$, což transponováno dává (1012). □

Vzah (1012), definující obecnou sdruženou kvazinewtonovskou metodu obsahuje dva volitelné vektory. Vektor p_i se nejčastěji volí tak, že $p_i = q_i$ (pak pokud $q_i \neq 0$ je výraz ve jmenovateli nenulový), nebo $p_i = z_i = J_{i+1}^T q_i$ (což odpovídá volbě $W = I$ v (1018)), nebo $p_i = (A_i - J_{i+1})d_i$ (pak dostaneme oboustrannou kvazinewtonovskou metodu (1014)). Nejznámější sdružené kvazinewtonovské metody dostaneme, položíme-li

$$q_i = (A_i - J_{i+1})d_i \tag{1019}$$

(tečná sdružená kvazinewtonovská metoda), nebo

$$q_i = A_i d_i - (f_{i+1} - f_i) \tag{1020}$$

(sečná sdružená kvazinewtonovská metoda), nebo

$$q_i = f_{i+1} \tag{1021}$$

(reziduální sdružená kvazinewtonovská metoda). Reziduální sdružená kvazinewtonovská metoda je uvedená v poznámce 368. Tato metoda je ekvivalentní sečné sdružené kvazinewtonovské metodě, pokud $A_i d_i = f_i$ (řešíme-li přesné soustavu lineárních rovnic a používáme-li jednotkovou délku kroku). Všechny sdružené kvazinewtonovské metody používají vektor $J_{i+1}^T q_i$, který lze spočítat pomocí zpětného automatického derivování (příklad 12 v oddílu 14.3). Tečná sdružená kvazinewtonovská metoda a oboustranná kvazinewtonovská metoda používají navíc vektor $J_{i+1} d_i$, který lze spočítat pomocí přímého automatického derivování (příklad 10 v oddílu 14.3), nebo numericky pomocí diferencí, a lze ho poměrně dobře aproximovat vektorem $y_i = f_{i+1} - f_i$.

Používáme-li reziduální sdruženou kvazinewtonovskou metodu, platí $J_{i+1}^T q_i = J_{i+1}^T f_{i+1} = g_{i+1}$, takže (1011) s $p_i = q_i = f_{i+1}$ lze zapsat ve tvaru

$$A_{i+1} = A_i - \frac{f_{i+1}(h_{i+1} - g_{i+1})^T}{f_{i+1}^T f_{i+1}}, \tag{1022}$$

kde $h_{i+1} = A_i^T f_{i+1}$. Aktualizaci reziduální oboustranné kvazinewtonovské metody (1013) lze aproximovat výrazem

$$A_{i+1} = A_i - \frac{(A_i d_i - y_i)(h_{i+1} - g_{i+1})^T}{(h_{i+1} - g_{i+1})^T d_i} \tag{1023}$$

(směrovou derivaci $J_{i+1} d_i$ nahrazujeme vektorem y_i). Metoda (1023) není oboustrannou kvazinewtonovskou metodou, neboť $y_i \neq J_{i+1} d_i$, ale její vlastnosti se podobají vlastnostem metody (1013), neboť $y_i \approx J_{i+1} d_i$. Poznamenejme, že metoda (1023) je kvazinewtonovskou metodou tvaru (978), kde $v_i = h_{i+1} - g_{i+1}$, takže se na ní vztahují tvrzení uvedená v oddílu 11.5. Změníme-li v (1023) jmenovatel tak, že

$$A_{i+1} = A_i - \frac{(A_i d_i - y_i)(h_{i+1} - g_{i+1})^T}{f_{i+1}^T (A_i d_i - y_i)}, \tag{1024}$$

dostaneme další aproximaci reziduální oboustranné kvazinewtonovské metody (1013). Pro tuto metodu platí $A_{i+1}^T f_{i+1} = J_{i+1}^T f_{i+1} = g_{i+1}$. Metody (1022)–(1024) vyžadují výpočet gradientu $g_{i+1} = J_{i+1}^T f_{i+1}$, ale není třeba ukládat Jacobiovu matici J_{i+1} ani počítat vektor $J_{i+1}d_i$.

Další užitečnou vlastností sdružených i oboustranných kvazinewtonovských metod je jejich invariantnost vzhledem k lineární transformaci proměnných.

Věta 255. *Nechť $\tilde{f}(\tilde{x}) = f(T^{-1}x)$, kde T je regulární čtvercová matice. Nechť $\tilde{x}_i \in R^n$, $i \in N$, je posloupnost generovaná sdruženou nebo oboustrannou kvazinewtonovskou metodou s počáteční maticí $\tilde{A}_1$ aplikovanou na soustavu rovnic $\tilde{f}(\tilde{x}) = 0$ a $x_i \in R^n$, $i \in N$, je posloupnost generovaná toutéž sdruženou nebo oboustrannou kvazinewtonovskou metodou aplikovanou na soustavu rovnic $f(x) = 0$. Pak pokud používáme stejný výběr délky kroku ($\tilde{\alpha}_i = \alpha_i, i \in N$) a pokud $A_1 = \tilde{A}_1 T^{-1}$, platí $x_i = T\tilde{x}_i$.*

Důkaz Snadno se dokáže (derivováním složeného zobrazení $\tilde{f}(\tilde{x}) = f(T^{-1}x)$), že platí $\tilde{J}(\tilde{x}) = J(x)T$. Ukážeme, že $A_i = \tilde{A}_i T^{-1}$ $\forall i \in N$ (podle předpokladu to platí pro $i = 1$). Pak

$$x_{i+1} = x_i - \alpha_i A_i^{-1} f_i = T\tilde{x}_i - \alpha_i T\tilde{A}_i^{-1} f_i = T(\tilde{x}_i - \alpha_i \tilde{A}_i^{-1}\tilde{f}_i) = T\tilde{x}_{i+1}.$$

Důkaz provedeme indukcí. Předpokládejme, že $A = \tilde{A}T^{-1}$ (platí to v první iteraci). Použijeme-li vztah $J_+ = \tilde{J}_+ T^{-1}$ a (1012), dostaneme

$$A_+ = A - \frac{pq^T(A - J_+)}{q^T p} = \tilde{A}T^{-1} - \frac{pq^T(\tilde{A} - \tilde{J}_+)T^{-1}}{q^T p} = \tilde{A}_+ T^{-1}.$$

Použijeme-li vztah $J_+ = \tilde{J}_+ T^{-1}$ a (1014), dostaneme

$$A_+ = A - \frac{(A - J_+)dq^T(A - J_+)}{q^T(A - J_+)d} = \tilde{A}T^{-1} - \frac{(\tilde{A} - \tilde{J}_+)T^{-1}T\tilde{d}q^T(\tilde{A} - \tilde{J}_+)T^{-1}}{q^T(\tilde{A} - \tilde{J}_+)T^{-1}T\tilde{d}} = \tilde{A}_+ T^{-1},$$

neboť $d = x_+ - x = T(\tilde{x}_+ - \tilde{x}) = T\tilde{d}$. □

Poznámka 371. Broydenova dobrá metoda uvedená v oddílu 11.5 ani metoda uvedená v poznámce 367 nejsou invariantní vzhledem k lineární transformaci proměnných.

Poznámka 372. Vynikající vlastností reziduální sdružené kvazinewtonovské metody (1011) je snadné určení přesného gradientu funkce $F = (1/2)\|f\|^2$. Podle poznámky 368 platí $h_i = A_i^T f_i = J_i^T f_i = g_i$ $\forall i \in N$, takže je splněn předpoklad A3b s $\overline{\vartheta} = 0$. V důsledku toho platí tvrzení lemmatu 99 s $\lambda = 0$. K tomu, aby reziduální sdružená Broydenova metoda byla globálně konvergentní tedy stačí, aby platilo $\|s_i\| \leq \overline{c}\|f_i\|$ $\forall i \in N$ (posnámka 350). To lze snadno zařídit jednoduchou úpravou algoritmu. Určíme směrový vektor s_i tak, aby platilo $A_i s_i + f_i = 0$. Pokud $\|s_i\| \leq \overline{c}\|f_i\|$, kde $\overline{c}$ je vhodně zvolená (velká) konstanta, určíme délku kroku $\alpha_i > 0$ tak, aby byla splněna některá z podmínek ($\overline{\text{S2a}}$)–($\overline{\text{S2c}}$) a položíme $x_{i+1} = x_i + \alpha_i s_i$ a $A_{i+1} = A_i - p_i f_{i+1}^T (A_i - J_{i+1})/f_{i+1}^T p_i$, kde p_i je vhodně zvolený vektor (například $p_i = f_{i+1}$). V opačném případě (pokud $\|s_i\| > \overline{c}\|f_i\|$), položíme $x_{i+1} = x_i$ a $A_{i+1} = J_{i+1}$.

Výsledkem úvah uvedených v této poznámce je následující věta.

Věta 256. *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům J1, J4, J5a a J6. Pak reziduální sdružená kvazinewtonovská metoda popsaná v poznámce 372 je globálně konvergentní.*

Ostatní sdružené kvazinewtonovské metody tuto výjimečnou vlastnost nemají. Proto je musíme realizovat pomocí algoritmu 27. Tyto úpravy mají své opodstatnění, neboť platí tato věta.

Věta 257. *Nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ taková, že pokud $\|x_1 - x^*\| \leq \overline{\delta}$ a $\|A_1 - J_1\| \leq \overline{\vartheta}$, posloupnost x_i, $i \in N$, určená sdruženou kvazinewtonovskou metodou (1019), (1020) (nebo (1021), když $\overline{\omega} = 0$) s jednotkovým výběrem délky kroku ($\alpha_i = 1$, $i \in N$) konverguje Q-superlineárně k bodu $x^* \in R^n$.*

Důkaz Věta 249 je speciálním důsledkem obecnějších vět 270 a 271 dokázáných v oddílu 12.6.

11.8 Tenzorové metody

V oddílu 8.2 jsme ukázali, jak lze pomocí aproximací B_k, $1 \leq k \leq m$, Hessových matic G_k, $1 \leq k \leq m$, urychlit konvergenci Gaussovy-Newtonovy metody. Místo lineárního modelu $l(x+s) = f(x) + J(x)s$, na kterém je založena Gaussova-Newtonova metoda, jsme použili kvadratický model

$$q(x+s) = f(x) + J(x)s + \frac{1}{2}s^T T s,$$

kde T je třírozměrná veličina (tenzor) mající prvky

$$T_{kij} = (B_k)_{ij} \approx (G_k(x))_{ij} = \frac{\partial^2 f_k(x)}{\partial x_i \partial x_j}$$

(v oddílu 8.2 jsme tenzor T nazaváděli, pracovali jsme pouze s maticemi B_k, $1 \leq k \leq m$). Podobný postup můžeme použít i v případě řešení soustav nelineárních rovnic. Místo soustavy lineárních rovnic $l(x+s) = f(x) + J(x)s = 0$, která definuje směrový vektor Newtonovy metody, můžeme řešit soustavu kvadratických rovnic $q(x+s) = f(x) + J(x)s + (1/2)s^T T s = 0$. Potíž je v tom, že tato soustava má stejnou dimenzi jako původní soustava, je také nelineární (kvadratická) a navíc nemusí mít řešení. Tuto potíž lze odstranit, pracujeme-li s maticemi získanými interpolací omezeného počtu funkčních hodnot (tyto matice nají omezenou hodnost).

V dalším výkladu budeme předpokládat, že matice $B_k^i \approx G_k(x_i)$, $1 \leq k \leq m$, splňují p interpolačních podmínek

$$f_k(x_i + d_j^i) = f_k(x_i) + g_k^T(x_i)d_j^i + \frac{1}{2}(d_j^i)^T B_k^i d_j^i, \qquad 1 \leq j \leq p,$$

kde $d_j^i = x_{i-j} - x_i$, $1 \leq j \leq p$. Abychom zjednodušili značení, budeme často index i vynechávat. V následující větě (a jejím důkazu) vynecháme i index k, takže symboly f a g budou označovat hodnotu a gradient funkce f_k.

Věta 258. *Matice B má minimální Frobeniovu normu na množině zadané interpolačními rovnostmi*

$$z_j \stackrel{\Delta}{=} 2(f(x+d_j) - f(x) - g^T(x)d_j) = d_j^T B d_j, \qquad 1 \leq j \leq p$$

právě tehdy, když

$$B = \sum_{j=1}^{p} u_j d_j d_j^T,$$

kde $u = M^{-1}z$. Přitom $z = [z_1, \ldots, z_p]^T$, $u = [u_1, \ldots, u_p]^T$ a

$$M = \begin{bmatrix} (d_1^T d_1)^2, & \ldots, & (d_1^T d_p)^2 \\ \ldots, & \ldots, & \ldots \\ (d_p^T d_1)^2, & \ldots, & (d_p^T d_p)^2 \end{bmatrix}.$$

Důkaz Nutnost dokážeme pomocí Lagrangeovy funkce

$$L(B,u) = \frac{1}{2}\sum_{k=1}^{n}\sum_{l=1}^{n} B_{kl}^2 + \sum_{i=1}^{p} u_i \left(z_i - \sum_{k=1}^{n}\sum_{l=1}^{n} d_i^T e_k B_{kl} e_l^T d_i \right).$$

Postačitelnost plyne z konvexity Frobeniovy normy. Derivujeme-li Lagrangeovu funkci podle proměnné B_{kl}, dostaneme

$$\frac{\partial L(B,u)}{\partial B_{kl}} = B_{kl} - \sum_{i=1}^{p} u_i d_i^T e_k e_l^T d_i,$$

kde e_k a e_l jsou odpovídající sloupce jednotkové matice. Nutné podmínky pro extrém mají tedy tvar

$$B_{kl} = \sum_{i=1}^{p} u_i d_i^T e_k e_l^T d_i, \qquad 1 \le k \le n, \quad 1 \le l \le n.$$

Dosadíme-li toto vyjádření do interpolačních rovností, dostaneme

$$\begin{aligned} z_j &= \sum_{k=1}^{n}\sum_{l=1}^{n} d_j^T e_k B_{kl} e_l^T d_j = \sum_{i=1}^{p} u_i \sum_{k=1}^{n}\sum_{l=1}^{n} d_j^T e_k d_i^T e_k e_l^T d_i e_l^T d_j \\ &= \sum_{i=1}^{p} u_i \left(\sum_{k=1}^{n} d_j^T e_k d_i^T e_k\right)\left(\sum_{l=1}^{n} e_l^T d_i e_l^T d_j\right) = \sum_{i=1}^{p} u_i (d_j^T d_i)^2. \end{aligned}$$

Nechť $z = [z_1, \ldots, z_p]^T$, $u = [u_1, \ldots, u_p]^T$ a M je matice uvedená ve větě 258. Pak lze předchozí soustavu rovnic zapsat ve tvaru $Mu = z$ takže $u = M^{-1}z$. □

Poznámka 373. Vzorce uvedené ve větě 258 můžeme zapsat v tenzorovém tvaru. Nechť $Z \in R^{n\times p}$ je matice jejímiž sloupci jsou vektory $2(f(x + d_l) - f(x) - J(x)d_l)$, $1 \le l \le p$, a $U \in R^{n\times p}$ je matice taková, že $U = ZM^{-1}$. Pak platí

$$T = \sum_{l=1}^{p} (Ue_l) \times d_l \times d_l, \tag{1025}$$

kde $\times$ značí tenzorový součin, takže

$$T_{kij} = \sum_{l=1}^{p} e_k^T U e_l e_i^T d_l e_j^T d_l, \qquad 1 \le k \le n, \quad 1 \le i \le n, \quad 1 \le j \le n.$$

Poznámka 374. Použijeme-li vzorec (1025) (s indexem j místo l), můžeme soustavu kvadratických rovnic, sloužících k určení směrového vektoru s, zapsat ve tvaru

$$q(x + s) = f(x) + J(x)s + \frac{1}{2}\sum_{j=1}^{p} Ue_j (d_j^T s)^2 = 0. \tag{1026}$$

Předpokládejme, že matice J je regulární a označme $\beta_j = d_j^T s$, $1 \le j \le p$. Pak platí

$$s = -J^{-1}\left(f - \frac{1}{2}\sum_{j=1}^{p} Ue_j \beta_j^2\right). \tag{1027}$$

Vynásobíme-li tento vztah postupně vektory d_i, $1 \le i \le p$, dostaneme soustavu kvadratických rovnic

$$w_i^T f + \beta_i + \frac{1}{2}\sum_{j=1}^{p} w_i^T U e_j \beta_j^2 = 0, \qquad 1 \le i \le p \tag{1028}$$

(s neznámými β_j, $1 \le j \le p$), kde $w_i = (J^{-1})^T d_i$, $1 \le i \le p$. Jelikož obvykle $p \ll n$, je řešení této soustavy mnohem snažší než řešení původní soustavy nelineárních rovnic.

V poznámce 374 předpokládáme, že matice J je regulární a soustava rovnic (1028) má řešení. V tomto případě má i soustava (1026) řešení, které dostaneme, dosadíme-li β_j, $1 \le j \le p$, do (1027). Nemá-li soustava (1026) řešení, určíme vektor s tak, aby minimalizoval normu $\|q(x+s)\|$. Předpokládejme nejprve, že matice J je regulární a označme $D \in R^{n\times p}$ matici, jejímiž sloupci jsou vektory d_i, $1 \le i \le p$, a $W \in R^{n\times p}$

matici, jejímiž sloupci jsou vektory $w_i = (J^{-1})^T d_i$, $1 \leq\ i \leq p$ (takže $W = (J^{-1})^T D$ a $W^T J s = D^T s = \beta$). Pak můžeme soustavu (1028) zapsat ve tvaru

$$\tilde{q}(\beta) = W^T q(x+s) = W^T f + \beta + \frac{1}{2} W^T U \beta^2 = 0, \tag{1029}$$

kde $\beta = [\beta_1, \ldots, \beta_p]$ a $\beta^2 = [\beta_1^2, \ldots, \beta_p^2]$.

Věta 259. *Nechť matice J je regulární a nechť vektor $\beta \in R^p$ minimalizuje normu $\|(W^T W)^{-1/2} \tilde{q}(\beta)\|$. Pak vektor*

$$s = -J^{-1} \left(f + \frac{1}{2} U \beta^2 - W (W^T W)^{-1} \tilde{q}(\beta) \right) \tag{1030}$$

minimalizuje normu $\|q(x+s)\|$.

Důkaz Jelikož pro libovolnou ortogonální matici Q, platí $\|Qq(x+s)\| = \|q(x+s)\|$, budeme minimalizovat normu $\|Qq(x+s)\|$, kde $Q = [V, Z]^T$ a kde V, Z jsou matice s ortonormálními sloupci takové, že

$$V = W(W^T W)^{-1/2}, \qquad W^T Z = D^T J^{-1} Z = 0$$

(vynásobením se snadno přesvědčíme, že $V^T V = I$). Použijeme-li (1029), dostaneme

$$Qq(x+s) = \begin{bmatrix} V^T q(x+s) \\ Z^T q(x+s) \end{bmatrix} = \begin{bmatrix} (W^T W)^{-1/2} W^T q(x+s) \\ Z^T q(x+s) \end{bmatrix} = \begin{bmatrix} (W^T W)^{-1/2} \tilde{q}(\beta) \\ Z^T q(x+s) \end{bmatrix}.$$

Z tohoto vyjádření je patrné, že pokud vektor β minimalizuje normu $\|(W^T W)^{-1/2} \tilde{q}(\beta)\|$ a pokud s je vektor takový, že $D^T s = \beta$ a $Z^T q(x+s) = 0$, minimalizuje tento vektor normu $\|Qq(x+s)\| = \|q(x+s)\|$. Položme

$$s = J^{-1} W (W^T W)^{-1} \beta + J^{-1} Z r, \tag{1031}$$

kde $r \in R^{n-p}$. Pak platí

$$D^T s = D^T J^{-1} W (W^T W)^{-1} \beta + D^T J^{-1} Z r = W^T W (W^T W)^{-1} \beta + W^T Z r = \beta,$$

a

$$\begin{aligned} Z^T q(x+s) &= Z^T \left(f + Js + \frac{1}{2} U \beta^2 \right) = Z^T f + Z^T W (W^T W)^{-1} \beta + Z^T Z r + \frac{1}{2} Z^T U \beta^2 \\ &= Z^T f + r + \frac{1}{2} Z^T U \beta^2, \end{aligned}$$

takže $Z^T q(x+s) = 0$, pokud $r = -Z^T (f + (1/2) U \beta^2)$. Z ortogonality matice Q plyne, že $VV^T + ZZ^T = I$, takže $ZZ^T = I - VV^T = I - W(W^T W)^{-1} W^T$ a tedy

$$Zr = -ZZ^T \left(f + \frac{1}{2} U \beta^2 \right) = -\left(f + \frac{1}{2} U \beta^2 \right) + W (W^T W)^{-1} W^T \left(f + \frac{1}{2} U \beta^2 \right)$$

Dosadíme-li tento vektor do (1031) a použijeme-li (1029), dostaneme

$$\begin{aligned} s &= J^{-1} W (W^T W)^{-1} \beta - J^{-1} \left(f + \frac{1}{2} U \beta^2 \right) + J^{-1} W (W^T W)^{-1} W^T \left(f + \frac{1}{2} U \beta^2 \right) \\ &= -J^{-1} \left(f + \frac{1}{2} U \beta^2 + W (W^T W)^{-1} \tilde{q}(\beta) \right). \end{aligned}$$

□

Poznámka 375. Víme-li, že existuje vektor s takový, že $Z^T q(x+s) = 0$, můžeme vzorec (1030) odvodit jednodušším způsobem. Jelikož

$$q(x+s) = Q^T Q q(x+s) = [V, Z] \begin{bmatrix} (W^T W)^{-1/2} \tilde{q}(\beta) \\ 0 \end{bmatrix} = V(W^T W)^{-1/2} \tilde{q}(\beta) = W(W^T W)^{-1} \tilde{q}(\beta),$$

platí

$$f(x) + J(x)s + \frac{1}{2} U \beta^2 = W(W^T W)^{-1} \tilde{q}(\beta),$$

odkud bezprostředně plyne (1030).

Poznámka 376. Nechť $W^T W = R^T R$, kde R je horní trojúhelníková matice. Pak minimalizace normy $\|(W^T W)^{-1/2} \tilde{q}(\beta)\|$ je ekvivalentní minimalizaci funkce

$$\frac{1}{2} \|(W^T W)^{-1/2} \tilde{q}(\beta)\|^2 = \frac{1}{2} \tilde{q}^T(\beta)(W^T W)^{-1} \tilde{q}(\beta) = \frac{1}{2} \tilde{q}^T(\beta)(R^T R)^{-1} \tilde{q}(\beta),$$

což je vážený součet čtverců, jehož minimum lze nalézt metodami popsanými v kapitole 8. Zdůrazněme, že vektor proměnných β má dimenzi $p \ll n$.

Poznámka 377. Je-li matice J singulární, řešíme příslušné soustavy lineární rovnic ve smyslu nejmenších čtverců, takže místo matice J^{-1} používáme pseudoinverzi $J^\dagger$ (věta 164). Pak můžeme psát $W = (J^\dagger)^T D$ a $s = -J^\dagger \left(f + (1/2) U \beta^2 - W(W^T W)^{-1} \tilde{q}(\beta) \right)$.

Poznámka 378. Zbývá ukázat, jak se volí číslo p (počet interpolačních podmínek). Pro husté úlohy menšího rozměru je výhodné, aby platilo $p \leq \sqrt{n}$. V tomto případě trojúhelníkový rozklad matice J spotřebuje zhruba $(1/3)n^3$ operací, určení vektorů w_i, $1 \leq i \leq p$, a s zhruba $n^2(p+1) \leq n^2(\sqrt{n}+1)$ operací a Choleského rozklad matice $W^T W$ zhruba $p^3 = n\sqrt{n}$ operací, takže pro $n \geq 10$ již operace spotřebované na trojúhelníkový rozklad matice J dominují. V případě rozsáhlých úloh, kdy je třeba řešit soustavy rovnic s maticí J iteračně, je třeba aby číslo p bylo co nejmenší. Ukazuje se, že vliv tenzorového členu se úspěšně projevuje již pro $p = 1$.

Zaměříme se nyní na případ, kdy $p = 1$. V tomto případě budeme používat označení $D = [d_1] = d$, $W = [w_1] = w = (J^{-1})^T d$, $Z = [z_1] = z = 2(f(x+d) - F(x) - J^T(x)d$ a $U = [u_1] = u = z/(d^T d)^2$, takže

$$\tilde{q}(\beta) = w^T f + \beta + \frac{1}{2} w^T u \beta^2. \tag{1032}$$

Věta 260. *Nechť $p = 1$. Pak pro směrový vektor určený tenzorovou metodou platí*

$$\begin{aligned} s &= -J^{-1} \left(f + \frac{1}{2} u \beta^2 \right), & w^T u w^T f \leq 1, \\ s &= -J^{-1} \left(f + \frac{1}{2} u \beta^2 - \frac{w}{w^T w} \tilde{q}(\beta) \right), & w^T u w^T f > 1, \end{aligned}$$

kde

$$\begin{aligned} \beta &= \frac{w^T f}{1 + \sqrt{1 - w^T u w^T f}}, & w^T u w^T f \leq 1, \\ \beta &= -\frac{1}{w^T u}, & w^T u w^T f > 1. \end{aligned}$$

Důkaz Vzorce pro směrový vektor plynou bezprostředně z (1030). Podle (1032) je rovnice $\tilde{q}(\beta) = 0$ kvadratickou rovnicí, která má řešení pokud její diskriminant $1 - w^T u w^T f$ je nezáporný. V tomto případě platí

$$\beta = \frac{-1 \pm \sqrt{1 - w^T u w^T f}}{w^T u} = \frac{w^T f}{1 \pm \sqrt{1 - w^T u w^T f}}.$$

(znaménko bereme tak, aby číslo β bylo v absolutní hodnotě co nejmenší). Pokud $1 - w^T u w^T f < 0$, rovnice $\tilde{q}(\beta) = 0$ nemá řešení a číslo β musíme určit minimalizací funkce $\tilde{q}(\beta)(w^T w)^{-1}\tilde{q}(\beta)$, nebo (což je ve skalárním případě totéž) funkce $\tilde{q}^2(\beta)$. Ale $(\tilde{q}^2(\beta))' = 2\tilde{q}(\beta)\tilde{q}'(\beta) = 0$ pokud buď $\tilde{q}(\beta) = 0$ (což jsme vyloučili) nebo $\tilde{q}'(\beta) = 1 + w^T u\beta = 0$, což dává $\beta = -1/w^T u$. □

Abychom dostali účinný algoritmus odolný vůči selhání, je třeba tenzorovou metodu kombinovat s Newtonovou metodou. Jedna z možností je použita v následujícím algoritmu.

Algoritmus 28. Data $0 < \underline{\rho} < 1$, $0 < \underline{\beta} \le \overline{\beta} < 1$, $\overline{\varepsilon} > 0$, $0 < \underline{k} \le \overline{k}$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$, vypočteme $f_1 = f(x_1)$ a položíme $i = 1$.

Krok 2 Pokud $\|f_i\| \le \overline{\varepsilon}$, ukončíme výpočet.

Krok 3 Vypočteme Jacobiovu matici $J_i = J(x_i)$. Určíme Newtonův směr $s_i^N = -J_i f_i$ a tenzorový směr $s_i^S = s_i^N - J^{-1}(f + (1/2)U\beta^2 - W(W^T W)^{-1}\tilde{q}(\beta))$, kde β je vektor, který minimalizuje normu $\|(W^T W)^{-1/2}\tilde{q}(\beta))\|$ (věta 259).

Krok 4 Položíme $x_{i+1} = x_i + s_i^S$ a vypočteme $f_{i+1} = f(x_{i+1})$. Jestliže $\|f(x_{i+1})\| < \|f_i\|$ přejdeme na krok 7.

Krok 5 Jestliže $f_i^T J_i s_i^S < 0$, položíme $s_i = s_i^S$. V opačném případě položíme $s_i = s_i^N$.

Krok 6a Položíme $\alpha_i^1 = 1$ a $k = 1$.

Krok 6b Položíme $x_{i+1} = x_i + \alpha_i^k s_i$ a vypočteme $f_{i+1} = f(x_{i+1})$ (pokud $s_i = s_i^S$ a $k = 1$, použijeme hodnotu f_{i+1} z kroku 4). Je-li splněna některá (vybraná) podmínka z ($\overline{\text{S2}}$), přejdeme na krok 7.

Krok 6c Pokud $s_i = s_i^N$ a $j > \overline{k}$, ukončíme výpočet (předčasné ukončení způsobené selháním Newtonovy metody). Pokud $s_i = s_i^S$ a $j > \underline{k}$, položíme $s_i = s_i^N$ a přejdeme na krok 6a. V ostatních případech určíme délku kroku α_i^{k+1} tak aby platilo $\underline{\beta}\alpha_i^k \le \alpha_i^{k+1} \le \overline{\beta}\alpha_i^k$, položíme $k := k + 1$ a přejdene na krok 6b.

Krok 7 Položíme $i := i + 1$ a přejdeme na krok 2.

11.9 Aktualizace ortogonálního rozkladu

Používáme-li kvazinewtonovské metody (978), je třeba určovat směrový vektor řešením soustavy rovnic $As + f = 0$, kde A je regulární čtvercová matice. V tomto případě je výhodné pracovat s ortogonálním rozkladem $A = QR$, kde Q je ortogonální čtvercová matice a R je horní trojúhelníková matice (ortogonální rozklad lze určit podle poznámky 275). Pak řešení soustavy rovnic $QRs + f = 0$ vyžaduje $O(n^2)$ aritmetických operací. Ukážeme nyní, jak lze určit ortogonální rozklad matice $\bar{A} = A + uv^T$ z ortogonálního rozkladu matice A s použitím $O(n^2)$ aritmetických operací.

Věta 261. *Nechť $\bar{A} = A + uv^T$, kde $A = QR$. Nechť $\tilde{u} = Q^T u$ a $\tilde{Q}^T$ je ortogonální matice (součin Givensových matic elementárních rotací) taková, že $\tilde{Q}^T\tilde{u} = \|\tilde{u}\|e_1$, přičemž matice $\tilde{R} = \tilde{Q}R$ je horní Hessenbergova. Nechť $\hat{Q}^T$ je ortogonální matice (součin Givensových matic elementárních rotací) taková, že matice $\bar{R} = \hat{Q}^T(\tilde{R} + \|\tilde{u}\|e_1 v^T)$ je horní trojúhelníková. Pak platí $\bar{A} = \bar{Q}\bar{R}$, kde $\bar{Q} = Q\tilde{Q}\hat{Q}$.*

Důkaz Jelikož matice Q je ortogonální, můžeme psát

$$A + uv^T = Q(R + Q^T uv^T) = Q(R + \tilde{u}v^T). \tag{1033}$$

Podle poznámky 271 existuje ortogonální matice $\tilde{Q}^T = \tilde{Q}_{12}^T\tilde{Q}_{23}^T \dots \tilde{Q}_{n-1,n}^T$ taková, že $\tilde{Q}^T\tilde{u} = \|\tilde{u}\|e_1$. Z konstrukce této matice plyne, že matice $\tilde{R} = \tilde{Q}^T R$ je horní Hessenbergova. Pak také matice $\tilde{R} + \|\tilde{u}\|e_1 v^T$ je horní Hessenbergova. Podle poznámky 272 existuje ortogonální matice $\hat{Q}^T = \hat{Q}_{n-1,n}^T \dots \hat{Q}_{23}^T\hat{Q}_{12}^T$ taková, že matice $\bar{R} = \hat{Q}^T(\tilde{R} + \|\tilde{u}\|e_1 v^T)$ je horní trojúhelníková. Po dosazení do (1033) dostaneme

$$\begin{aligned} \bar{A} &= A + uv^T = Q(R + \tilde{u}v^T) = Q\tilde{Q}(\tilde{Q}^T R + \tilde{Q}^T\tilde{u}v^T) = Q\tilde{Q}(\tilde{R} + \|\tilde{u}\|e_1 v^T) \\ &= Q\tilde{Q}\hat{Q}\hat{Q}^T(\tilde{R} + \|\tilde{u}\|e_1 v^T) = Q\tilde{Q}\hat{Q}\bar{R} = \bar{Q}\bar{R}. \end{aligned}$$

□

11.10 Numerické porovnání

K testování metod pro řešení systémů nelineárních rovnic bylo použito 62 úloh, obsahujících 200 rovnic o 200 neznámých, ze sbírky TEST37 zmíněné v oddílu 1.5. V tabulce 19 jsou uvedeny výsledky získané těmito metodami:

`TRNM-DER` - Newtonova metoda s Jacobiovou maticí počítanou analyticky,
`TRNM-DIF` - Newtonova metoda s Jacobiovou maticí počítanou numericky,
`TRGB-DER` - Broydenova dobrá metoda (979) s Jacobiovou maticí počítanou analyticky,
`TRGB-DIF` - Broydenova dobrá metoda (979) s Jacobiovou maticí počítanou numericky,
`TRT1-DER` - Optimálně podmíněná Toddova metoda (989) s výběrem parametru ϑ podle věty 244 s Jacobiovou maticí počítanou analyticky,
`TRT1-DIF` - Optimálně podmíněná Toddova metoda (989) s výběrem parametru ϑ podle věty 244 s Jacobiovou maticí počítanou numericky,
`TRT2-DER` - Optimálně podmíněná Toddova metoda (989) s výběrem parametru ϑ podle věty 245 s Jacobiovou maticí počítanou analyticky,
`TRT2-DIF` - Optimálně podmíněná Toddova metoda (989) s výběrem parametru ϑ podle věty 245 s Jacobiovou maticí počítanou numericky,
`TRRA-DER` - Reziduální sdružená kvazinewtonovská metoda (1022),
`TRR1-DER` - Reziduální oboustranná kvazinewtonovské metoda (1013),
`TRR2-DER` - Aproximace (1023) reziduální oboustranné kvazinewtonovské metody (1013),
`TRR3-DER` - Aproximace (1024) reziduální oboustranné kvazinewtonovské metody (1013).

Všechny uvedené metody jsou metodami s lokálně omezeným krokem (metody spádových směrů dosahovaly horších výsledků). Newtonova metoda používá k řešení lineárních rovnic LU rozklad nesymetrické matice a kvazinewtonovské metody používají aktualizaci QR rozkladu popsanou v oddílu 11.9.

Metoda	NIT	NFV	NFJ	NDC	selhání	čas
TRNM-DER	1516	1701	1516	1349	-	7.00
TRNM-DIF	1217	244354	-	1082	-	13.44
TRGB-DER	3129	3514	290	286	-	7.14
TRGB-DIF	2612	46002	-	213	-	7.92
TRT1-DER	3077	3555	278	273	1	7.37
TRT1-DIF	2998	53959	-	252	1	9.25
TRT2-DER	2625	3048	294	289	-	6.73
TRT2-DIF	2214	48735	-	228	-	7.72
TRRA-DER	4439	4758	5034	336	-	8.84
TRR1-DER	1736	1923	2049	186	-	4.68
TRR2-DER	1641	1819	1925	166	-	4.39
TRR3-DER	2257	2417	2522	165	-	5.19

Tabulka 14: TEST37 – 62 úloh

Tabulka 14 obsahuje celkový počet iterací `NIT`, celkový počet použitých funkčních hodnot `NFV`, celkový počet použitých Jacobiových matic (nebo gradientů funkce (949)) `NFJ`, celkový počet maticových rozkladů `NDC`, počet selhání a celkový čas výpočtu. K selhání došlo, když bylo nalezeno lokální minimum funkce (949)), které nebylo řešením soustavy rovnic.

Z výsledků uvedených v této tabulce lze vyvodit několik závěrů:

- Jsou-li prvky Jacobiovy matice zadány analyticky, je Newtonova metoda velmi efektivní a vyžaduje nejmenší počet funkčních hodnot a gradientů. V každém iteračním kroku je třeba řešit soustavu lineárních rovnic o n neznámých, což vyžaduje $O(n^3)$ aritmetických operací a pro velká n to může zpomalovat výpočet.
- Diferenční verze Newtonovy metody je pro husté systémy rovnic neefektivní, neboť numerický výpočet Jacobiovy matice v každém iteračním kroku vyžaduje vyčíslení velkého počtu funkčních hodnot.
- Broydenova dobrá metoda s analytickým výpočtem Jacobiovy matice po přerušení iteračního procesu konverguje pomaleji než Newtonova metoda, takže potřebuje větší počet iterací k nalezení řešení. Jelikož se používají aktualizace QR rozkladu (vyžadující pouze $O(n^2)$ aritmetických operací) je účinnost tato metody srovnatelná s účinností Newtonovy metody.
- Broydenova dobrá metoda s numerickým výpočtem Jacobiovy matice po přerušení iteračního procesu je mnohem efektivnější než diferenční verze Newtonovy metody, neboť Jacobiova matice se počítá jen ve zhruba desetině iteračních kroků.
- Optimálně podmíněná Toddova metoda s výběrem parametru ϑ podle věty 245 je mírně lepší než Broydenova dobrá metoda, která jinak patří k nejefektivnějším kvazinewtonovským metodám.
- Reziduální sdružené kvazinewtonovské metody vyžadují analytické zadání výrazů pro první derivace (nebo použití automatického derivování). Reziduální oboustranná kvazinewtonovská metoda a její aproximace konvergují téměř stejně rychle jako Newtonova metoda. Používají však aktualizace QR rozkladu, takže čas potřebný k nalezení řešení je nejnižší ve srovnání s ostatními testovanými metodami.

12 Metody pro rozsáhlé soustavy nelineárních rovnic

Rozsáhlé systémy nelineárních rovnic nemůžeme řešit metodami, které vyžadují uchovávání velkých hustých matic. Pro řešení takových systémů se používají metody, které jsou založeny na podobných myšlenkách jako optimalizační metody popsané v kapitolách 9 a v kapitole 10.

12.1 Kvazinewtonovské metody s omezenou pamětí

Kvazinewtonovské metody s omezenou pamětí používají omezený počet kroků kvazinewtonovských metod popsaných v oddílu 11.5. Při jejich popisu budeme používat množinu

$$M = \{l \in N : \ l = (j-1)m + 1, \ j \in N\}, \tag{1034}$$

kde $m \in N$, vzorec (995) a standardní označení $d_i = x_{i+1} - x_i$, $y_i = f_{i+1} - f_i$, $i \in N$.

Definice 86. *Řekneme, že základní metoda pro řešení soustav nelineárních rovnic je přímou m-krokovou kvazinewtonovskou metodou s omezenou pamětí, jestliže* $s_i = -S_i f_i$, *kde* $S_i = S_l = J_l^{-1}$ *pro* $i = l \in M$ *a*

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) v_i^T S_i}{v_i^T S_i y_i} = (I + u_i v_i^T) S_i, \qquad u_i = \frac{d_i - S_i y_i}{v_i^T S_i y_i} \tag{1035}$$

pro $l \le i \le l + m - 2$, *kde* $v_i \in R^n$ *je zvolený vektor.*

Poznámka 379. Pokud položíme $v_i = d_i$, dostaneme Broydenovu dobrou metodu. Položíme-li $v_i = e_k$, dostaneme přímou metodu aktualizace sloupců.

Definice 87. *Řekneme, že základní metoda pro řešení soustav nelineárních rovnic je inverzní m-krokovou kvazinewtonovskou metodou s omezenou pamětí, jestliže* $s_i = -S_i f_i$, *kde* $S_i = S_l = J_l^{-1}$ *pro* $i = l \in M$ *a*

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} = S_i + u_i z_i^T, \qquad u_i = \frac{d_i - S_i y_i}{z_i^T y_i} \tag{1036}$$

pro $l \le i \le l + m - 2$, *kde* $z_i \in R^n$ *je zvolený vektor.*

Poznámka 380. Pokud položíme $z_i = y_i$, dostaneme Broydenovu špatnou metodu. Položíme-li $z_i = e_k$, dostaneme inverzní metodu aktualizace sloupců.

K realizaci kvazinewtonovských metod s omezenou pamětí můžeme (tak jako v případě metod s proměnnou metrikou s omezenou pamětí) použít buď vektorové nebo maticové reprezentace. Vektorové reprezentace používají směrové vektory $s_{i+1} = -S_{i+1} f_{i+1} = -p_{i+1}^{i+1}$, kde $p_l^{i+1} = S_l f_{i+1}$ a kde vektory p_{j+1}^{i+1}, $l \le j \le i$, se určují pomocí přímých rekurentních vztahů, ve kterých vystupují již použité vektory u_j, v_j, z_j, $l \le j \le i-1$ (které jsou uloženy v paměti počítače) a nové vektory u_i, v_i, z_i Nejprve odvodíme rekurentní vztahy pro přímé kvazinewtonovské metody s omezenou pamětí.

Věta 262. *Nechť* $p_l^{i+1} = S_l f_{i+1}$ *a*

$$p_{j+1}^{i+1} = p_j^{i+1} + u_j v_j^T p_j^{i+1}, \qquad l \le j \le i,$$

kde

$$u_i = \frac{d_i - (p_i^{i+1} + s_i)}{v_i^T (p_i^{i+1} + s_i)}, \qquad s_i = -S_i f_i,$$

a v_i *je zvolený vektor. Pak platí* $p_{j+1}^{i+1} = S_{j+1} f_{i+1}$, $l \le j \le i$, *takže* $s_{i+1} = -S_{i+1} f_{i+1} = -p_{i+1}^{i+1}$.

Důkaz Důkaz provedeme indukcí. Předpokládejme, že $p_j^{i+1} = S_j f_{i+1}$ pro nějaký index $l \leq j \leq i$ (platí to zcela jistě pro $j = l$). Podle indukčního předpokladu a podle definice 86 lze psát

$$p_{j+1}^{i+1} = p_j^{i+1} + u_j v_j^T p_j^{i+1} = (I + u_j v_j^T) S_j f_{i+1} = S_{j+1} f_{i+1},$$

čímž je indukční krok dokončen. Vzorec pro vektor u_i plyne z toho, že

$$u_i = \frac{d_i - S_i y_i}{v_i^T S_i y_i} = \frac{d_i - S_i(f_{i+1} - f_i)}{v_i^T S_i (f_{i+1} - f_i)} = \frac{d_i - (p_i^{i+1} + s_i)}{v_i^T (p_i^{i+1} + s_i)}.$$

□

Podobné rekurentní vztahy dostaneme pro inverzní kvazinewtonovské metody s omezenou pamětí.

Věta 263. *Nechť $p_l^{i+1} = S_l f_{i+1}$ a*

$$p_{j+1}^{i+1} = p_j^{i+1} + u_j z_j^T f_{i+1}, \qquad l \leq j \leq i,$$

kde

$$u_i = \frac{d_i - (p_i^{i+1} + s_i)}{z_i^T y_i}, \qquad s_i = -S_i f_i,$$

a z_i je zvolený vektor. Pak platí $p_{j+1}^{i+1} = S_{j+1} f_{i+1}$, $l \leq j \leq i$, takže $s_{i+1} = -S_{i+1} f_{i+1} = -p_{i+1}^{i+1}$.

Důkaz Důkaz provedeme indukcí. Předpokládejme, že $p_j^{i+1} = S_j f_{i+1}$ pro nějaký index $l \leq j \leq i$ (platí to zcela jistě pro $j = l$). Podle indukčního předpokladu a podle definice 86 lze psát

$$p_{j+1}^{i+1} = p_j^{i+1} + u_j z_j^T f_{i+1} = (S_j + u_j z_j^T) f_{i+1} = S_{j+1} f_{i+1},$$

čímž je indukční krok dokončen. Vzorec pro vektor u_i plyne z toho, že

$$u_i = \frac{d_i - S_i y_i}{z_i^T y_i} = \frac{d_i - S_i(f_{i+1} - f_i)}{z_i^T y_i} = \frac{d_i - (p_i^{i+1} + s_i)}{z_i^T y_i}.$$

□

Poznámka 381. Pro $i = l \in M$ pokládáme $S_i = S_l = J_l^{-1}$. Jacobiovu matici J_l neinvertujeme. Místo toho používáme trojúhelníkový rozklad $J_l = L_l U_l$. Pak vektor $p_l^{i+1} = S_l f_{i+1}$ získáme řešením soustavy lineárních rovnic $L_l U_l \, p_l^{i+1} = f_{i+1}$

Kvazinewtonovské metody s omezenou pamětí můžeme také realizovat pomocí maticových reprezentací. Pro jejich odvození budeme tak jako v oddílu 9.3 předpokládat, že $i \leq m$ a budeme používat označení $D_i = [d_1, \ldots, d_i]$, $Y_i = [y_1, \ldots, y_i]$, $V_i = [v_1, \ldots, v_i]$, $Z_i = [z_1, \ldots, z_i]$. Abychom zjednodušili zápis budeme v důkazech index i vynechávat a index $i + 1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $D = [d_1, \ldots, d_{i-1}]$, $Y = [y_1, \ldots, y_{i-1}]$, $V = [v_1, \ldots, v_{i-1}]$, $Z = [z_1, \ldots, z_{i-1}]$, takže $D_i = [D, d]$, $Y_i = [Y, y]$, $V_i = [V, v]$, $Z_i = [Z, z]$.

Nejprve odvodíme maticové reprezentace pro přímé kvazinewtonovské metody s omezenou pamětí. Pro tento účel označíme R_i horní trojúhelníkovou matici řádu i takovou, že $(R_i)_{kl} = v_k^T d_l$, $k \leq l$, a $(R_i)_{kl} = 0$, $k > l$. V důkazech budeme používat označení

$$R_i = \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}.$$

Lemma 111. *Nechť A_1 je regulární matice a nechť platí (978) s $v_i^T d_i \neq 0$ pro libovolný index $1 \leq i \leq m$. Pak lze psát*

$$A_{i+1} = A_1 + (Y_i - A_1 D_i) R_i^{-1} V_i^T. \tag{1037}$$

Důkaz Pro $i = 1$ je (1037) ekvivalentní s (978). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (1037) s j místo i platí pro všechny indexy j menší než i. Pro index i můžeme (1037) zapsat ve tvaru

$$A_+ = A_1 + [Y - A_1 D, y - A_1 d] \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}^{-1} \begin{bmatrix} V^T \\ v^T \end{bmatrix}.$$

Jelikož platí

$$\begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}^{-1} = \begin{bmatrix} R^{-1}, & -\dfrac{R^{-1} V^T d}{v^T d} \\ 0, & \dfrac{1}{v^T d} \end{bmatrix}$$

(což lze snadno ověřit vynásobením), můžeme psát

$$A_+ = A_1 + (Y - A_1 D) R^{-1} V^T \left(I - \frac{d v^T}{v^T d} \right) + (y - A_1 d) \frac{v^T}{v^T d} = A + \frac{(y - Ad) v^T}{v^T d},$$

což je právě vztah (978). □

Poznámka 382. Vyjádření (1037) používáme nejčastěji ve spojení s iteračním řešením soustavy rovnic $A_i s_i + f_i = 0$, $i \in N$. Pokud $i > m$, pokládáme $A_i = A_l = J_l$ pro $i = l \in M$ a

$$A_i = A_l + (Y_k - A_l D_k) R_k^{-1} V_k^T \tag{1038}$$

pro $l < i \leq l + m - 1$, kde $D_k = [d_l, \ldots, d_{i-1}]$, $Y_k = [y_l, \ldots, y_{i-1}]$, $V_k = [v_l, \ldots, v_{i-1}]$ a R_k je horní trojúhelníková matice řádu $k = i - l$, jejíž nenulové prvky jsou shodné s prvky matice $V_k^T D_k$. Poznamenejme, že matice V_k se obvykle neukládá (pro Broydenovu dobrou metodu platí $V_k = D_k$ a pro přímou metodu aktualizace sloupců stačí ukládat indexy prvků s maximální absolutní hodnotou sloupců matice D_k). Místo matice Y_k ukládáme matici $U_k = Y_k - A_l D_k$ a součin $A_i p$, $p \in R^n$, počítáme podle vzorce $A_i p = A_l p + U_k R_k^{-1} V_k^T p$.

Věta 264. *Nechť S_1 je regulární matice a nechť platí (995) pro libovolný index $1 \leq i \leq m$. Pak lze psát*

$$S_{i+1} = S_1 + (D_i - S_1 Y_i)(C_i - L_i + V_i^T S_1 Y_i)^{-1} V_i^T S_1, \tag{1039}$$

kde L_i je dolní trojúhelníková matice taková, že $(L_i)_{kl} = 0$, $k < l$, a $(L_i)_{kl} = v_k^T l_j$, $k \geq l$, a C_i je diagonální matice řádu i taková, že $(C_i)_{kl} = v_k^T d_l$, $k = l$, a $(C_i)_{kl} = 0$, $k \neq l$.

Důkaz Přímou inverzí vztahu (1037) (použitím důsledku 8, kde $H = A_1$, $U = (Y_i - A_1 D_i) R_i^{-1}$ a $V = V_i$), dostaneme

$$A_{i+1}^{-1} = A_1^{-1} - A_1^{-1}(Y_i - A_1 D_i)(R_i + V_i^T A_1^{-1}(Y_i - A_1 D_i))^{-1} V_i^T A_1^{-1}.$$

Zřejmě $R_i - V_i^T D_i = C_i - L_i$ a jelikož $A_1^{-1} = S_1$ a $A_{i+1}^{-1} = S_{i+1}$, můžeme poslední vzorec přepsat ve tvaru

$$S_{i+1} = S_1 + (D_i - S_1 Y_i)(C_i - L_i + V_i^T S_1 Y_i)^{-1} V_i^T S_1.$$

□

Nyní odvodíme maticové reprezentace pro inverzní kvazinewtonovské metody s omezenou pamětí. Tentokrát označíme R_i horní trojúhelníkovou matici řádu i takovou, že $(R_i)_{kl} = z_k^T y_l$, $k \leq l$, a $(R_i)_{kl} = 0$, $k > l$. V důkazech budeme používat označenmí

$$R_i = \begin{bmatrix} R, & Z^T y \\ 0, & z^T y \end{bmatrix}.$$

Věta 265. *Nechť S_1 je regulární matice a nechť platí (995) s $z_i^T y_i \neq 0$ pro libovolný index $1 \leq i \leq m$. Pak lze psát*

$$S_{i+1} = S_1 + (D_i - S_1 Y_i) R_i^{-1} Z_i^T. \tag{1040}$$

Důkaz Tvrzení věty plyne z duality. Porovnáme-li (978) s (995), vidíme, že v (1037) stačí provést záměnu $A_1 \to S_1$, $V_i \to Z_i$, $D_i \to Y_i$ a $Y_i \to D_i$. □

Poznámka 383. Vyjádření (1040) používáme k určení směrového vektoru $s_i = -S_i f_i$, $i \in N$. Pokud $i > m$, pokládáme $S_i = S_l = J_l^{-1}$ pro $i = l \in M$ a

$$S_i = S_l + (D_k - S_l Y_k) R_k^{-1} Z_k^T \tag{1041}$$

pro $l < i \leq l + m - 1$, kde $D_k = [d_l, \ldots, d_{i-1}]$, $Y_k = [y_l, \ldots, y_{i-1}]$, $Z_k = [z_l, \ldots, z_{i-1}]$ a R_k je horní trojúhelníková matice řádu $k = i - l$, jejíž nenulové prvky jsou shodné s prvky matice $Z_k^T Y_k$. Poznamenejme, že matice Z_k se obvykle neukládá (pro Broydenovu špatnou metodu platí $Z_k = Y_k$ a pro inverzní metodu aktualizace sloupců stačí ukládat indexy prvků s maximální absolutní hodnotou sloupců matice Y_k). Místo matice D_k ukládáme matici $U_k = D_k - S_l Y_k$ a součin $S_i f_i$ počítáme podle vzorce $S_i f_i = S_l f_i + U_k R_k^{-1} Z_k^T f_i$ (obvykle se používá trojúhelníkový rozklad $J_l = L_l U_l$, takže vektor $S_l f_i$ lze získat řešením soustavy rovnic $L_l U_l (S_l f_i) + f_i = 0$. analogický postup lze použít ke konstrukci předpodmiňovače popsaného v oddílu 12.2.

12.2 Diferenční verze Newtonovy metody pro husté úlohy

Diferenční verze nepřesné Newtonovy metody se vyznačují tím, že se systémy lineárních rovnic řeší nepřesně iteračními metodami. V případě hustých úloh se nepoužívá matice $A = J$ a násobení $q = Ap = J(x)p$ se nahražuje numerickým derivováním

$$J(x)p \approx \frac{f(x + \delta p) - f(x)}{\delta},$$

kde $\delta = \varepsilon / \|p\|$ je vhodná diference (obvykle $\varepsilon = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost). Jinak se algoritmy nemění. Jestliže výpočet vektoru $f(x)$ vyžaduje $O(n)$ operací, je tento způsob úspornější než násobení matice vektorem (obecně $O(n^2)$ operací). Navíc není třeba počítat žádné derivace. Iterační metody pro řešení systémů lineárních rovnic však nesmí používat transponovou matici $A^T = J^T$, což poněkud omezuje jejich výběr (iterační metody pro řešení systémů lineárních rovnic jsou popsány v oddílu 12.8).

Poznámka 384. Snadno se dokáže tvrzení, které je analogií věty 204. Nechť zobrazení $f : \mathcal{D} \to R$ splňuje předpoklad J6. Nechť $q = J(x)p$ a

$$\tilde{q} = \frac{f(x + \delta p) - f(x)}{\delta}, \qquad \delta = \frac{\varepsilon}{\|p\|},$$

kde $x \in \mathcal{D}$ a $x + \delta p \in \mathcal{D}$. Pak platí

$$\|\tilde{q} - q\| \leq \frac{1}{2} \varepsilon \overline{G} \|p\|.$$

Nevýhodou metod studovaných v tomto oddílu je skutečnost, že počet vnitřních iterací zvolené iterační metody, tedy i počet vyčíslení hodnot zobrazení f, může být značně velký, je-li matice $J = J(x)$ špatně podmíněná. Proto je účelné iterační metody vhodně předpodmínit. Potíž je v tom, že neznáme matici J, takže není možné použít standardní postupy.

Tak jako v oddílu 9.7 se zaměříme zejména na pásové předpodmiňovače. Jednou z možností, jak konstruovat pásové předpodmiňovače, je předpokládat, že Jacobiova matice má pásovou strukturu a určovat její prvky numerickým derivováním. K určení všech prvků pásové matice, kde l je šířka pásu, stačí použít l diferencí hodnot zobrazení, tedy spočítat v každém kroku Newtonovy metody l hodnot zobrazení navíc. V tomto oddílu se budeme zabývat pouze tridiagonálními předpodmiňovači.

Věta 266. *Předpokládejme, že Jacobiova matice zobrazení f je tridiagonální matice tvaru*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \dots & 0 & 0 \\ \gamma_2 & \alpha_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & \dots & \gamma_n & \alpha_n \end{bmatrix}. \tag{1042}$$

Položme $v_1 = [\delta_1, 0, 0, \delta_4, 0, 0, 0, \dots]$, $v_2 = [0, \delta_2, 0, 0, \delta_5, 0, \dots]$, $v_3 = [0, 0, \delta_3, 0, 0, \delta_6, \dots]$, *kde* $\delta_i = \varepsilon\overline{\delta}_i$, $1 \le i \le n$, *přičemž* $\varepsilon = \sqrt{\varepsilon_M}$ *a buď* $\overline{\delta}_i = \sqrt{l/n}$ *($l = 3$ je šířka pásu) nebo* $\overline{\delta}_i = \max(|x_i|, 1)$ *pro* $1 \le i \le n$. *Pak platí*

$$\begin{aligned}
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, \qquad & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_{i-1}}, \qquad & &\mathrm{mod}(i,3) = 1, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, \qquad & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_{i-1}}, \qquad & &\mathrm{mod}(i,3) = 2, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_i}, \qquad & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_{i-1}}, \qquad & &\mathrm{mod}(i,3) = 0,
\end{aligned}$$

kde veličiny, v jejichž vzorcích vystupují indexy $i < 1$ nebo $i > n$, nepočítáme.

Důkaz Použitím (1042) se snadno přesvědčíme, že platí

$$Tv_1 = [\alpha_1, \gamma_1, \beta_3, \alpha_4, \gamma_4, \dots]^T, \quad Tv_2 = [\beta_1, \alpha_2, \gamma_2, \beta_4, \alpha_5, \dots]^T, \quad Tv_2 = [0, \beta_2, \alpha_3, \gamma_3, \beta_5, \dots]^T,$$

kde se vyskytují všechny prvky matice T. Použijeme-li stejné úvahy jako v důkazu věty 211, dostaneme dokazované tvrzení. □

K předpodmínění vybrané iterační metody realizující diferenční verzi Newtonovy metody lze též použít kvazinewtonovskou metodu s omezenou pamětí popsanou v poznámce 383. Jelikož kvazinewtonovská metoda vyžaduje dobrou počáteční aproximaci S_i matice J_i^{-1}, postupujeme tak, že pro $i = l \in M$ (M je množina definovaná vztahem (1034)) pokládáme $S_i = S_l = T_l^{-1}$, kde T_l je tridiagonální matice (1042) spočtená podle věty 266, a pro $l \le i \le l + m - 2$ používáme vzorec (1041).

12.3 Diferenční verze Newtonovy metody pro řídké úlohy

Rozsáhlé soustavy nelineárních rovnic se obvykle vyznačují tím, že jejich Jacobiovy matice obsahují málo (typicky $O(n)$) nenulových prvků. Protože počet nulových prvků je obvykle mnohem větší, je účelné ukládat pouze nenulové prvky a jen s nimi provádět aritmetické operace. Nesymetrické řídké matice lze ukládat různým způsobem. V tomto oddílu budeme používat pouze komprimované ukládání po řádcích, které používá tři pole $num(A)$, $adr(A)$, $col(A)$, jejichž význam je popsán v oddílu 10.1.

Diferenční verze Newtonovy metody pro řídké úlohy lze rozdělit do dvou skupin (sloupcové a řádkové metody) podle toho jakým způsobem je organizován přibližný výpočet derivací. Sloupcové metody se používají zejména tehdy, je-li výhodné počítat všechny složky zobrazení f současně. Řádkové metody jsou algoritmicky jednodušší a lze je použít v případě, kdy počítáme hodnoty funkcí $f_i(x)$, $1 \le i \le n$ postupně (v cyklu s indexem i). Použití diferenčních verzí Newtonovy metody je podloženo teorií uvedenou v oddílu 11.4 (lemma 103).

Sloupcové metody jsou založeny na aproximaci sloupců Je_j, $1 \leq j \leq n$, Jacobiovy matice J pomocí diferenčních vzorců

$$J(x)e_j \approx \frac{f(x+\delta_j e_j) - f(x)}{\delta_j}, \tag{1043}$$

kde $\delta_j = \varepsilon\overline{\delta}_j$, přičemž $\varepsilon = \sqrt{\varepsilon_M}$ a buď $\overline{\delta}_j = 1$ nebo $\overline{\delta}_j = \max(|x_j|, 1)$. Je-li matice J řídká může nastat případ, kdy pomocí jedné diference vektorů funkčních hodnot určíme více sloupců této matice. Rozdělme sloupce matice J do l disjunktních skupin $\mathcal{S}_k \subset \{1, \ldots, n\}$, $1 \leq k \leq l$, tak aby submatice $J(\mathcal{S}_k)$, složené ze sloupců matice J patřících do skupin $\mathcal{S}_k$, měly v každém řádku nanejvýš jeden nenulový prvek. Nechť v_k, $1 \leq k \leq l$, jsou vektory takové, že

$$(v_k)_j = e_j^T v_k = \delta_j \Longleftrightarrow j \in \mathcal{S}_k,$$

takže pro libovolný řádkový index $1 \leq i \leq n$ existuje právě jeden sloupcový index $j \in \mathcal{S}_k$ takový, že $J_{ij}(x) = (f_i(x+v_k) - f_i(x))/\delta_j$. Všechny sloupce matice J tedy můžeme určit pomocí l diferencí

$$f(x+v_k) - f(x) \approx Jv_k, \qquad 1 \leq k \leq l$$

(pomocí vektoru v_k určíme prvky submatice $J(\mathcal{S}_k)$). Získání rozkladu $\{1, \ldots, n\} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_l$, takového, aby počet skupin l byl minimální je NP těžká úloha, kterou nelze v obecném případě vyřešit v rozumném čase. Proto se, tak jako ve většině kombinatorických úloh používají algoritmy, které jsou poměrně jednoduché, rychlé a dávají dobrou aproximaci optimálního řešení. Jeden takový algoritmus je popsán v [26]. Je také možné použít volně dostupný zdrojový program v jazyce Fortran [23] popsaný v [22]. Poznamenejme, že sloupcové metody nelze použít, má-li Jacobiova matice husté řádky.

Řádkové metody určují jednotlivé nenulové prvky Jacobiovy matice podle vzorců

$$(J(x))_{ij} \approx \frac{f_i(x+\delta_j e_j) - f_i(x)}{\delta_j}. \tag{1044}$$

Pro každý index $1 \leq i \leq n$, se počítají jen ty diference, které odpovídají nenulovým prvkům v i-tém řádku Jacobiovy matice (jde v podstatě o numerický výpočet redukovaných gradientů zavedených v definici 74). Předpokládejme, že výpočet hodnoty každé z funkcí f_i, $1 \leq i \leq n$, je zhruba stejně náročný. Pak je k výpočtu všech prvú Jacobiovy matice zapotřebí zhruba stejný počet oprerací jako pro $(n_1 + \cdots + n_n)/n$ vyčíslení zobrazení f (zde n_i je počet nenulových prvků v i-tém řádku Jacobiovy matice). Sloupcové metody vyžadují přinejmenším $\max(n_1, \ldots, n_n)$ vyčíslení zobrazení f, což je více než v případě řádkových metod. Není-li tedy společný výpočet hodnot $f_i(x)$, $1 \leq i \leq n$ významně výhodnější než výpočet v cyklu, je lepší používat řádkové metody. Řádkové metody lze navíc použít i tehdy, má-li Jacobiova matice husté řádky.

12.4 Kvazinewtonovské metody pro řídké úlohy

Kvazinewtonovské metody pro řídké úlohy používají kvazinewtonovské aktualizace, které zachovávají strukturu řídké Jacobiovy matice. Označme

$$\begin{aligned} \mathcal{V}_Q &= \{A \in R^{n\times n} : Ad = y\}, \\ \mathcal{V}_J &= \{A \in R^{n\times n} : J_{ij} = 0 \Rightarrow A_{ij} = 0\}. \end{aligned}$$

Tak jako v oddílu 9.7 můžeme definovat operátory ortogonální projekce $\mathcal{P}_Q$, $\mathcal{P}_J$ do lineárních variet $\mathcal{V}_Q$, $\mathcal{V}_J$ předpisem

$$\begin{aligned} \mathcal{P}_Q A &= \min_{\tilde{A}\in\mathcal{V}_Q} \|\tilde{A} - A\|_F, \\ \mathcal{P}_J A &= \min_{\tilde{A}\in\mathcal{V}_J} \|\tilde{A} - A\|_F. \end{aligned}$$

Podobně můžeme definovat operátor ortogonální projekce $\mathcal{P}_{QJ}$ do $\mathcal{V}_Q \cap \mathcal{V}_J$. Podle věty 220 platí

$$\mathcal{P}_{QJ}A = \mathcal{P}_J(A + ud^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = y - (\mathcal{P}_J A)d$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \|d^i\|^2 e_i e_i^T,$$

kde d^i, $1 \le i \le n$, jsou vektory takové, že $d^i_j = d_j$, $J_{ij} \neq 0$ a $d^i_j = 0$, $J_{ij} = 0$. Označíme-li $A_+ = \mathcal{P}_{QJ}A$ a předpokládáme-li, že matice A má stejné rozložení nenulových prvků jako matice J (takže $\mathcal{P}_J A = A$), můžeme vzorec $A_+ = \mathcal{P}_J(A + ud^T)$ zapsat ve tvaru

$$A_+ = A + \sum_{i=1}^{n} \frac{e_i^T(y - Ad)e_i(d^i)^T}{(d^i)^T d^i}, \tag{1045}$$

kde členy s $d^i = 0$ odpadnou. Z vyjádření (1045) plyne, že každý člen v uvedeném součtu mění pouze jeden řádek matice A. Platí

$$e_i^T A_+ = e_i^T A + \frac{e_i^T(y - Ad)}{(d^i)^T d^i}(d^i)^T, \qquad 1 \le i \le n.$$

Metoda, která používá aktualizaci (1045) se nazývá Schubertovou metodou a jelikož je zobecněním Broydenovy dobré metody, má podobné vlastnosti jako Broydenova dobrá metoda. Standardní Schubertova metoda, která používá aktualizaci (1045) v každém iteračním kroku, není globálné konvergentní. Tuto metodu je třeba realizovat pomocí algoritmu 27. Je však možné dokázat, že standardní Schubertova metoda konverguje lokálně Q-superlineárně.

Lemma 112. *Nechť A_+ je matice určená podle (1045). Pak pro libovolnou matici $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_J$ platí*

$$\|A_+ - \tilde{J}\|_F^2 \le \|A - \tilde{J}\|_F^2 - \frac{\|y - Ad\|^2}{\|d\|^2}.$$

Důkaz Jelikož $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_J$, $\mathcal{P}_{QJ}$ je operátor ortogonální projekce do $\mathcal{V}_Q \cap \mathcal{V}_J$ a $A_+ = \mathcal{P}_{QJ}A$, můžeme použít Pythagorovu větu

$$\|A_+ - \tilde{J}\|_F^2 = \|A - \tilde{J}\|_F^2 - \|A_+ - A\|_F^2.$$

Jelikož $\mathcal{V}_Q \cap \mathcal{V}_J \subset \mathcal{V}_Q$, platí $A_+ d = y$, takže

$$\|y - Ad\| = \|(A_+ - A)d\| \le \|A_+ - A\|\|d\| \le \|A_+ - A\|_F\|d\|,$$

což po dosazení do předchozí rovnosti dává dokazované tvrzení. □

Lemma 113. *Nechť A_+ je matice určená podle (1045) a nechť je splněn předpoklad J6. Pak*

$$\|A_+ - J_+\|_F \le \|A - J\|_F + \overline{G}\sqrt{n}\|d\|.$$

Důkaz Podle věty o střední hodnotě (tvrzení 6) platí

$$y = f_+ - f = \tilde{J}d, \qquad \tilde{J} = \int_0^1 J(x + \lambda d)\mathrm{d}\lambda.$$

Použijeme-li předpoklad J6, můžeme psát

$$\|\tilde{J} - J\|_F \leq \sqrt{n} \int_0^1 \|J(x + \lambda d) - J(x)\| \mathrm{d}\lambda \leq \overline{G}\sqrt{n}\|d\| \int_0^1 \lambda d\lambda \leq \frac{1}{2}\overline{G}\sqrt{n}\|d\| \tag{1046}$$

a podobným způsobem dostaneme

$$\|\tilde{J} - J_+\|_F \leq \frac{1}{2}\overline{G}\sqrt{n}\|d\|. \tag{1047}$$

Podle lemmatu 112 platí

$$\begin{aligned} \|A_+ - J_+\|_F &\leq \|A_+ - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \leq \|A - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \\ &\leq \|A - J\|_F + \|\tilde{J} - J\|_F + \|\tilde{J} - J_+\|_F, \end{aligned}$$

což s použitím (1046)–(1047) dává dokazované tvrzení. □

Věta 267. *Nechť je splněn předpoklad J6 a nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$, $\overline{\vartheta} > 0$ taková, že pokud $\|x_1 - x^*\| \leq \overline{\delta}$, $\|A_1 - J_1\| \leq \overline{\vartheta}$ a pokud platí*

$$\begin{aligned} \|A_i d_i + f_i\| &\leq \overline{\omega}\|f_i\|, \\ x_{i+1} &= x_i + d_i, \\ A_{i+1} &= \mathcal{P}_{QJ} A_i \end{aligned}$$

$\forall i \in N$, kde $0 \leq \overline{\omega} < 1$, konverguje posloupnost x_i, $i \in N$, k bodu $x^* \in R^n$.

Důkaz (a) Výsledky dosažené v částech (a)–(b) důkazu věty 235 můžeme přeformulovat (pomocí okolí) tak, že existují čísla $\delta > 0$, $\vartheta > 0$ taková, že pokud $\|x - x^*\| \leq \delta$, $\|(A - J(x))d\| \leq \vartheta\|d\|$ a $\|Ad + f(x)\| \leq \overline{\omega}\|f(x)\|$, kde $0 \leq \overline{\omega} < 1$, platí

$$\frac{1 - \overline{\omega}}{\overline{J}}\|f(x)\| \leq \|d\| \leq \frac{1 + \overline{\omega}}{\underline{J}}\|f(x)\|, \tag{1048}$$

kde $0 < \underline{J} < \|J^{-1}(x^*)\|^{-1} \leq \|J(x^*)\| < \overline{J}$, a

$$\|f(x + d)\| \leq r\|f(x)\|, \tag{1049}$$

kde $\overline{\omega} < r < 1$. Zdůrazněme, že číslo $0 \leq \overline{\omega} < 1$ může být libovolné zatímco čísla $\delta > 0$ a $\vartheta > 0$ mohou vycházet malá.

(b) Zvolme čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ tak, aby platilo

$$\left(1 + \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r}\right) \overline{\delta} \leq \delta$$

a

$$\overline{\vartheta}\sqrt{n} + \overline{G}\sqrt{n} \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \overline{\delta} \leq \vartheta.$$

Nechť $\|x_1 - x^*\| \leq \overline{\delta}$ a $\|A_1 - J_1\| \leq \overline{\vartheta}$. Dokážeme indukcí, že pro libovolný index $i \in N$ platí $\|x_i - x^*\| \leq \delta$ a $\|A_i - J_i\| \leq \vartheta$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme platnost tohoto tvrzení pro $1 \leq i \leq k$. Pak podle (1048), (1049) a podle předpokladu J6 platí

$$\sum_{i=1}^{k} \|d_i\| \leq \frac{1 + \overline{\omega}}{\underline{J}} \sum_{i=1}^{k} \|f_i\| \leq \frac{1 + \overline{\omega}}{\underline{J}} \|f_1\| \sum_{i=1}^{k} r^{i-1} \leq \frac{1}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \|f_1\| \leq \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \|x_1 - x^*\|, \tag{1050}$$

takže

$$\|x_{k+1}-x^*\| \le \|x_1-x^*\| + \sum_{i=1}^{k}\|d_i\| \le \|x_1-x^*\| + \frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1-x^*\| \le \left(1+\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\right)\overline{\delta} \le \delta$$

a použijeme-li lemma 113, dostaneme

$$\|A_{k+1}-J_{k+1}\| \le \|A_{k+1}-J_{k+1}\|_F \le \|A_1-J_1\|_F + \overline{G}\sqrt{n}\sum_{i=1}^{k}\|d_i\| \le \overline{\vartheta}\sqrt{n} + \overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\overline{\delta} \le \vartheta.$$

(c) Podle (a)–(b) platí $\|f_i\| \le r^{i-1}\|f_1\|$ $\forall i \in N$, kde $\overline{\omega} < r < 1$, takže $\sum_{i=1}^{\infty}\|f_i\| < \infty$, $\sum_{i=1}^{\infty}\|d_i\| < \infty$ a tedy i $\|f_i\| \to 0$, $\|d_i\| \to 0$ a $x_i \to x^*$. □

Věta 268. *Nechť jsou splněny předpoklady věty 267 a nechť navíc* $\|A_i d_i + f_i\|/\|f_i\| \to 0$. *Pak* $x_i \to x^*$ *Q-superlineárně.*

Důkaz (a) Podle lemmatu 112 platí

$$\begin{aligned}\frac{\|y-Ad\|^2}{\|d\|^2} &\le \|A-\tilde{J}\|_F^2 - \|A_+-\tilde{J}\|_F^2\\ &= \left(\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F\right)\left(\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F\right)\\ &\le \overline{M}\left(\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F\right).\end{aligned}$$

Existence konstanty $\overline{M} > 0$ plyne z toho, že

$$\begin{aligned}\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F &\le \|A-J\|_F + \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\|\\ &\le 2\|A-J\|_F + 2\overline{G}\sqrt{n}\|d\|\\ &\le 2\|A-J\|_F + 2\overline{G}\sqrt{n}\left(\|x^+-x^*\| + \|x-x^*\|\right),\end{aligned}$$

takže podle části (b) důkazu věty 267 platí

$$\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F \le 2\sqrt{n}\vartheta + 4\overline{G}\sqrt{n}\delta \triangleq \overline{M}.$$

Dále lze psát

$$\|A_+-J_+\|_F \le \|A_+-\tilde{J}\|_F + \|J_+-\tilde{J}\|_F,$$

takže podle lemmatu 113 platí

$$\begin{aligned}\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F &\le \|A-J\|_F + \|J-\tilde{J}\|_F - \|A_+-J_+\|_F + \|J_+-\tilde{J}\|_F \le\\ &\le \|A-J\|_F - \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\|,\end{aligned}$$

což podle (1050) dává

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\|y_i-A_id_i\|^2}{\|d_i\|^2} &\le \overline{M}\|A_1-J_1\|_F + \overline{M}\,\overline{G}\sqrt{n}\sum_{i=1}^{\infty}\|d_i\|\\ &\le \overline{M}\|A_1-J_1|_F + \overline{M}\,\overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1-x^*\| < \infty,\end{aligned}$$

neboli

$$\lim_{i\to\infty} \frac{\|y_i - A_i d_i\|}{\|d_i\|} = 0 \tag{1051}$$

(b) Použijeme-li (1051) a (1046), dostaneme

$$\frac{\|(A_i - J_i)d_i\|}{\|d_i\|} \le \frac{\|(A_i - \tilde{J}_i)d_i\|}{\|d_i\|} + \|\tilde{J}_i - J_i\| \le \frac{\|y_i - A_i d_i\|}{\|d_i\|} + \frac{1}{2}\overline{G}\sqrt{n}\|d_i\|,$$

což spolu s (1051) a $\|d_i\| \to 0$ (podle části (c) důkazu věty 267) dává $\|(A_i - J_i)d_i\|/\|d_i\| \to 0$. Jelikož předpokládáme, že také $\|A_i s_i + f_i\|/\|f_i\| \to 0$, lze použít větu 235, podle které $x_i \to x^*$ Q-superlinárně. □

12.5 Sdružené kvazinewtonovské metody pro řídké úlohy

Sdružené kvazinewtonovské metody pro řídké úlohy používají sdružené kvazinewtonovské aktualizace, které zachovávají strukturu řídké Jacobiovy matice. Označme

$$\begin{aligned} \mathcal{V}_A &= \{A \in R^{n\times n} : A^T q = z\}, \\ \mathcal{V}_J &= \{A \in R^{n\times n} : J_{ij} = 0 \Rightarrow A_{ij} = 0\}, \end{aligned}$$

kde $z = J_+^T q$. Tak jako v oddílu 9.7 můžeme definovat operátory ortogonální projekce $\mathcal{P}_A$, $\mathcal{P}_J$ do lineárních variet $\mathcal{V}_A$, $\mathcal{V}_J$ předpisem

$$\begin{aligned} \mathcal{P}_A A &= \min_{\tilde{A}\in\mathcal{V}_A} \|\tilde{A} - A\|_F, \\ \mathcal{P}_J A &= \min_{\tilde{A}\in\mathcal{V}_J} \|\tilde{A} - A\|_F. \end{aligned}$$

Podobně můžeme definovat operátor ortogonální projekce $\mathcal{P}_{AJ}$ do $\mathcal{V}_A \cap \mathcal{V}_J$.

Věta 269. *Nechť $A \in R^{n\times n}$ a nechť $\mathcal{P}_{AJ}$ je operátor orthogonální projekce do $\mathcal{V}_A \cap \mathcal{V}_J$ Pak*

$$\mathcal{P}_{AJ}A = \mathcal{P}_J(A + qu^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = z - (\mathcal{P}_J A)^T q$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \|q^i\|^2 e_i e_i^T.$$

kde q^i, $1 \le i \le n$, jsou vektory takové, že $q_j^i = q_j$, $J_{ji} \ne 0$ a $q_j^i = 0$, $J_{ji} = 0$. Označíme-li $A_+ = \mathcal{P}_{AJ}A$ a předpokládáme-li, že matice A má stejné rozložení nenulových prvků jako matice J (takže $\mathcal{P}_J A = A$), můžeme vzorec $A_+ = \mathcal{P}_J(A + qu^T)$ zapsat ve tvaru

$$A_+ = A - \sum_{i=1}^{n} \frac{q^T(A - J_+)e_i q^i e_i^T}{(q^i)^T q^i}, \tag{1052}$$

kde členy s $q^i = 0$ odpadnou.

Důkaz Položíme-li ve větě 220 $B = A^T$, $G = J^T$, $d = q$, $y = z$, dostaneme

$$A_+^T = \mathcal{P}_{QG}B = \mathcal{P}_G(B + ud^T) = \mathcal{P}_J(A^T + uq^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = z - (\mathcal{P}_J A)^T q = (J_+ - (\mathcal{P}_J A))^T q$. Má-li matice A stejné rozložení nenulových prvků jako matice J, můžeme tento vzorec zapsat ve tvaru

$$A_+^T = A^T + \sum_{i=1}^{n} \frac{e_i^T (J_+^T - A^T) q e_i (q^i)^T}{(q^i)^T q^i},$$

což po úpravě dává (1052). □

Z vyjádření (1052) plyne, že každý člen v uvedeném součtu mění pouze jeden sloupec matice A. Platí

$$A_+ e_i = Ae_i - \frac{q^T (A - J_+) e_i}{(q^i)^T q^i} q^i, \qquad 1 \le i \le n.$$

Sdružené kvazinewtonovské metody pro řídké úlohy je třeba realizovat pomocí algoritmu 27. Pokud však položíme $q_i = f_{i+1}$, dostaneme reziduální sdruženou kvazinewtonovskou metodu, pro kterou platí $A_{i+1}^T f_{i+1} = J_{i+1}^T f_{i+1}$. Tato metoda realizovaná podle poznámky 372 je globálně konvergentní. Ukážeme, že sdružená kvazinewtonovská metoda, používající aktualizaci (1052), je lokálně Q-superlineárně konvergentní (bez přerušování iteračního procesu).

Lemma 114. *Nechť A_+ je matice určená podle (1052). Pak platí*

$$\|A_+ - J_+\|_F^2 \le \|A - J_+\|_F^2 - \frac{\|(A - J_+)^T q\|^2}{\|q\|^2}.$$

Důkaz Jelikož $J_+ \in \mathcal{V}_A \cap \mathcal{V}_J$, $\mathcal{P}_{AJ}$ je operátor ortogonální projekce do $\mathcal{V}_A \cap \mathcal{V}_J$ a $A_+ = \mathcal{P}_{AJ} A$, můžeme použít Pythagorovu větu

$$\|A_+ - J_+\|_F^2 = \|A - J_+\|_F^2 - \|A - A_+\|_F^2.$$

Jelikož $\mathcal{V}_A \cap \mathcal{V}_J \subset \mathcal{V}_A$, platí $A_+^T q = J_+^T q$, takže

$$\|(A - J_+)^T q\| = \|(A - A_+)^T . q\| \le \|A - A_+\| \|q\| \le \|A - A_+\|_F \|q\|,$$

což po dosazení do předchozí rovnosti dává dokazované tvrzení. □

Lemma 115. *Nechť A_+ je matice určená podle (1052) a nechť je splněn předpoklad J6. Pak*

$$\|A_+ - J_+\|_F \le \|A - J\|_F + \overline{G}\sqrt{n}\|d\|.$$

Důkaz Použijeme-li předpoklad J6, dostaneme

$$\|J_+ - J\|_F \le \sqrt{n}\|J_+ - J\| \le \overline{G}\sqrt{n}\|d\|. \tag{1053}$$

Podle lemmatu 114 pak platí

$$\begin{aligned} \|A_+ - J_+\|_F &\le \|A - J_+\|_F \le \|A - J\|_F + \|J_+ - J\|_F \\ &\le \|A - J\|_F + \overline{G}\sqrt{n}\|d\|. \end{aligned}$$

□

Věta 270. *Nechť je splněn předpoklad J6 a nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$, $\overline{\vartheta} > 0$ taková, že pokud $\|x_1 - x^*\| \le \overline{\delta}$, $\|A_1 - J_1\| \le \overline{\vartheta}$ a pokud platí*

$$\begin{aligned} \|A_i d_i + f_i\| &\le \overline{\omega}\|f_i\|, \\ x_{i+1} &= x_i + d_i, \\ A_{i+1} &= \mathcal{P}_{AJ} A_i \end{aligned}$$

$\forall i \in N$, kde $0 \le \overline{\omega} < 1$, konverguje posloupnost x_i, $i \in N$, k bodu $x^* \in R^n$.

Důkaz Důkaz tohoto tvrzení je prakticky stejný jako důkaz věty 267 (místo matic $\tilde{J}_i$ $i \in N$, používáme matice J_{i+1}, $i \in N$, a místo lemmatu 113 používáme lemma 115). □

Věta 271. *Nechť jsou splněny předpoklady věty 270 a nechť navíc* $\|A_i d_i + f_i\|/\|f_i\| \to 0$. *Pak pokud vektory* q_i, $i \in N$, *vybíráme podle (1019), (1020) (nebo (1021), když* $\overline{\omega} = 0$*), konverguje* $x_i \to x^*$ *Q-superlineárně.*

Důkaz (a) Podle lemmatu 114 platí

$$\begin{aligned}\frac{\|(A-J_+)^T q\|^2}{\|q\|^2} &\leq \|A-J_+\|_F^2 - \|A_+ - J_+\|_F^2 \\ &= (\|A-J_+\|_F - \|A_+-J_+\|_F)\,(\|A-J_+\|_F + \|A_+-J_+\|_F) \\ &\leq \overline{M}\,(\|A-J_+\|_F - \|A_+-J_+\|_F)\,.\end{aligned}$$

Existence konstanty $\overline{M} > 0$ plyne z toho, že

$$\begin{aligned}\|A-J_+\|_F + \|A_+-J_+\|_F &\leq \|A-J\|_F + \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\| \\ &\leq 2\|A-J\|_F + 2\overline{G}\sqrt{n}\|d\| \\ &\leq 2\|A-J\|_F + 2\overline{G}\sqrt{n}\left(\|x^+-x^*\| + \|x-x^*\|\right) \\ &\leq 2\sqrt{n}\vartheta + 4\overline{G}\sqrt{n}\delta \triangleq \overline{M}.\end{aligned}$$

Dále lze psát

$$\begin{aligned}\|A-J_+\|_F - \|A_+-J_+\|_F &\leq \|A-J\|_F + \|J_+-J\|_F - \|A_+-J_+\|_F \\ &\leq \|A-J\|_F - \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\|,\end{aligned}$$

což podle (1050) dává

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\|(A_i-J_{i+1})^T q_i\|^2}{\|q_i\|^2} &\leq \overline{M}\|A_1-J_1\|_F + \overline{M}\,\overline{G}\sqrt{n}\sum_{i=1}^{\infty}\|d_i\| \\ &\leq \overline{M}\|A_1-J_1|_F + \overline{M}\,\overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1-x^*\| < \infty,\end{aligned}$$

neboli

$$\lim_{i\to\infty}\frac{\|(A_i-J_{i+1})^T q_i\|}{\|q_i\|} = 0. \tag{1054}$$

Dále platí

$$\frac{\|(A_i-J_i)^T q_i\|}{\|q_i\|} \leq \frac{\|(A_i-J_{i+1})^T q_i\|}{\|q_i\|} + \|J_{i+1}-J_i\| \leq \frac{\|(A_i-J_{i+1})^T q_i\|}{\|q_i\|} + \overline{G}\sqrt{n}\|d_i\|,$$

což spolu s (1054) a $\|d_i\| \to 0$ dává

$$\lim_{i\to\infty}\frac{\|(A_i-J_i)^T q_i\|}{\|q_i\|} = 0. \tag{1055}$$

(b) Ukážeme, že pro vektory q_i, $i \in N$, zvolené podle (1019) nebo (1020) platí

$$\lim_{i\to\infty} \frac{\|q_i - (A_i - J_i)d_i\|}{\|d_i\|} = 0. \tag{1056}$$

Použijeme-li (1019), dostaneme

$$\begin{aligned} \|q_i - (A_i - J_i)d_i\| &= \|(A_i - J_{i+1})d_i - (A_i - J_i)d_i\| = \|(J_i - J_{i+1})d_i\| \\ &\leq \|J_i - J_{i+1}\|\|d_i\| \leq \overline{G}\sqrt{n}\|d_i\|^2, \end{aligned}$$

což spolu s $\|d_i\| \to 0$ dává (1056). Použijeme-li (1020), dostaneme

$$\begin{aligned} \|q_i - (A_i - J_i)d_i\| &= \|(A_i d_i - (f_{i+1} - f_i) - (A_i - J_i)d_i\| = \|J_i d_i - (f_{i+1} - f_i)\| \\ &= \|f_i + J_i d_i - (f_i + J_i d_i + o(\|d_i\|))\| = o(\|d_i\|) \end{aligned}$$

(používáme dva členy Taylorova rozvoje), což spolu s $\|d_i\| \to 0$ dává (1056).

(c) Ukážeme, že z (1055) a (1056) plyne

$$\lim_{i\to\infty} \frac{\|(A_i - J_i)d_i\|}{\|d_i\|} = 0. \tag{1057}$$

Platí

$$\|q_i\|^2 = q_i^T(q_i - (A_i - J_i)d_i + (A_i - J_i)d_i) \leq \|q_i\|\|q_i - (A_i - J_i)d_i\| + \|(A_i - J_i)^T q_i\|\|d_i\|,$$

takže

$$\frac{\|q_i\|}{\|d_i\|} \leq \frac{\|q_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|(A_i - J_i)^T q_i\|}{\|q_i\|}$$

a

$$\frac{\|(A_i - J_i)d_i\|}{\|d_i\|} \leq \frac{\|q_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|q_i\|}{\|d_i\|} \leq 2\frac{\|q_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|(A_i - J_i)^T q_i\|}{\|q_i\|},$$

což spolu s (1055) a (1056) dává (1057). Z (1057) a $\|A_i d_i + f_i\|/\|f_i\| \to 0$ plyne superlineární konvergence (věta 235). □

12.6 Metody založené na aktualizaci nesymetrického trojúhelníkového rozkladu

Soustavu lineárních rovnic $As + f = 0$ můžeme řešit buď přímo nebo iteračně. Přímé řešení je založeno na použití nesymetrického trojúhelníkového rozkladu

$$PA = LU,$$

kde P je permutační matice, která se vybírá tak, aby počet nově vzniklých nenulových prvků byl co nejmenší, L je dolní trojúhelníková matice s jednotkami na hlavní diagonále a U je horní trojúhelníková matice. Nalezení permutační matice P a následné určení struktury trojúhelníkových matic L a U se nazývá symbolickou faktorizací. Na rozdíl od řídkého Choleského rozkladu (oddíl 9.6) nestačí provádět symbolickou faktorizaci pouze na začátku iteračního procesu, neboť permutace řádků (výběr pivotů) může ovlivnit stabilitu eliminačního procesu. Dá se tedy konstatovat, že nesymetrický trojúhelníkový rozklad je časově dosti náročný, takže je výhodné omezit jeho provádění. Tato myšlenka je základem metod založených na aktualizaci nesymetrického trojúhelníkového rozkladu. Na rozdíl od Schubertovy metody, kde se matice A^+ vybírá tak, aby byla splněna kvazinewtonovská podmínka $A_+ d = y$, $d = x_+ - x$, $y = f_+ - f$, se pokládá $PA_+ = LU_+$ a matice U_+ se vybírá tak, aby byla splněna kvazinewtonovská podmínka

$$U_+ d = L^{-1} P y.$$

Jelikož musí být zároveň zachována struktura horní trojúhelníkové matice, můžeme použít postup popsaný v oddílu 12.4 (vzorec (1045)). Výsledkem je aktualizace

$$U_+ = U + \sum_{i=1}^{n} \frac{e_i^T (L^{-1} P y - U d) e_i (d^i)^T}{(d^i)^T d^i}, \tag{1058}$$

kde d^i, $1 \leq i \leq n$, jsou vektory takové, že $d^i_j = d_j$, $U_{ij} \neq 0$ a $d^i_j = 0$, $U_{ij} = 0$ (členy s $d^i = 0$ odpadnou). Metoda, která používá aktualizaci (1058) se nazývá Dennisovou-Marwilovou metodou. Obvykle se realizuje tak, že se provede nesymetrický trojúhelníkový rozklad $PJ = LU$ pak se v m po sobě následujících iteračních krocích použije aktualizace (1058). Po m aktualizacích (1058) nebo po vynuceném přerušení iteračního procesu se opět provede nesymetrický trojúhelníkový rozklad $PJ = LU$.

Ještě jednodušší metodou je metoda škálování řádků. V tomto případě se pokládá $PA_+ = D_+ LU$ a diagonální matice D_+ se vybírá tak, aby byla splněna kvazinewtonovská podmínka

$$D_+ LU d = P y.$$

Zapíšeme-li tuto podmínku ve tvaru

$$\sum_{i=1}^{n} D_+ e_i e_i^T LU d = P y$$

a přihlédneme-li k tomu, že matice D_+ je diagonální, můžeme psát

$$e_i^T D_+ e_i e_i^T LU d = e_i^T P y,$$

$1 \leq i \leq n$, takže

$$PA_+ = D_+ LU, \qquad e_i^T D_+ e_i = \frac{e_i^T P y}{e_i^T LU d}. \tag{1059}$$

Také metodu škálování řádků je třeba po m iteračních krocích přerušovat s tím, že se provede nesymetrický trojúhelníkový rozklad $PJ = LU$.

12.7 Nedokonalé diferenční verze Newtonovy metody

Nedokonalé diferenční verze Newtonovy metody jsou založeny na myšlence, že se přibližný výpočet derivací provádí pouze v některých iteračních krocích. Nejjednodušší je Shamanského metoda, kdy se položí $A = J$ a pak se v m po sobě jdoucích iteračních krocích používá tatáž matice ($A_+ = A$). Důmyslnější metody jsou založeny na podobném principu jako sloupcové diferenční verze Newtonovy metody. Opět se určí rozklad $\{1, \ldots, n\} = \mathcal{S}_1 \cup \cdots \cup \mathcal{S}_k$ sloupců matice J do k disjunktních skupin $\mathcal{S}_i$, $1 \leq i \leq k$, tak, aby submatice $J(\mathcal{S}_i)$, složené ze sloupců matice J patřících do skupin $\mathcal{S}_i$, měly v každém řádku nanejvýš jeden nenulový prvek (oddíl 12.3). Pak se v každém iteračním kroku určují sloupce matice J patřící pouze do jedné skupiny a ostatní sloupce se nemění. Konkrétněji, nechť $l = \mathrm{mod}_k i$ ($\mathrm{mod}_k i$ je zbytek po dělení čísla i číslem k). V i-tém iteračním kroku se použije vektor v_i takový, že

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_l$$

a pomocí diference

$$\frac{f(x + \delta v_i) - f(x)}{\delta} \approx J v_i$$

se určí sloupce matice J patřící do skupiny $\mathcal{S}_l$. Sloupce patřící do ostatních skupin se ponechají beze změny.

Tuto metodu, která se nazývá Liovou metodou, lze kombinovat se Schubertovou metodou tak, že se v každém iteračním kroku po určení sloupců matice J, patřících do skupiny $\mathcal{S}_l$, provede navíc aktualizace (1045).

12.8 Iterační řešení systémů lineárních rovnic s nesymetrickou maticí

Pro řešení systému lineárních rovnic $As + f = 0$ s nesymetrickou maticí A existuje celá řada iteračních metod. Můžeme je zhruba rozdělit na dvě skupiny:

(1) Metody s krátkými rekurentními vztahy.
(2) Metody s dlouhými rekurentními vztahy.

Výhodou metod s krátkými rekurentními vztahy (jsou to dvojčlenné nebo trojčlenné rekurence) je nízký počet numerických operací a ukládaných hodnot (je jich $O(n)$). Nevýhodou těchto metod je možnost selhání (dělení nulou) během iteračního procesu. Metody s dlouhými rekurentními vztahy mají opačné vlastnosti. V n-tém iteračním kroku se pracuje s n vektory dimenze n, což vyžaduje $O(n^2)$ numerických operací a ukládaných hodnot (teoreticky je zapotřebí k získání řešení n iteračních kroků). Zato nedochází k selhání během iteračního procesu (každý jeho krok je korektně definován).

V tomto textu, který si nečiní nároky na úplnost, se budeme zabývat pouze zhlazenou metodou CGS používající krátké rekurentní vztahy a metodou GMRES používající dlouhé rekurentní vztahy.

Definice 88. *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$, $\tilde{f} \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \quad f_1 = f, \quad \tilde{f}_1 = \tilde{f}, \quad p_1 = -f_1, \quad \tilde{p}_1 = -\tilde{f}_1$$

a

$$q_i = Ap_i, \quad \tilde{q}_i = A^T\tilde{p}_i, \quad \alpha_i = \tilde{f}_i^T f_i/\tilde{p}_i^T q_i,$$

$$s_{i+1} = s_i + \alpha_i p_i,$$

$$f_{i+1} = f_i + \alpha_i q_i, \quad \tilde{f}_{i+1} = \tilde{f}_i + \alpha_i\tilde{q}_i, \quad \beta_i = \tilde{f}_{i+1}^T f_{i+1}/\tilde{f}_i^T f_i,$$

$$p_{i+1} = -f_{i+1} + \beta_i p_i, \quad \tilde{p}_{i+1} = -\tilde{f}_{i+1} + \beta_i\tilde{p}_i$$

pro $1 \le i \le n$, nazveme metodu bikonjugovaných gradientů (BCG) určenou maticí $A \in R^{n\times n}$ a vektory $f \in R^n$, $\tilde{f} \in R^n$.

Věta 272. *Uvažujme metodu bikonjugovaných gradientů určenou regulární maticí $A \in R^{n\times n}$ a vektory $f \in R^n$, $\tilde{f} \in R^n$. Nechť $\tilde{f}_i^T f_i \neq 0$ a $\tilde{p}_i^T q_i \neq 0$, $1 \le i \le n$. Pak platí $f_{n+1} = 0$ a vektor s_{n+1} je řešením soustavy rovnic $As + f = 0$.*

Důkaz Předpokládejme, že $\tilde{f}_i^T f_i \neq 0$ a $\tilde{p}_i^T q_i \neq 0$, $1 \le i \le n$. Dokážeme indukcí, že platí

$$\tilde{p}_j^T f_i = p_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1, \tag{1060}$$

$$\tilde{f}_j^T f_i = f_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1, \tag{1061}$$

$$\tilde{p}_j^T q_i = p_j^T \tilde{q}_i = 0 \quad \forall 1 \le j < i \le n. \tag{1062}$$

Z (1061) plyne, že vektory f_i, $1 \le i \le n$ (a také $\tilde{f}_i$, $1 \le i \le n$), jsou lineárně nezávislé. Jestliže totiž $\lambda_1 f_1 + \cdots + \lambda_n f_n = 0$, pak pro $1 \le i \le n$ platí

$$\tilde{f}_i^T \left(\sum_{j=1}^{n} \lambda_j f_j \right) = \lambda_i \tilde{f}_i^T f_i = 0$$

a jelikož $\tilde{f}_i^T f_i \neq 0$, musí být $\lambda_i = 0$. Podobně z (1062) plyne, že vektory p_i, $1 \leq i \leq n$ (a také $\tilde{p}_i$, $1 \leq i \leq n$), jsou lineárně nezávislé. Jelikož $f_{n+1} = As_{n+1} + f$ (plyne to z rekurentních vztahů metody BCG), vektory $\tilde{f}_i$, $1 \leq i \leq n$, jsou lineárně nezávislé a

$$\tilde{f}_j^T f_{n+1} = 0 \quad \forall 1 \leq j \leq n,$$

musí platit $f_{n+1} = As_{n+1} + f = 0$. Pro $i = 1$ (1060)–(1062) platí, neboť není co dokazovat.

(a) Nechť $i \leq n$. Podle indukčních předpokladů (1060) a (1062) platí

$$\tilde{p}_j^T f_{i+1} = \tilde{p}_j^T f_i + \alpha_i \tilde{p}_j^T q_i = 0,$$

$$p_j^T \tilde{f}_{i+1} = p_j^T \tilde{f}_i + \alpha_i p_j^T \tilde{q}_i = 0$$

$\forall 1 \leq j < i$. Z (1060) a (1062) pak plyne

$$\tilde{p}_i^T f_{i+1} = \tilde{p}_i^T f_i + \alpha_i \tilde{p}_i^T q_i = -\tilde{f}_i^T f_i + \beta_{i-1} \tilde{p}_{i-1}^T f_i + \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T q_i} \tilde{p}_i^T q_i = 0,$$

$$p_i^T \tilde{f}_{i+1} = p_i^T \tilde{f}_i + \alpha_i p_i^T \tilde{q}_i = -f_i^T \tilde{f}_i + \beta_{i-1} p_{i-1}^T \tilde{f}_i + \frac{f_i^T \tilde{f}_i}{p_i^T \tilde{q}_i} p_i^T \tilde{q}_i = 0.$$

Je tedy $\tilde{p}_j^T f_{i+1} = 0$, $p_j^T \tilde{f}_{i+1} = 0$ $\forall 1 \leq j \leq i$.

(b) Nechť $i \leq n$. Z rekurentních vztahů metody BCG plyne

$$\tilde{f}_1 = -\tilde{p}_1,$$

$$\tilde{f}_j = -\tilde{p}_j + \beta_{j-1} \tilde{p}_{j-1} \quad \forall 1 < j \leq i,$$

$$f_1 = -p_1,$$

$$f_j = -p_j + \beta_{j-1} p_{j-1} \quad \forall 1 < j \leq i,$$

takže podle (a) platí

$$\tilde{f}_1^T f_{i+1} = -\tilde{p}_1^T f_{i+1} = 0,$$

$$\tilde{f}_j^T f_{i+1} = -\tilde{p}_j^T f_{i+1} + \beta_{j-1} \tilde{p}_{j-1} f_{i+1} = 0 \quad \forall 1 < j \leq i,$$

$$f_1^T \tilde{f}_{i+1} = -p_1^T \tilde{f}_{i+1} = 0,$$

$$f_j^T \tilde{f}_{i+1} = -p_j^T \tilde{f}_{i+1} + \beta_{j-1} p_{j-1} \tilde{f}_{i+1} = 0 \quad \forall 1 < j \leq i.$$

(c) Nechť $i < n$. Z rekurentních vztahů metody BCG a z (a) plyne

$$\begin{aligned}
\tilde{p}_j^T q_{i+1} &= \tilde{p}_j^T A p_{i+1} = -\tilde{p}_j^T A f_{i+1} + \beta_i \tilde{p}_j^T A p_i \\
&= -(\tilde{f}_{j+1} - \tilde{f}_j)^T f_{i+1}/\alpha_j + \beta_i \tilde{p}_j^T q_i = 0, \\
p_j^T \tilde{q}_{i+1} &= p_j^T A^T \tilde{p}_{i+1} = -p_j^T A^T \tilde{f}_{i+1} + \beta_i p_j^T A^T \tilde{p}_i \\
&= -(f_{j+1} - f_j)^T \tilde{f}_{i+1}/\alpha_j + \beta_i p_j^T \tilde{q}_i = 0
\end{aligned}$$

$\forall 1 \leq j < i$. Použijeme-li navíc (b), dostaneme

$$\begin{aligned}
\tilde{p}_i^T q_{i+1} &= -\frac{1}{\alpha_i}(\tilde{f}_{i+1} - \tilde{f}_i)^T f_{i+1} + \beta_i \tilde{p}_i^T q_i = -\frac{\tilde{p}_i^T q_i}{\tilde{f}_i^T f_i} \tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} \tilde{p}_i^T q_i = 0, \\
p_i^T \tilde{q}_{i+1} &= -\frac{1}{\alpha_i}(f_{i+1} - f_i)^T \tilde{f}_{i+1} + \beta_i p_i^T \tilde{q}_i = -\frac{p_i^T \tilde{q}_i}{\tilde{f}_i^T f_i} \tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} p_i^T \tilde{q}_i = 0,
\end{aligned}$$

takže $\tilde{p}_j^T q_{i+1} = 0$ a $p_j^T \tilde{q}_{i+1} = 0$ $\forall 1 \leq j \leq 1$. □

Poznámka 385. Iterační proces metody BCG může skončit dříve než po n krocích. Buď $f_k = 0$ pro nějaký index $k \leq n$ (takže dostaneme řešení soustavy rovnic $As + f = 0$ po méně než n krocích) nebo $f_k \neq 0$ a $\tilde{f}_k^T f_k = 0$ (principiální selhání společné všem metodám odvozeným z nesymetrického Lanczosova procesu) nebo $f_k \neq 0$ a $\tilde{p}_k^T q_k = 0$ (selhání vlastní metodě BCG). V běžných případech k selhání nedochází (je vyjímečné), mohou však nastávat potíže se stabilitou, pokud $f_k \neq 0$ a $\tilde{f}_k^T f_k \approx 0$ nebo $f_k \neq 0$ a $\tilde{p}_k^T q_k \approx 0$.

Lemma 116. *Nechť jsou splněny předpoklady věty 272. Pak vektory f_j, $1 \leq j \leq i \leq n$, (a také vektory p_j, $1 \leq j \leq i \leq n$) tvoří bázi v Krylovově podprostoru*

$$\mathcal{K}_i = \mathrm{span}\{f,\ Af, \ldots, A^{i-1} f\}$$

a vektory $\tilde{f}_j$, $1 \leq j \leq i \leq n$ (a také vektory $\tilde{p}_j$, $1 \leq j \leq i \leq n$) tvoří bázi v Krylovově podprostoru

$$\tilde{\mathcal{K}}_i = \mathrm{span}\{\tilde{f},\ (A^T)\tilde{f}, \ldots, (A^T)^{i-1}\tilde{f}\}.$$

Důkaz (indukcí) pro $i = 1$ je tvrzení zřejmé. Předpokládejme, že tvrzení platí pro nějaký index $i < n$. Jelikož $f_i \in \mathcal{K}_i$ a $p_i \in \mathcal{K}_i$, dostaneme $f_{i+1} = f_i + \alpha_i A p_i \in \mathcal{K}_{i+1}$ a $p_{i+1} = -f_{i+1} + \beta_i p_i \in \mathcal{K}_{i+1}$, a jelikož vektory f_j, $1 \leq j \leq i+1$ (a také vektory p_j, $1 \leq j \leq i+1$) jsou lineárně nezávislé (důkaz věty 272), tvoří tam bázi. Jelikož $\tilde{f}_i \in \tilde{\mathcal{K}}_i$ a $\tilde{p}_i \in \tilde{\mathcal{K}}_i$, dostaneme $\tilde{f}_{i+1} = \tilde{f}_i + \alpha_i A^T \tilde{p}_i \in \tilde{\mathcal{K}}_{i+1}$ a $\tilde{p}_{i+1} = -\tilde{f}_{i+1} + \beta_i \tilde{p}_i \in \tilde{\mathcal{K}}_{i+1}$, a jelikož vektory $\tilde{f}_j$, $1 \leq j \leq i+1$ (a také vektory $\tilde{p}_j$, $1 \leq j \leq i+1$) jsou lineárně nezávislé (důkaz věty 272), tvoří tam bázi. □

Poznámka 386. Nechť jsou splněny předpoklady věty 272. Pak platí

$$\begin{aligned}
f_i &= \varphi_i(A) f, \qquad & \tilde{f}_i &= \varphi_i(A^T)\tilde{f}, \\
p_i &= -\psi_i(A) f, \qquad & \tilde{p}_i &= -\psi_i(A^T)\tilde{f}
\end{aligned}$$

$\forall 1 \leq i \leq n+1$, kde φ_i a ψ_i jsou maticové polynomy stupně nejvýše $i-1$. Tyto polynomy lze počítat pomocí rekurentních vztahů $\varphi_1 = I$, $\psi_1 = I$ a

$$\begin{aligned}
\varphi_{i+1} &= \varphi_i - \alpha_i A \psi_i, \\
\psi_{i+1} &= \varphi_{i+1} + \beta_i \psi_i
\end{aligned}$$

$1 \le i \le n$. Plyne to bezprostředně z rekurentních vztahů metody BCG. Koeficienty α_i a β_i, $1 \le i \le n$, lze vyjádřit pomocí polynomů φ_i a ψ_i, $1 \le i \le n$, tak, že

$$\alpha_i = \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T A p_i} = \frac{\tilde{f}^T \varphi_i^2(A) f}{\tilde{f}^T A \psi_i^2(A) f}, \qquad \beta_i = \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} = \frac{\tilde{f}^T \varphi_{i+1}^2(A) f}{\tilde{f}^T \varphi_i^2(A) f},$$

neboť matice A a polynom $\psi_i(A)$ komutují). Jelikož koeficienty α_i a β_i, $1 \le i \le n$ lze použít také k určení polynomů $\varphi_i^2(A)$ a $\psi_i^2(A)$, $1 \le i \le n$, můžeme definovat nový iterační proces $\overline{s}_i \in R^n$, $1 \le i \le n+1$ tak, aby platilo $\overline{f}_i = A\overline{s}_i + f = \varphi_i^2(A) f$, $1 \le i \le n+1$.

Lemma 117. *Nechť maticové polynomy φ_i a ψ_i splňují rekurentní vztahy*

$$\varphi_1 = I, \quad \psi_1 = I$$

a

$$\begin{aligned} \varphi_{i+1} &= \varphi_i - \alpha_i A \psi_i, \\ \psi_{i+1} &= \varphi_{i+1} + \beta_i \psi_i \end{aligned}$$

pro $1 \le i \le n$. Pak maticové polynomy φ_i^2 a ψ_i^2 splňují rekurentní vztahy

$$\varphi_1^2 = I, \quad \psi_1^2 = I, \quad \varphi_1 \psi_1 = I$$

a

$$\begin{aligned} \varphi_{i+1}\psi_i &= \varphi_i \psi_i - \alpha_i A \psi_i^2, \\ \varphi_{i+1}^2 &= \varphi_i^2 - \alpha_i A(\varphi_i \psi_i + \varphi_{i+1}\psi_i), \\ \varphi_{i+1}\psi_{i+1} &= \varphi_{i+1}^2 + \beta_i \varphi_{i+1}\psi_i, \\ \psi_{i+1}^2 &= \varphi_{i+1}\psi_{i+1} + \beta_i(\varphi_{i+1}\psi_i + \beta_i \psi_i^2) \end{aligned}$$

pro $1 \le i \le n$.

Důkaz Vynásobíme-li rekurentní vztah pro φ_{i+1} polynomem ψ_i, dostaneme

$$\varphi_{i+1}\psi_i = \varphi_i \psi_i - \alpha_i A \psi_i^2.$$

Umocníme-li vztah pro φ_{i+1}, dostaneme

$$\begin{aligned} \varphi_{i+1}^2 &= \varphi_i^2 - 2\alpha_i A \varphi_i \psi_i + \alpha_i^2 A^2 \psi_i^2 = \varphi_i^2 - \alpha_i A(2\varphi_i\psi_i - \alpha_i A \psi_i^2)) \\ &= \varphi_i^2 - \alpha_i A(\varphi_i \psi_i + \varphi_{i+1}\psi_i). \end{aligned}$$

Vynásobíme-li rekurentní vztah pro ψ_{i+1} polynomem φ_{i+1}, dostaneme

$$\varphi_{i+1}\psi_{i+1} = \varphi_{i+1}^2 + \beta_i \varphi_{i+1}\psi_i.$$

Umocníme-li vztah pro ψ_{i+1}, dostaneme

$$\begin{aligned} \psi_{i+1}^2 &= \varphi_{i+1}^2 + 2\beta_i \varphi_{i+1}\psi_i + \beta_i^2 \psi_i^2 = \varphi_{i+1}^2 + \beta_i \varphi_{i+1}\psi_i + \beta_i(\varphi_{i+1}\psi_i + \beta_i \psi_i^2) \\ &= \varphi_{i+1}\psi_{i+1} + \beta_i(\varphi_{i+1}\psi_i + \beta_i \psi_i^2). \end{aligned}$$

□

Položíme-li nyní $\overline{f}_i = \varphi_i^2 f$, $p_i = \psi_i^2 f$, $v_i = A\psi_i^2 f = Ap_i$, $u_i = \varphi_i \psi_i f$, $q_i = \varphi_{i+1}\psi_i f = u_i - \alpha_i v_i$, dostaneme rekurentní vztahy, které jsou základem metody CGS.

Definice 89. *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$, $\tilde{f} \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$\overline{s}_1 = 0, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

a

$$\begin{aligned} v_i &= Ap_i, \quad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i, \\ q_i &= u_i - \alpha_i v_i, \\ \overline{s}_{i+1} &= \overline{s}_i - \alpha_i(u_i + q_i), \\ \overline{f}_{i+1} &= \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i, \\ u_{i+1} &= \overline{f}_{i+1} + \beta_i q_i, \\ p_{i+1} &= u_{i+1} + \beta_i(q_i + \beta_i p_i) \end{aligned}$$

pro $1 \le i \le n$, nazveme umocněnou metodou sdružených gradientů (CGS) určenou maticí $A \in R^{n\times n}$ a vektory $f \in R^n$, $\tilde{f} \in R^n$.

Poznámka 387. Jsou-li splněny předpoklady věty 272 platí

$$\|\overline{f}_i\| = \|\varphi_i^2(A)f\| \le \|\varphi_i(A)\|\|\varphi_i(A)f\| = \|\varphi_i(A)\|\|f_i\|,$$

$1 \le i \le n+1$, takže metoda CGS najde řešení soustavy rovnic $As+f = 0$ po nejvýše n krocích ($\|f_{n+1}\| = 0$ podle věty 272).

Výhodou metody CGS je to, že nepoužívá transponovanou matici, což je nutné pro konstrukci diferenčních verzí nepřímé Newtonovy metody, kdy se násobení $J(x)v$ nahrazuje diferencí $(f(x+\delta v)-f(x))/\delta$. Nevýhodou metody CGS (stejně jako metody BCG) je to, že není založena na žádném minimalizačním principu. Normy reziduí nemají monotonní průběh a mohou dosti silně oscilovat. Proto se používají další úpravy metody CGS založené na zhlazení norem reziduí.

Lemma 118. *Nechť $\overline{f}_i$, $i \in N$, je posloupnost reziduí určená metodou CGS. Nechť $f_1 = \overline{f}_1$ a*

$$\begin{aligned} \lambda_i &= -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2}, \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}), \end{aligned}$$

$1 \le i \le n$. Pak platí

$$\lambda_i = \arg\min_{\lambda\in R} \|\overline{f}_{i+1} + \lambda(f_i - \overline{f}_{i+1})\|,$$

$1 \le i \le n$, takže $\|f_{i+1}\| \le \|f_i\|$ (normy reziduí monotonně klesají) a $\|f_{i+1}\| \le \|\overline{f}_{i+1}\|$ (řešení je nalezeno po nejvýše n krocích).

Důkaz Zřejmě pro $1 \le i \le n$ platí

$$\|f_{i+1}\|^2 = \|\overline{f}_{i+1}\|^2 + 2\lambda_i\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1}) + \lambda_i^2\|f_i - \overline{f}_{i+1}\|^2,$$

Tato kvadratická funkce nabývá minima pro $\lambda_i = -\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})/\|f_i - \overline{f}_{i+1}\|^2$. □

Rekurentní vztahy pro f_i (lemma 118) spolu s odpovídajícími rekurentními vztahy pro s_i jsou základem jednoduše zhlazené metody CGS.

Definice 90. *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$, $\tilde{f} \in R^n$. Pak iterační proces*

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

a

$$\begin{gathered}
v_i = Ap_i, \quad \alpha_i = \tilde{f}^T f_i / \tilde{f}^T v_i, \\
q_i = u_i - \alpha_i v_i, \\
\overline{s}_{i+1} = \overline{s}_i - \alpha_i (u_i + q_i), \\
\overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T f_{i+1} / \tilde{f}^T \overline{f}_i, \\
u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\
p_{i+1} = u_{i+1} + \beta_i (q_i + \beta_i p_i), \\
\lambda_i = -\frac{\overline{f}_{i+1}^T (f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2}, \\
s_{i+1} = \overline{s}_{i+1} + \lambda_i (s_i - \overline{s}_{i+1}), \\
f_{i+1} = \overline{f}_{i+1} + \lambda_i (f_i - \overline{f}_{i+1})
\end{gathered}$$

pro $1 \le i \le n$, nazveme jednoduše zhlazenou metodou CGS (SSCGS) určenou maticí $A \in R^{n\times n}$ a vektory $f \in R^n$, $\tilde{f} \in R^n$.

Ačkoliv normy reziduí jednoduše zhlazené metody CGS mají monotonní průběh, pro konstrukci metod s lokálně omezeným krokem je vhodnější dvojnásobně zhlazená metoda CGS.

Definice 91. *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$, $\tilde{f} \in R^n$. Pak iterační proces*

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

a

$$\begin{gathered}
v_i = Ap_i, \quad \alpha_i = \tilde{f}^T \overline{f}_i / \overline{f}^T v_i, \\
q_i = u_i - \alpha_i v_i, \\
\overline{s}_{i+1} = \overline{s}_i - \alpha_i (u_i + q_i), \\
\overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T \overline{f}_{i+1} / \tilde{f}^T \overline{f}_i, \\
u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\
p_{i+1} = u_{i+1} + \beta_i (q_i + \beta_i p_i), \\
[\lambda_i, \mu_i]^T = \arg \min_{[\lambda,\mu]^T \in R^2} \|\overline{f}_{i+1} + \lambda (f_i - \overline{f}_{i+1}) + \mu v_i\|, \\
s_{i+1} = \overline{s}_{i+1} + \lambda_i (s_i - \overline{s}_{i+1}) + \mu_i p_i, \\
f_{i+1} = \overline{f}_{i+1} + \lambda_i (f_i - \overline{f}_{i+1}) + \mu_i v_i
\end{gathered}$$

pro $1 \le i \le n$, nazveme dvojnásobně zhlazenou metodou CGS (DSCGS) určenou maticí $A \in R^{n\times n}$ a vektory $f \in R^n$, $\tilde{f} \in R^n$.

Poznámka 388. Vektor $[\lambda_i, \mu_i]^T$ realizující minimum normy $\|f_{i+1}\|$ můžeme určit podle vzorce

$$\begin{bmatrix} \lambda_i \\ \mu_i \end{bmatrix} = -(V_i^T V_i)^{-1} V_i^T \overline{f}_{i+1},$$

kde $V_i = [f_i - \overline{f}_{i+1}, v_i] \in R^{n\times 2}$ (odvození tohoto vzorce je analogické odvození vzorce pro λ_i v lemmatu 118). Dosadíme-li toto vyjádření do vztahu pro f_{i+1}, dostaneme $f_{i+1} = P_i \overline{f}_{i+1}$, kde $P_i = I - V_i(V_i^T V_i)^{-1} V_i^T$ je matice ortogonální projekce do podprostoru generovaného vektory $f_i - \overline{f}_{i+1}$ a v_i.

Metody CGS, SSCGS, DSCGS lze modifikovat tak, že se používá předpodmínění. Vzhledem k tomu, že při nepřesném řešení soustavy rovnic $As + f = 0$ nás zajímá reziduum $As + f$, používá se pravé předpodmínění, což znamená, že se řeší soustava rovnic $AC^{-1}\hat{s} + f = 0$ s předpodmíňovací maticí C^{-1} a pak se pokládá $s = C^{-1}\hat{s}$. Jelikož úpravy metod CGS, SSCGS, DSCGS jsou prakticky stejné uvedeme pouze předpodmíněnou verzi metody DSCGS, která používá rekurentní vztahy

$$s_1 = 0, \quad \overline{s}_1 = 0, \quad f_1 = f, \quad \overline{f}_1 = f, \quad p_1 = f, \quad u_1 = f$$

a

$$\begin{gathered}
v_i = AC^{-1}p_i, \quad \alpha_i = \tilde{f}^T \overline{f}_i / \tilde{f}^T v_i, \\
q_i = u_i - \alpha_i v_i, \\
\overline{s}_{i+1} = \overline{s}_i + \alpha_i C^{-1}(u_i + q_i), \\
\overline{f}_{i+1} = \overline{f}_i + \alpha_i AC^{-1}(u_i + q_i), \quad \beta_i = \tilde{f}^T \overline{f}_{i+1} / \tilde{f}^T \overline{f}_i, \\
u_{i+1} = \overline{f}_{i+1} + \beta_i q_i, \\
p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i), \\
[\lambda_i, \mu_i]^T = -(V_i^T V_i)^{-1} V_i^T \overline{f}_{i+1}, \\
s_{i+1} = \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}) + \mu_i C^{-1} p_i, \\
f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) + \mu_i v_i
\end{gathered}$$

pro $1 \le i \le n$, (zde $V_i = [f_i - \overline{f}_{i+1}, v_i] \in R^{n\times 2}$).

Předpodmiňovací matice se obvykle volí tak, aby platilo $C \approx A$. Pak matice $AC^{-1} \approx I$ je lépe podmíněná. Velmi účinné je předpodmiňování pomocí neúplného trojúhelníkového rozkladu

$$P(A + E) = LU,$$

kde L je dolní trojúhelníková matice s jednotkami na hlavní diagonále, U je horní trojúhelníková matice, P je permutační matice a E je matice zahrnující vliv potlačování nově vznikajících nenulových prvků. Permutační matice se volí tak, aby matice PA měla nenulové prvky (pivoty) na hlavní diagonále.

Nyní se budeme zabývat metodou GMRES, která patří mezi metody s dlouhými rekurentními vztahy. Princip metody GMRES spočívá v tom, že se generují ortogonální vektory q_i, $1 \le i \le n$, tak, že q_j $1 \le j \le i$, tvoří bázi v Krylovově podprostoru $\mathcal{K}_i$. Vektor $s_{i+1} \in R^n$ se volí tak, aby platilo

$$s_{i+1} = \arg\min_{s\in\mathcal{K}_i} \|As + f\|. \tag{1063}$$

Metoda GMRES je tedy založena na minimalizačním principu, což znamená, že normy reziduí monotonně klesají.

Ortonormální vektory q_i, $1 \le i \le n$ se generují pomocí Gramova-Schmidtova ortogonalizačního procesu. Klasický Gramův-Schmidtův ortogonalizační proces používá rekurentní vztahy

$$\beta_1 q_1 = f$$

a

$$\left.\begin{array}{c} q_{i+1}^1 = Aq_i, \\ \alpha_{ji} = q_j^T q_{i+1}^1, \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \end{array}\right\} 1 \le j \le i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$

$1 \le i \le n-1$, kde koeficienty β_i, $1 \le i \le n$ se vybírají tak, aby vektory q_i, $1 \le i \le n$, měly jednotkovou normu. Stabilnější je modifikovaný Gramův-Schmidtův ortogonalizační proces

$$\beta_1 q_1 = f$$

a

$$\left.\begin{array}{c} q_{i+1}^1 = Aq_i, \\ \alpha_{ji} = q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \end{array}\right\} 1 \le j \le i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$

$1 \le i \le n-1$. Gramův-Schmidtův ortogonalizační proces generující ortonormální báze Krylovových podprostorů $\mathcal{K}_i$, $1 \le i \le n$, se také nazývá Arnoldiovým procesem určeným maticí $A \in R^{n\times n}$ a vektorem $f \in R^n$.

Označíme-li $Q_i = [q_1, q_2, \ldots, q_i]$ a

$$H_i = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1i} \\ \beta_2 & \alpha_{22} & \ldots & \alpha_{2i} \\ 0 & \beta_3 & \ldots & \alpha_{3i} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \beta_{i+1} \end{bmatrix}$$

($H_i \in R^{(i+1)\times i}$ je horní Hessenbergova matice), můžeme Arnoldiův proces zapsat v maticovém tvaru

$$AQ_i = Q_{i+1} H_i.$$

Položíme-li $s_{i+1} = Q_i z_i$, kde $z_i \in R^n$, platí

$$\|As_{i+1} + f\| = \|AQ_i z_i + f\| = \|Q_{i+1} H_i z_i + Q_{i+1}(\beta_1 e_1)\| = \|H_i z_i + \beta_1 e_1\|,$$

takže minimalizační podmínku můžeme zapsat ve tvaru

$$z_i = \arg\min_{z\in R^i} \|H_i z + \beta_1 e_1\|. \tag{1064}$$

Věta 273. *Nechť $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \le j \le i$, $\mathcal{K}_i = \mathcal{K}_{i+1}$ a nechť platí (1063). Pak $As_{i+1} + f = 0$.*

Důkaz Uvažujme Arnoldiův proces určený regulární maticí $A \in R^{n\times n}$ a vektorem f. Jestliže $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \le j \le i$ a $\mathcal{K}_i = \mathcal{K}_{i+1}$, pak vektory q_i, $1 \le j \le i$, jsou lineárně nezávislé a $\beta_{i+1} = 0$. Platí tedy

$$AQ_i = Q_i \overline{H}_i,$$

kde $\overline{H}_i \in R^{i\times i}$ je horní Hessenbergova matice, která vznikne z matice $H_i \in R^{(i+1)\times i}$ vyškrtnutím posledního řádku. Jelikož matice AQ_i má lineárně nezávislé sloupce a A je regulární), je matice $\overline{H}_i$ regulární a existuje řešení soustavy rovnic $\overline{H}_i z_i + \beta_1 e_1 = 0$. Položíme-li $s_{i+1} = Q_i z_i$ platí

$$\|As_{i+1} + f\| = \|\overline{H}_i z_i + \beta_1 e_1\| = 0.$$

□

Metoda GMRES nalezne řešení soustavy rovnic $As + f = 0$ po nejvýše n krocích. Jestliže totiž $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j < n$, pak nutně $\mathcal{K}_n = \mathcal{K}_{n+1} = R^n$. Metoda GMRES nemůže selhat, neboť $\beta_{i+1} = 0$ implikuje $As_{i+1} + f = 0$.

Abychom mohli určit vektor z_i vyhovující podmínce (1064), je třeba provést ortogonální rozklad

$$P_i(H_i z_i + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z_i + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix},$$

kde $P_i = \overline{P}_i \overline{P}_{i-1} \dots \overline{P}_1$ je součin Givensových matic elementárních rotací a

$$R_i = \begin{bmatrix} \rho_{11} & \rho_{12} & \dots & \rho_{1i} \\ 0 & \rho_{22} & \dots & \rho_{2i} \\ 0 & 0 & \dots & \rho_{ii} \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_i \end{bmatrix}.$$

Je to postup, který byl již použit v metodě LSQR (oddíl 10.7), proto ho nebudeme znovu odvozovat. Uvedeme pouze výsledné rekurentní vztahy metody GMRES.

Definice 92. *Nechť $A \in R^{n \times n}$ je regulární matice a $f \in R^n$. Pak iterační proces*

$$\beta_1 q_1 = f, \qquad \overline{\eta}_1 = \beta_1$$

a

$$\begin{gathered} q_{i+1}^1 = A q_i, \\ \overline{\alpha}_{1i} = q_1^T q_{i+1}^1, \qquad q_{i+1}^2 = q_{i+1}^1 - \overline{\alpha}_{1i} q_1, \\ \left.\begin{aligned} \alpha_{ji} = q_j^T q_{i+1}^j, \qquad q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \\ \rho_{j-1i} = \lambda_{j-1} \overline{\alpha}_{j-1i} + \tau_{j-1} \alpha_{ji}, \\ \overline{\alpha}_{ji} = -\lambda_{j-1} \overline{\alpha}_{j-1i} + \tau_{j-1} \alpha_{ji}, \end{aligned} \right\} 1 < j \leq i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \\ \rho_{ii} = \sqrt{\overline{\alpha}_{ii}^2 + \beta_{i+1}^2}, \\ \lambda_i = \frac{\overline{\alpha}_{ii}}{\rho_{ii}}, \qquad \tau_i = \frac{\beta_{i+1}}{\rho_{ii}}, \\ \eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i, \end{gathered}$$

$1 \leq i \leq n$, nazveme metodou GMRES určenou maticí $A \in R^{n \times n}$ a vektorem $f \in R^n$.

Používáme-li metodu GMRES, můžeme minimalizační podmínku přepsat ve tvaru

$$z_i = \arg \min_{z \in R^n} \left\| \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix} \right\|.$$

Platí tedy $R_i z_i + h_i = 0$ (matice R_i je horní trojúhelníková) a položíme-li $s_{i+1} = Q_i z_i$, platí $\| As_{i+1} + f \| = |\overline{\eta}_{i+1}|$. Čísla $|\overline{\eta}_i|$, $1 \leq i \leq n+1$, jsou tedy normy reziduí $f_i = As_i + f$, $1 \leq i \leq n+1$. Jakmile metoda GMRES získá dostatečně malé rezidium (platí-li $|\overline{\eta}_{i+1}| \leq \overline{\omega} \|f\|$), můžeme proces ukončit a položit $s_{i+1} = Q_i z_i$, kde $R_i z_i + h_i = 0$.

Metodu GMRES můžeme různým způsobem modifikovat. Generujeme-li ortonormální bázi v posunutých Krylovových podprostorech

$$A\mathcal{K}_i = \operatorname{span}\{Af, \dots, A^i f\},$$

odpadne použití ortogonálního rozkladu. Vektory q_j, $1 \leq j \leq i$ se opět určují pomocí Gramova-Schmidtova ortogonalizačního procesu, takže platí

$$AQ_{i-1} = Q_i H_{i-1},$$

kde $H_{i-1} \in R^{i\times(i-1)}$ je horní Hessenbergova matice. Zvolíme-li vektor q_1 tak, že $\beta_1 q_1 = Af$, můžeme psát

$$[Af, AQ_{i-1}] = Q_i[\beta_1 e_1, H_{i-1}] = Q_i R_i,$$

kde

$$R_i = \begin{bmatrix} \beta_1 & \alpha_{11} & \alpha_{12} & \dots & \alpha_{1i-1} \\ 0 & \beta_2 & \alpha_{22} & \dots & \alpha_{2i-1} \\ 0 & 0 & \beta_3 & \dots & \alpha_{3i-1} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \beta_i \end{bmatrix}$$

($R_i \in R^{i\times i}$ je horní trojúhelníková matice). Položíme-li

$$s_{i+1} = [f, Q_{i-1}]z_i,$$

platí $s_{i+1} \in \mathcal{K}_i$, neboť vektory f a q_j, $1 \leq j \leq i-1$, jsou lineárně nezávislé. Dále platí

$$\|As_{i+1} + f\| = \|[Af, AQ_{i-1}]z_i + f\| = \|Q_i R_i z_i + f\|,$$

takže minimalizační podmínku můžeme zapsat ve tvaru

$$z_i = \arg\min_{z\in R^i} \|Q_i R_i z + f\|.$$

Normální soustava rovnic pro tento problém nejmenších čtverců má tvar $R_i^T Q_i^T Q_i R_i z_i + R_i^T Q_i^T f = 0$, takže

$$R_i z_i + Q_i^T f = 0,$$

což po dosazení do vzorce pro reziduum dává

$$f_{i+1} = As_{i+1} + f = (I - Q_i Q_i^T)f = f_i - q_i q_i^T f.$$

Jelikož z ortogonality plyne $q_i^T Q_{i-1} = 0$, můžeme psát $q_i^T f_i = q_i^T (I - Q_{i-1}Q_{i-1}^T)f = q_i^T f$, což dává

$$f_{i+1} = f_i - q_i q_i^T f_i.$$

Tento vzorec zlepšuje stabilitu modifikované metody GMRES. Shrneme-li dosažené výsledky, můžeme modifikovanou metodu GMRES definovat takto.

Definice 93. *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$. Pak iterační proces*

$$f_1 = f, \qquad \beta_1 q_1 = Af$$

a

$$\gamma_i = q_i^T f_i,$$
$$f_{i+1} = f_i - \gamma_i q_i,$$
$$q_{i+1}^1 = Aq_i,$$
$$\left.\begin{array}{l} \alpha_{ji} = q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \end{array}\right\} 1 \leq j \leq i,$$

$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$

$1 \le i \le n-1$, *nazveme modifikovanou metodou GMRES určenou maticí* $A \in R^{n\times n}$ *a vektorem* $f \in R^n$.

Jakmile modifikovaná metoda GMRES získá dostatečně malé rezidium (platí-li $\|f_{i+1}\| \le \overline{\omega}\|f\|$), můžeme proces ukončit a položit $s_{i+1} = [f, Q_{i-1}]z_i$, kde

$$\begin{bmatrix} \beta_1 & \alpha_{11} & \dots & \alpha_{1i-1} \\ 0 & \beta_2 & \dots & \alpha_{2i-1} \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \beta_i \end{bmatrix} z_i = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_i \end{bmatrix}$$

Poznámka 389. Základní i modifikovanou metodu GMRES lze snadno předpodmiňovat (používá se pravé předpodmínění). V tomto případě se místo matice A používá matice AC^{-1} a vektor $s_{i+1} \in R^n$ se určuje podle vzorce

$$s_{i+1} = -C^{-1}Q_i R_i^{-1} h_i$$

(základní metoda) nebo

$$s_{i+1} = -C^{-1}[f, Q_{i-1}]R_i^{-1}Q_i^T f$$

(modifikovaná metoda). Předpodmiňovací matice C^{-1} se opět volí tak, aby platilo $C \approx A$ (použije se například neúplný LU rozklad).

12.9 Metody s lokálně omezeným krokem

Poznámka 390. Zhlazenou metodu CGS nebo metodu GMRES můžeme použít ke konstrukci nepřesných metod s lokálně omezeným krokem. V tomto případě se generuje posloupnost vektorů $s_{i+1} \in R^n$, $1 \le i \le n$, které aproximují řešení soustavy rovnic $As + f = 0$, a pak se pokládá $s = s_{i+1}$, pokud $\|s_{i+1}\| \le \Delta$ a $\|f_{i+1}\| \le \overline{\omega}\|f\|$, $0 \le \overline{\omega} < 1$, nebo $s = s_i + \alpha_i(s_{i+1} - s_i)$ a $\|s\| = \Delta$, pokud $\|s_j\| < \Delta$, $\|f_j\| > \overline{\omega}\|f\|$, $1 \le j \le i$, a $\|s_{i+1}\| \ge \Delta$. Tato volba zřejmě splňuje podmínky ($\overline{\text{T1a}}$), ($\overline{\text{T1b}}$) metody s lokálně omezeným krokem (definice 81). Navíc je třeba zformulovat předpoklady, aby byla splněna i podmínka ($\overline{\text{T1c}}$), neboli

$$\|f\| - \|As + f\| \ge 2\underline{\sigma}\|As\|),$$

kde $\underline{\sigma}$ je nějaká konstanta. V dalším textu budeme předpokládat, že matice A splňuje podmínku $\|I - A\| \le \overline{\nu} < 1$, což lze docílit vhodným předpodmíněním (místo matice A se používá matice AC^{-1} taková, že $\|I - AC^{-1}\| \le \overline{\nu} < 1$).

Lemma 119. *Nechť* $\|I - A\| \le \overline{\nu} < 1$ *a nechť* $s_{i+1} \in R^n$, $i = 1, \dots, n$, *jsou vektory generované metodou GMRES nebo dvojnásobně zhlazenou metodou CGS. Pak*

$$\|f\|^2 - \|r_{i+1}\|^2 \ge \underline{\eta}^2\|f\|^2,$$

kde $\underline{\eta} = (1 - \overline{\nu})/(1 + \overline{\nu})$.

Důkaz (a) Nejprve ukážeme, že

$$|f^T Af| \ge \frac{1 - \overline{\nu}}{1 + \overline{\nu}}\|f\|\|Af\| = \underline{\eta}\|f\|\|Af\|.$$

Podle předpokladu platí

$$\begin{aligned} |f^T Af| &= |f^T f - f^T(I - A)f| \ge |f^T f| - |f^T(I - A)f| \\ &\ge \|f\|^2 - \|I - A\|\|f\|^2 \ge (1 - \overline{\nu})\|f\|^2 \end{aligned}$$

a

$$\|Af\| \leq \|f\| + \|I - A\|\|f\| \leq (1 + \overline{\nu})\|f\|,$$

což dohromady dává dokazovanou nerovnost.

(b) Protože posloupnost norem reziduí metody GMRES i dvojnásobně zhlazené metody CGS je nerostoucí, stačí dokázat, že

$$\|f\|^2 - \|r_2\|^2 \geq \underline{\eta}^2\|f\|^2.$$

Uvažejme nejprve metodu GMRES. Jelikož $s_1 = 0$ a $\mathcal{K}_1 = \text{span}\{f\}$, platí

$$\|r_2\| = \min_{\mu \in R} \|A(\mu f) + f\|.$$

Z podmínky optimality

$$\mu_1 \stackrel{\Delta}{=} \arg\min_{\mu \in R} \|A(\mu f) + f\|^2 = \arg\min_{\mu \in R} (\mu^2 \|Af\|^2 + 2\mu f^T Af + \|f\|^2)$$

dostaneme $\mu_1 = -f^T Af / \|Af\|^2$, takže pro normu residua r_2 platí

$$\|r_2\|^2 = \frac{(f^T Af)^2}{\|Af\|^4}\|Af\|^2 - 2\frac{(f^T Af)^2}{\|Af\|^2} + \|f\|^2 = \|f\|^2 - \frac{(f^T Af)^2}{\|Af\|^2\|f\|^2}\|f\|^2.$$

Tato nerovnost spolu s (a) dokazuje tvrzení lemmatu pro metodu GMRES. Uvažujme nyní dvojnásobně zhlazenou metodu CGS. Pak platí

$$\|r_2\| = \min_{[\lambda,\mu]^T \in R^2} \|\overline{r}_2 + \lambda(f - \overline{r}_2) + \mu v_1\| \leq \min_{\mu \in R} \|f + \mu v_1\| = \min_{\mu \in R} \|f + \mu Af\|$$

(po dosazení $\lambda = 1$) což dává stejný výsledek jako v případě metody GMRES. □

Lemma 120. *Nechť jsou splněny předpoklady lemmatu 119 a nechť $s \in \mathcal{R}^n$ je vektor určený metodou GMRES nebo dvojnásobně zhlazenou metodou CGS tak jako v poznámce 390. Pak platí*

$$\|f\| - \|As + f\| \geq 2\underline{\sigma}\|As\|,$$

kde $2\underline{\sigma} = \underline{\eta}^2/8$.

Důkaz (a) Nechť $\|s_{i+1}\| < \Delta$ a $\|r_{i+1}\| \leq \overline{\omega}\|f\|$. Pak podle lemmatu 119 platí

$$2\|f\|\left(\|f\| - \|r_{i+1}\|\right) \geq \|f\|^2 - \|r_{i+1}\|^2 \geq \underline{\eta}^2\|f\|^2,$$

což dohromady z odhadem $\underline{A}\|s\| \leq \|As\| \leq 2\|f\|$ dává

$$\|f\| - \|r_{i+1}\| \geq \frac{1}{2}\underline{\eta}^2\|f\| \geq \frac{1}{4}\underline{\eta}^2\|As\|.$$

(b) Nechť $\|s_{i+1}\| \geq \Delta$ a $i > 1$. Pak platí $s = \tau_i s_{i+1} + (1 - \tau_i)s_i$ s $0 < \tau_i \leq 1$, takže

$$\|As + f\| = \|\tau_i(As_{i+1} + f) + (1 - \tau_i)(As_i + f)\| \leq \tau_i\|r_{i+1}\| + (1 - \tau_i)\|r_i\|$$

a lemma 119 spolu s odhadem $\underline{A}\|s\| \leq \|As\| \leq 2\|f\|$ dává

$$\|f\| - \|As + f\| \geq \tau_i\left(\|f\| - \|r_{i+1}\|\right) + (1 - \tau_i\left(\|f\| - \|r_i\|\right) \geq \frac{1}{2}\underline{\eta}^2\|f\| \geq \frac{1}{4}\underline{\eta}^2\|As\|.$$

(c) Nechť $\|s_{i+1}\| \geq \Delta$ a $i = 1$. Pak platí $s = \tau_1 s_2$, kde $0 < \tau_1 \leq 1$. Můžeme tedy psát

$$\begin{aligned}
\|f\|^2 - \|As + f\|^2 &= \|f\|^2 - \tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 - \|f\|^2 \\
&= -\tau_1^2\|As_2\|^2 - 2\tau_1 f^T As_2 \geq \tau_1\left(-\|As_2\|^2 - 2f^T As_2\right) \\
&= \tau_1\left(\|f\|^2 - \|As_2 + f\|^2\right)
\end{aligned}$$

(neboť $\tau_1^2 \leq \tau_1$ pro $0 < \tau_1 \leq 1$), nebo

$$\begin{aligned} 2\|f\|\left(\|f\| - \|As+f\|\right) &\geq \|f\|^2 - \|As+f\|^2 \geq \tau_1\left(\|f\|^2 - \|r_2\|^2\right) \\ &\geq \tau_1\|f\|\left(\|f\| - \|r_2\|\right), \end{aligned}$$

takže

$$\|f\| - \|As+f\| \geq \frac{1}{2}\tau_1\left(\|f\| - \|r_2\|\right) \geq \frac{1}{4}\tau_1\underline{\eta}^2\|f\|$$

jako v případě (a). Platí tedy

$$2\|f\| \geq \|r_2 - f\| = \|As_2\|,$$

což po dosazení do předchozí nerovnosti dává

$$\|f\| - \|As+f\| \geq \frac{1}{8}\tau_1\underline{\eta}^2\|As_2\| = \frac{1}{8}\underline{\eta}^2\|As\|.$$

□

Věta 274. *Nechť $\|I - A_i\| \leq \overline{\nu} < 1$, $i \in N$ a nechť $s_i \in R^n$, $i \in N$, jsou směrové vektory určené metodou GMRES nebo dvojnásobně zhlazenou metodou CGS tak jako v poznámce 390. Pak jsou splněny podmínky (*$\overline{\text{T1a}}$*)–(*$\overline{\text{T1c}}$*) a směrové vektory $s_i \in R^n$, $i \in N$, můžeme použít ke konstrukci nepřesné metody s lokálně omezeným krokem. Aplikujeme-li tuto metodu na funkci $f : \mathcal{D} \to R^n$ vyhovující předpokladům J1 a J4–J6 a splňují-li matice A_i, $i \in N$ podmínky uvedené v lemmatu 102, platí $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 120 a věty 238. □

Metodu GMRES nebo dvojnásobně zhlazenou metodu CGS můžeme také použít ke konstrukci metod, které se nazývají metodami psí nohy. V tomto případě se generují vektory $s_{i+1} \in R^n$, $1 \leq i \leq m$, kde $m \ll n$ (obvykle $1 \leq m \leq 3$). Jestliže pro nějaký index $1 \leq i \leq m$ platí $\|s_{i+1}\| \leq \Delta$ a $\|f_{i+1}\| \leq \overline{\omega}\|f\|$, $0 \leq \overline{\omega} < 1$, pokládáme $s = s_{i+1}$. Jestliže pro nějaký index $1 \leq i \leq m$ platí $\|s_j\| < \Delta$, $\|f_j\| > \overline{\omega}\|f\|$, $1 \leq j \leq i$, a $\|s_{i+1}\| \geq \Delta$, pokládáme $s = s_i + \alpha_i(s_{i+1} - s_i)$ tak, že $\|s\| = \Delta$. Nenastane-li ani jeden z těchto případů určíme pomocí některé přímé eliminační metody řešení $s^* \in R^n$ soustavy rovnic $As + f = 0$ a pokládáme $s = s_{m+1} + \alpha_{m+1}(s^* - s_{m+1})$. Jednoduše se dá ukázat (podobně jako v důkazu lemmatu 120), že pokud platí $\Delta \geq \underline{\gamma}\|f\|$ nebo $|f^T Af| \geq \underline{\varepsilon}\|f\|\|Af\|$, je splněna podmínka ($\overline{\text{T1c}}$).

12.10 Numerické porovnání

K testování metod pro řešení rozsáhlých systémů nelineárních rovnic bylo použito 44 úloh, obsahujících 1000 rovnic o 1000 neznámých, ze sbírky TEST18 zmíněné v oddílu 1.5 a popsané v práci [112]. Nejprve porovnáme účinnost metod, které používají analytické vyjádření pro prvky Jacobiovy matice a to Newtonovu metodu pro nelineární rovnice `LSNMJ-xx` s metodami `TRGN-xx`, `TRGNS-xx`, `TRNMS-xx` a `LSVMP-xx`, testovanými v oddílu 10.8 (tabulka 11) aplikovanými na součet čtverců (949). Metody, jejichž označení začíná písmeny TR jsou realizovány jako metody s lokálně omezeným krokem a písmena LS označují metody spádových směrů. Přitom číslice xx udávají číslo použitého algoritmu. U metod pro řešení systémů nelineárních rovnic jsou to buď písmena LU, označující přímou metodu používající řídký trojúhelníkový rozklad nesymetrické matice, nebo písmena CGS, označující dvojnásobně zhlazenou metodu CGS (definice 91), nebo písmena GMR, označující metodu GMRES (definice 92) přerušovanou vždy po 50 krocích.

Metoda	NIT	NFV	NFG	NCG	selhání	čas
LSNMJ-LU	625	1619	625	–	2	3.15
LSNMJ-CGS	610	879	610	1538	–	2.49
LSNMJ-GMR	609	897	609	1521	1	2.09
TRGN-11	2511	2899	2555	–	1	17.22
TRGNS-11	4188	4875	4232	–	2	23.30
TRNMS-11	5228	53685	52773	–	11	72.05
LSVMP-3	24876	48852	48852	–	10	212.77

Tabulka 15: TEST18 – 44 úloh

Tabulka 15 obsahuje celkový počet iterací NIT, celkový počet použitých funkčních hodnot NFV, celkový počet použitých gradientů NFG, celkový počet iterací NCG zhlazené metody CGS nebo metody GMRES, počet selhání a celkový čas výpočtu. Selhání bylo dvojího druhu. Buď bylo naleno lokální minimum funkce (949), které nebylo řešením dané úlohy, nebo byl překročen maximální počer iterací.

Z výsledků uvedených v této tabulce lze vyvodit několik závěrů:

- Newtonova metoda pro řešení soustavy nelineárních rovnic je mnohem efektivnější než Gaussova-Newtonova metoda aplikovaná na součet čtverců (949) nebo její modifikace. Gaussovu-Newtonovu metodu není třeba modifikovat, neboť řešíme úlohy s nulovými rezidui. Zcela nevhodné jsou metody pro minimalizaci obecné účelové funkce.
- Iterační určování směrového vektoru se zdá být výhodnější než přímé použití LU rozkladu.

Nyní porovnáme účinnost metod, které používají pouze hodnoty dílčích funkcí. V tabulce 16 jsou uvedeny výsledky získané těmito metodami:

LSNM1-xx - Diferenční verze Newtonovy metody pro nelineární rovnice s aproximací řádků Jacobiovy matice podle vzorců (1044),

LSNM2-xx - Diferenční verze Newtonovy metody pro nelineární rovnice s aproximací sloupců Jacobiovy matice podle vzorců (1043),

LSNMS-xx - metoda škálování řádků v LU rozkladu Jacobiovy matice podle vztahů (1059),

LSNMC-xx - nedokonalá diferenční verze Newtonovy metody popsaná v oddílu 12.7,

LSQNS-xx - Schubertova kvazinewtonovská metoda používající vzorec (1045),

LSQLB-xx - Broydenova dobrá metoda s omezenou pamětí používající vzorec (1038), kde $v_i = d_i$, $l < i \leq l + m - 1$ a $m = 5$ (poznámka 379),

LSQLC-xx - metoda aktualizace sloupců s omezenou pamětí používající vzorec (1038), kde $v_i = e_k$, $l < i \leq l + m - 1$ a $m = 5$ (poznámka 379),

LSQLI-xx - inverzní metoda aktualizace sloupců s omezenou pamětí používající vzorec (1036), kde $z_i = e_k$, $l < i \leq l + m - 1$ a $m = 5$ (věta 263 a poznámka 381).

Všechny uvedené metody jsou spádovými metodami (metody s lokálně omezeným krokem dosahovaly horších výsledků). Místo znaků xx dosazujeme písmena LU, CGS, GMR tak jako v předchozí tabulce.

Metoda	NIT	NFV	NCG	selhání	čas
LSNM1-LU	570	2942	–	1	3.27
LSNM1-CGS	576	2988	1398	–	3.36
LSNM1-GMR	575	2983	1372	1	2.99
LSNM2-LU	570	3105	–	1	3.33
LSNM2-CGS	576	3157	1398	–	3.41
LSNM2-GMR	575	3152	1372	1	3.04
LSNMS-LU	848	3653	–	1	3.18
LSNMC-LU	672	6028	–	1	6.56
LSQNS-LU	842	2213	–	1	3.38
LSQNS-CGS	876	2114	2550	-	3.38
LSQNS-GMR	1068	2243	12972	2	8.22
LSQLB-CGS	756	2012	2073	1	3.12
LSQLC-CGS	770	2255	2022	-	3.19
LSQLI-LU	747	1667	–	1	1.88
LSQLI-CGS	972	2328	1565	–	2.83

Tabulka 16: TEST18 – 44 úloh

Z výsledků uvedených v této tabulce lze vyvodit několik závěrů:

- Diferenční verze Newtonovy metody jsou velmi efektivní. Řádkový i sloupcový způsob výpočtu diferencí dávají prakticky stejné výsledky. Jelikož použití vzorců (1044) je algoritmicky jednodušší, byl tento postup použit ve všech dalších úpravách Newtonovy metody.
- Iterační určování směrového vektoru pomocí dvojnásobně zhlazené metody CGS se zdá být výhodnější než přímé použití LU rozkladu. Metoda GMRES přerušovaná vždy po 50 krocích je méně robustní (metodu GMRES je třeba přerušovat, neboť vyžaduje uchovávání všech předchozích vektorů).
- Metody s proměnnou metrikou s omezenou pamětí LSQLB-xx, LSQLC-xx a zejména LSQLI-xx dávají nejlepší výsledky.

13 Optimalizace dynamických systémů

Uvažujeme úlohu spočívající v minimalizaci účelové funkce

$$F(x) = \int_{t_0}^{t_1} F_A(x, y(x,t), t)\,\mathrm{d}t + F_T(x, y(x,t_1)), \tag{1065}$$

kde

$$\frac{\mathrm{d}y(x,t)}{\mathrm{d}t} = f_S(x, y(x,t), t), \qquad y(x,t_0) = f_I(x)\ .$$

Přitom $x \in R^n$, $y : R^n \times [t_0, t_1] \to R^{n_S}$, $F : R^n \to R$, $F_A : R^n \times R^{n_S} \times [t_0, t_1] \to R$, $F_T : R^n \times R^{n_S} \to R$, $f_S : R^n \times R^{n_S} \times [t_0, t_1] \to R^{n_S}$, $f_I : R^n \to R^{n_S}$. Minimalizovaná funkce je tedy integrálem, ve kterém vystupuje řešení soustavy obyčejných diferenciálních rovnic prvního řádu s počátečními podmínkami. Tuto soustavu diferenciálních rovnic nazýváme stavovým systémem. Z praktických důvodů je výhodné počítat integrál společně s řešením soustavy diferenciálních rovnic. V tomto případě pokládáme

$$F(x) = \tilde{F}_A(x, t_1) + F_T(x, y(x,t_1)), \tag{1066}$$

kde

$$\frac{\mathrm{d}y(x,t)}{\mathrm{d}t} = f_S(x, y(x,t), t), \qquad y(x,t_0) = f_I(x), \tag{1067}$$

$$\frac{\mathrm{d}\tilde{F}_A(x,t)}{\mathrm{d}t} = F_A(x, y(x,t), t), \qquad \tilde{F}_A(x,t_0) = 0. \tag{1068}$$

Celkem se řeší n_S+1 diferenciálních rovnic v přímém směru (začínáme v bodě t_0, kde jsou zadány počáteční podmínky $y(x,t_0) = f_I(x)$ a $\tilde{F}_A(x,t_0) = 0$). Stačí spočítat hodnoty funkcí $y(x,t_1)$ a $F_T(x, y(x,t_1))$ na konci intervalu a tyto hodnoty dosadit do (1066). Je samozřejmě nutné, aby soustava diferenciálních rovnic (1067)–(1068) měla řešení.

Nechť $\mathcal{D} \subset R^n$ je otevřená množina obsahující všechny body $x_i \in R^n$, $i \in N$, generované metodou použitou pro hledání minima funkce F, $y(x,t)$ je řešení systému rovnic (1067) a

$$\mathcal{D}_y = \{y = y(x,t) : x \in \mathcal{D},\ t \in [t_0, t_1]\}.$$

V dalším výkladu budeme používat následující předpoklady

Předpoklad D1. *Funkce $\tilde{F}_A(x,y)$ a $F_T(x,y)$ jsou zdola omezené na $\mathcal{D} \times \mathcal{D}_y$.*

Předpoklad D2. *Existuje spojité a omezené řešení systému (1067)–(1068) na intervalu $[t_0, t_1]$, kdykoliv $x \in \mathcal{D}$.*

Poznámka 391. Předpoklad D2 je poměrně silný, ale v praktických úlohách bývá často splněn. Jestliže pro dané $x \in \mathcal{D}$ je zobrazení $f_S(x, y(x,t), t)$ lipschitzovské podle proměnné y a spojité podle proměnné t, existuje spojité řešení systému (1067)–(1068) na nějakém maximálním intervalu $t_0 \le t < \bar{t}$. Pokud $t_1 < \bar{t}$, existuje spojité a omezemé řešení systému (1067)–(1068) na intervalu $[t_0, t_1]$. V opačném případě takové řešení neexistuje.

Předpoklad D3. *Funkce $F_A(x,y,t)$ a zobrazení $f_S(x,y,t)$ jsou spojité a omezené na intervalu $[t_0, t_1]$, kdykoliv $x \in \mathcal{D}$ a $y \in \mathcal{D}_y$, a jsou spojitě diferencovatelné podle proměnných x a y, přičemž derivace jsou omezené. Funkce $F_T(x,y)$ a zobrazení $f_I(x)$ jsou spojitě diferencovatelné podle svých proměnných, přičemž derivace jsou omezené.*

Předpoklad D4. *Funkce $F_A(x,y,t)$ a zobrazení $f_S(x,y,t)$ jsou spojité a omezené na intervalu $[t_0, t_1]$, kdykoliv $x \in \mathcal{D}$ a $y \in \mathcal{D}_y$, a jsou dvakrát spojitě diferencovatelné podle proměnných x a y, přičemž druhé derivace jsou omezené. Funkce $F_T(x,y)$ a zobrazení $f_I(x)$ jsou dvakrát spojitě diferencovatelné podle svých proměnných, přičemž druhé derivace jsou omezené.*

Poznámka 392. Je-li splněn předpoklad D3, jsou uvedená zobrazení lipschitzovská vzhledem k proměnným $x \in \mathcal{D}$ a $y \in \mathcal{D}_y$. Je-li splněn předpoklad D4, jsou derivace uvedených zobrazení lipschitzovské vzhledem k proměnným $x \in \mathcal{D}$ a $y \in \mathcal{D}_y$. Protože celkový počet funkcí je konečný, budeme používat společnou horní mez $\overline{K}$ a společnou lipschitzovskou konstantu $\overline{L}$.

13.1 Výpočet gradientu

Abychom mohli použít metody sdružených gradientů nebo metody s proměnnou metrikou, je třeba určovat gradienty minimalizované funkce. Gradient funkce $F(x)$ definované vztahy (1066)–(1068) lze určit buď přímým nebo zpětným výpočtem. Aby bylo možné derivovat příslušné vztahy, je třeba splnit předpoklady D2 a D3.

Nejprve popíšeme přímý výpočet gradientu. Abychom zjednodušili zápis, budeme psát y místo $y(x,t)$ (pro hodnoty t_0 a t_1 zápis $y(x,t_0)$ a $y(x,t_1)$ ponecháme). Nechť $u(x,t) = \mathrm{d}y(x,t)/\mathrm{d}x$, takže $u : R^n \times [t_0,t_1] \to R^{n_S \times n}$. Označíme-li $g^T(x) = \mathrm{d}F(x)/\mathrm{d}x$ a $\tilde{g}_A^T(x,t) = \mathrm{d}\tilde{F}_A(x,t)/\mathrm{d}x$, pak derivováním vztahů (1066)–(1068) dostaneme

$$g^T(x) = \tilde{g}_A^T(x,t_1) + \frac{\partial F_T(x,y(x,t_1))}{\partial x} + \frac{\partial F_T(x,y(x,t_1))}{\partial y} u(x,t_1), \tag{1069}$$

kde

$$\frac{\mathrm{d}u(x,t)}{\mathrm{d}t} = \frac{\partial f_S(x,y,t)}{\partial x} + \frac{\partial f_S(x,y,t)}{\partial y} u(x,t), \qquad u(x,t_0) = \frac{\mathrm{d}f_I(x)}{\mathrm{d}x}, \tag{1070}$$

$$\frac{\mathrm{d}\tilde{g}_A^T(x,t)}{\mathrm{d}t} = \frac{\partial F_A(x,y,t)}{\partial x} + \frac{\partial F_A(x,y,t)}{\partial y} u(x,t), \qquad \tilde{g}_A^T(x,t_0) = 0. \tag{1071}$$

K současnému určení hodnoty a gradientu minimalizované funkce je tedy třeba řešit $(n_S+1)(n+1)$ diferenciálních rovnic (1070)–(1071) v přímém směru.

Nyní odvodíme zpětný výpočet gradientu. Nechť $p : [t_0,t_1] \to R^{n_S}$ je libovolné zobrazení (jehož přesný tvar budeme specifikovat později) a $y : R^n \times [t_0,t_1] \to R^{n_S}$ je řešení systému (1067), takže platí $f_S(x,y,t) - \mathrm{d}y(x,t)/\mathrm{d}t = 0$ pro $t \in [t_0,t_1]$. Použijeme-li (1065), můžeme psát

$$F(x) = \int_{t_0}^{t_1} \left[F_A(x,y,t) + p^T(t) \left(f_S(x,y,t) - \frac{\mathrm{d}y(x,t)}{\mathrm{d}t} \right) \right] \mathrm{d}t + F_T(x,y(x,t_1))$$

a použitím pravidla integrování per partes $u^T v' = (u^T v)' - (u^T)' v$, kde $u = p(t)$, $v = y(x,t)$, dostaneme

$$\begin{aligned} F(x) &= \int_{t_0}^{t_1} \left[F_A(x,y,t) + p^T(t) f_S(x,y,t) + \frac{\mathrm{d}p^T(t)}{\mathrm{d}t} y(x,t) \right] \mathrm{d}t \\ &+ p^T(t_0) y(x,t_0) - p^T(t_1) y(x,t_1) + F_T(x,y(x,t_1)). \end{aligned}$$

Nyní můžeme $F(x)$ derivovat podle x, takže

$$\begin{aligned} g^T(x) &= \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial x} \right. \\ &+ \left. \left(\frac{\partial F_A(x,y,t)}{\partial y} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial y} + \frac{\mathrm{d}p^T(t)}{\mathrm{d}t} \right) \frac{\mathrm{d}y(x,t)}{\mathrm{d}x} \right] \mathrm{d}t \\ &+ p^T(t_0) \frac{\mathrm{d}f_I(x)}{\mathrm{d}x} + \frac{\partial F_T(x,y(x,t_1))}{\partial x} + \left(\frac{\partial F_T(x,y(x,t_1))}{\partial y} - p^T(t_1) \right) \frac{\mathrm{d}y(x,t_1)}{\mathrm{d}x}, \end{aligned}$$

neboť podle (1070) platí $dy(x,t_0)/dx = df_I(x)/dx$. Zvolíme-li funkci $p(t) = p(x,t)$ tak, aby se vynulovaly závorky u $\mathrm{d}y(x,t)/\mathrm{d}x$ a $\mathrm{d}y(x,t_1)/\mathrm{d}x$, čili tak, že

$$-\frac{\mathrm{d}p^T(x,t)}{\mathrm{d}t} = \frac{\partial F_A(x,y,t)}{\partial y} + p^T(x,t)\frac{\partial f_S(x,y,t)}{\partial y}, \qquad p^T(x,t_1) = \frac{\partial F_T(x,y(x,t_1))}{\partial y}, \tag{1072}$$

platí

$$g^T(x) = \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(x,t)\frac{\partial f_S(x,y,t)}{\partial x}\right] \mathrm{d}t + p^T(x,t_0)\frac{\mathrm{d}f_I(x)}{\mathrm{d}x} + \frac{\partial F_T(x,y(x,t_1))}{\partial x}. \tag{1073}$$

Dohromady to lze zapsat takto

$$g(x) = \tilde{g}(x,t_0) + \left(\frac{\mathrm{d}f_I(x)}{\mathrm{d}x}\right)^T p(x,t_0), \tag{1074}$$

kde

$$-\frac{\mathrm{d}p(x,t)}{\mathrm{d}t} = \left(\frac{\partial F_A(x,y,t)}{\partial y}\right)^T + \left(\frac{\partial f_S(x,y,t)}{\partial y}\right)^T p(x,t), \qquad p(x,t_1) = \left(\frac{\partial F_T(x,y(x,t_1))}{\partial y}\right)^T, \tag{1075}$$

$$\frac{\mathrm{d}\tilde{g}(x,t)}{\mathrm{d}t} = \left(\frac{\partial F_A(x,y,t)}{\partial x}\right)^T + \left(\frac{\partial f_S(x,y,t)}{\partial x}\right)^T p(x,t), \qquad \tilde{g}(x,t_1) = \left(\frac{\partial F_T(x,y(x,t_1))}{\partial x}\right)^T. \tag{1076}$$

K současnému určení hodnoty a gradientu minimalizované funkce je tedy třeba řešit $n_S + 1$ diferenciálních rovnic (1067)–(1068) v přímém směru a $n_S + n$ diferenciálních rovnic (1075)–(1076) ve zpětném směru.

Poznámka 393. Při zpětném výpočtu gradientu lze postupovat dvojím způsobem.

(1) Zvolíme n_A uzlových bodů $\tilde{t}_1 = t_0 < \tilde{t}_2 < \cdots < \tilde{t}_{n_A} = t_1$ a při řešení v přímém směru ukládáme vektory $y(x,\tilde{t}_i)$, $1 \le i \le n_A$, vypočtené v těchto uzlových bodech. Při řešení ve zpětném směru pak vektory $y(x,t)$, $t \in [t_0,t_1]$, interpolujeme z hodnot $y(x,\tilde{t}_i)$, $1 \le i \le n_A$ a dosazujeme je do výrazů na pravé straně rovnic (1075)–(1076). Je tedy třeba ukládat dalších $n_S n_A$ hodnot v paměti počítače.

(2) Použijeme vektor $y(x,t_1)$ získaný řeršením rovnic (1070)–(1071) jako koncovou podmínku a řešíme ve zpětném směru dalších n_S diferenciálních rovnic

$$\frac{\mathrm{d}y(x,t)}{\mathrm{d}t} = f_S(x,y(x,t),t), \qquad y(x,t_1) = y(x,t_1), \tag{1077}$$

tedy celkem $2n_S + n$ diferenciálních rovnic (1070)–(1071) a (1077). Jako kontrolu přesnosti výpočtu můžeme použít rovnost $y(x,t_0) = f_I(x)$.

13.2 Výpočet Hessovy matice

Abychom mohli použít Newtonovu metodu je třeba určovat Hessovy matice minimalizované funkce. Hessovu matici funkce $F(x)$ definované vztahy (1066)–(1068) lze určit buď přímým nebo zpětným výpočtem. Aby bylo možné dvakrát derivovat příslušné vztahy, je třeba splnit předpoklady D2 a D4.

Nejprve popíšeme přímý výpočet Hessovy matice. Nechť $v(x,t) = \mathrm{d}u(x,t)/\mathrm{d}x = \mathrm{d}^2y(x,t)/\mathrm{d}x^2$, takže $v : R^n \times [t_0,t_1] \to R^{n_S \times n \times n}$ ($v(x,t)$ je tedy tenzorová veličina, která má tři indexy k, i, j, přičemž platí $v_{kij}(x,t) = \partial^2 y_k(x,t)/\partial x_i \partial x_j$). Označíme-li $G(x) = \mathrm{d}^2F(x)/\mathrm{d}x^2$ a $\tilde{G}_A(x) = \mathrm{d}^2\tilde{F}_A(x,t)/\mathrm{d}x^2$, pak derivováním vztahů (1069)–(1071) dostaneme

$$\begin{aligned} G(x) &= \tilde{G}_A(x,t_1) + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x^2} \\ &+ \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x \partial y}u(x,t_1) + u^T(x,t_1)\frac{\partial^2 F_T(x,y(x,t_1))}{\partial y \partial x} \\ &+ u^T(x,t_1)\frac{\partial^2 F_T(x,y(x,t_1))}{\partial y^2}u(x,t_1) + \frac{\partial F_T(x,y(x,t_1))}{\partial y}v(x,t_1), \end{aligned} \tag{1078}$$

kde

$$\begin{aligned}
\frac{\mathrm{d}v(x,t)}{\mathrm{d}t} &= \frac{\partial^2 f_S(x,y,t)}{\partial x^2} + \frac{\partial^2 f_S(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 f_S(x,y,t)}{\partial y \partial x} \\
&+ u^T(x,t) \frac{\partial^2 f_S(x,y,t)}{\partial y^2} u(x,t) + \frac{\partial f_S(x,y,t)}{\partial y} v(x,t), \quad v(x,t_0) = \frac{\mathrm{d}^2 f_I(x)}{\mathrm{d}x^2}, \qquad (1079) \\
\frac{\mathrm{d}\tilde{G}_A(x,t)}{\mathrm{d}t} &= \frac{\partial^2 F_A(x,y,t)}{\partial x^2} + \frac{\partial^2 F_A(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y \partial x} \\
&+ u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y^2} u(x,t) + \frac{\partial F_A(x,y,t)}{\partial y} v(x,t), \quad \tilde{G}_A(x,t_0) = 0. \qquad (1080)
\end{aligned}$$

K současnému určení hodnoty, gradientu a Hessovy matice minimalizované funkce je tedy třeba řešit $(n_S+1)(n^2+n+1)$ diferenciálních rovnic (1070)–(1071) a (1079)–(1080) v přímém směru. Poznamenejme, že v (1079) vystupují tenzorové veličiny, takže je třeba dávat pozor na pořadí násobení. Platí například

$$\begin{aligned}
\left(\frac{\partial^2 f_S(x,y,t)}{\partial x \partial y} u(x,t) \right)_{kij} &= \sum_{p=1}^{n_S} \frac{\partial^2 f_k(x,y,t)}{\partial x_i \partial y_p} u_{pj}(x,t), \\
\left(u^T(x,t) \frac{\partial^2 f_S(x,y,t)}{\partial y^2} u(x,t) \right)_{kij} &= \sum_{p=1}^{n_S} \sum_{q=1}^{n_S} u_{pi}(x,t) \frac{\partial^2 f_k(x,y,t)}{\partial y_p \partial y_q} u_{qj}(x,t), \\
\left(\frac{\partial f_S(x,y,t)}{\partial y} v(x,t) \right)_{kij} &= \sum_{p=1}^{n_S} \frac{\partial f_k(x,y,t)}{\partial y_p} v_{pij},
\end{aligned}$$

kde $f_S(x,y,t) = [f_1(x,y,t), \ldots, f_{n_S}(x,y,t)]^T$.

Nyní odvodíme zpětný výpočet Hessovy matice. Nechť $u(t) \in R^{n_S \times n}$ je libovolné zobrazení a $p(x,t)$ je řešení soustavy diferenciálních rovnic (1072). Rovnici (1073) můžeme zapsat ve tvaru

$$\begin{aligned}
g^T(x) &= \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial x} \right. \\
&+ \left. \left(\frac{\partial F_A(x,y,t)}{\partial y} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial y} + \frac{dp^T(x,t)}{dt} \right) u(t) \right] dt \\
&+ p^T(x,t_0) \frac{df_I(x)}{dx} + \frac{\partial F_T(x,y(x,t_1))}{\partial x}
\end{aligned}$$

(neboť výraz v závorce u $u(t)$ je podle (1072) identicky nulový) a použitím pravidla integrování per partes dostaneme

$$\begin{aligned}
g^T(x) &= \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial x} \right. \\
&+ \left. \left(\frac{\partial F_A(x,y,t)}{\partial y} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial y} \right) u(t) - p^T(x,t) \frac{du(t)}{dt} \right] dt \\
&+ p^T(x,t_0) \frac{df_I(x)}{dx} + \frac{\partial F_T(x,y(x,t_1))}{\partial x} + \frac{\partial F_T(x,y(x,t_1)}{\partial y} u(t_1) - p^T(x,t_0) u(t_0),
\end{aligned}$$

neboť podle (1072) platí $p^T(x,t_1) = \partial F_T(x,y(x,t_1))/\partial y$. Tuto rovnici můžeme derivovat podle x, čímž dostaneme

$$
\begin{aligned}
G(x) &= \int_{t_0}^{t_1} \left[\frac{\partial^2 F_A(x,y,t)}{\partial x^2} + \frac{\partial^2 F_A(x,y,t)}{\partial x \partial y} \frac{dy(x,t)}{dx} + \frac{dp^T(x,t)}{dx} \frac{\partial f_S(x,y,t)}{\partial x} \right. \\
&+ \sum_{k=1}^{n_S} p_k(x,t) \left(\frac{\partial^2 f_k(x,y,t)}{\partial x^2} + \frac{\partial^2 f_k(x,y,t)}{\partial x \partial y} \frac{dy(x,t)}{dx} \right) \\
&+ u^T(t) \left(\frac{\partial^2 F_A(x,y,t)}{\partial y \partial x} + \frac{\partial^2 F_A(x,y,t)}{\partial y^2} \frac{dy(x,t)}{dx} \right) + \frac{dp^T(x,t)}{dx} \frac{\partial f_S(x,y,t)}{\partial y} u(t) \\
&+ \left. \sum_{k=1}^{n_S} p_k(x,t)\, u^T(t) \left(\frac{\partial^2 f_k(x,y,t)}{\partial y \partial x} + \frac{\partial^2 f_k(x,y,t)}{\partial y^2} \frac{dy(x,t)}{dx} \right) - \frac{dp^T(x,t)}{dx} \frac{du(t)}{dt} \right] dt \\
&+ \frac{dp^T(x,t_0)}{dx} \frac{df_I(x)}{dx} + p^T(x,t_0) \frac{d^2 f_I(x)}{dx^2} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x^2} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x \partial y} \frac{dy(x,t_1)}{dx} \\
&+ u^T(t_1) \left(\frac{\partial^2 F_T(x,y(x,t_1))}{\partial y \partial x} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y^2} \frac{dy(x,t_1)}{dx} \right) - \frac{dp^T(x,t_0)}{dx} u^T(t_0),
\end{aligned}
$$

kde $dp^T(x,t)/dt = (dp(x,t)/dt)^T$. Zvolíme-li funkci $u(t) = u(x,t)$ tak aby se vynulovaly členy s $dp(x,t)/dx$ a $dp(x,t_0)/dx$, dostaneme

$$
\frac{du(x,t)}{dt} = \frac{\partial f_S(x,y,t)}{\partial x} + \frac{\partial f_S(x,y,t)}{\partial y} u(x,t), \qquad u(x,t_0) = \frac{df_I(x)}{dx},
$$

takže podle (1070) platí $u(x,t) = dy(x,t)/dx$. Dosadíme-li získané výsledky do výrazu pro $G(x)$, můžeme psát

$$
\begin{aligned}
G(x) &= \int_{t_0}^{t_1} \left[\frac{\partial^2 F_A(x,y,t)}{\partial x^2} + \frac{\partial^2 F_A(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y \partial x} \right. \\
&+ u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y^2} u(x,t) + \sum_{k=1}^{n_S} p_k(x,t) \left(\frac{\partial^2 f_k(x,y,t)}{\partial x^2} \right. \\
&+ \left. \left. \frac{\partial^2 f_k(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 f_k(x,y,t)}{\partial y \partial x} + u^T(x,t) \frac{\partial^2 f_k(x,y,t)}{\partial y^2} u(x,t) \right) \right] dt \\
&+ p^T(x,t_0) \frac{d^2 f_I(x)}{dx^2} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x^2} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x \partial y} u(x,t_1) \\
&+ u(x,t_1) \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y \partial x} + u(x,t_1) \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y^2} u(x,t_1),
\end{aligned}
$$

Dohromady to lze zapsat takto

$$
G(x) = \tilde{G}(x,t_0) + p^T(x,t_0) \frac{d^2 f_I(x)}{dx^2} \tag{1081}
$$

kde

$$\begin{aligned}
\frac{d\tilde{G}(x,t)}{dt} &= \frac{\partial^2 F_A(x,y,t)}{\partial x^2} + \frac{\partial^2 F_A(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y \partial x} \\
&+ u^T(x,t) \frac{\partial^2 F_A(x,y,t)}{\partial y^2} u(x,t) \\
&+ \sum_{k=1}^{n_S} p_k(x,t) \left(\frac{\partial^2 f_k(x,y,t)}{\partial x^2} + \frac{\partial^2 f_k(x,y,t)}{\partial x \partial y} u(x,t) + u^T(x,t) \frac{\partial^2 f_k(x,y,t)}{\partial y \partial x} \right. \\
&+ \left. u^T(x,t) \frac{\partial^2 f_k(x,y,t)}{\partial y^2} u(x,t) \right) \qquad (1082) \\
\tilde{G}(x,t_1) &= \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x \partial y} u(x,t_1) + u(x,t_1) \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y \partial x} \\
&+ u(x,t_1) \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y^2} u(x,t_1).
\end{aligned}$$

K současnému určení hodnoty a gradientu minimalizované funkce je tedy třeba řešit $(n_S + 1)(n + 1)$ diferenciálních rovnic (1067)–(1068) a (1070)–(1071) v přímém směru a $n_S + n^2$ diferenciálních rovnic (1075) a (1087) ve zpětném směru.

Poznámka 394. Při zpětném výpočtu Hessovy matice lze postupovat podobně jako v poznámce 391. Použijeme-li první způsob, je třeba při řešení v přímém směru ukládat vektory $y(x, \tilde{t}_i)$ a matice $u(x, \tilde{t}_i)$ v uzlových bodech $\tilde{t}_i$, $1 \le i \le n_A$, tedy celkem $n_A n_S (n + 1)$ hodnot. Použijeme-li druhý způsob, je třeba ve zpětném směru řešit dalších $n\, n_S$ diferenciálních rovnic

$$\frac{\mathrm{d}u(x,t)}{\mathrm{d}t} = \frac{\partial f_S(x,y,t)}{\partial x} + \frac{\partial f_S(x,y,t)}{\partial y} u(x,t), \qquad u(x,t_1) = u(x,t_1), \qquad (1083)$$

tedy celkem $n_S(n + 2) + n^2$ diferenciálních rovnic (1070)–(1071), (1077) a (1083). Z těchto úvah plyne, že zpětný výpočet Hessovy matice není výhodný. V prvním případě je třeba uklaádat $n_A n_S (n + 1)$ hodnot navíc a ve druhém případě je celkový počet diferenciálních rovnic stejný jako při přímém výpočtu Hessovy matice, ale většina z těchto rovnic se řeší ve zpětném směru.

13.3 Aproximace Hessovy matice pro kritérium nejmenších čtverců

Úloha optimalizace dynamických systémů bývá často formulována tak, že se hledá optimální vektor x v (1067) tak, aby průběh funkce $y(x,t)$ na intervalu $[t_0, t_1]$ co nejlépe aproximoval předepsaný průběh $z(t)$ a aby v koncovém bodě intervalu co nejpřesněji platilo $y(x,t_1) = z(t_1)$. V tomto případě se nejčastěji používá analogie součtu čtverců, kdy funkce F_A a F_T mají tvar

$$\begin{aligned}
F_A(y(x,t),t) &= \frac{1}{2}(y(x,t) - z(t))^T W_A(t)(y(x,t) - z(t)), \\
F_T(y(x,t_1)) &= \frac{1}{2}(y(x,t_1) - z(t_1))^T W_T(y(x,t_1) - z(t_1)),
\end{aligned}$$

kde $z : [t_0, t_1] \to R^{n_S}$ je předpsaný průběh a $W_A : [t_0, t_1] \to R^{n_S \times n_S}$, $W_T \in R^{n_S \times n_S}$ jsou symetrické pozitivně definitní váhové matice (obecně $W_T \ne W_A(t_1)$). Tyto funkce nezávisejí na vektoru x a platí

$$\begin{aligned}
\frac{\partial F_A(y,t)}{\partial y} &= W_A(t)(y(x,t) - z(t)), \qquad \frac{\partial^2 F_A(y,t)}{\partial y^2} = W_A(t), \\
\frac{\partial F_T(y(x,t_1))}{\partial y} &= W_T(y(x,t_1) - z(t_1)), \qquad \frac{\partial^2 F_T(y(x,t_1))}{\partial y^2} = W_T.
\end{aligned}$$

Dosadíme-li tyto vztahy do vzorců uvedených v oddílu 13.1, dostaneme zjednodušené rovnice pro výpočet gradientu účelové funkce.

Speciální tvar funkcí F_A a F_T je výhodný pro odvození dobré aproximace Hessovy matice postupem, který odpovídá Gaussově-Newtonově metodě. Dosadíme-li vztahy pro funkce F_A a F_T do vzorců (1078) a (1080), dostaneme

$$G(x) = \tilde{G}_A(x,t_1) + u^T(x,t_1)W_T u(x,t_1) + (y(x,t_1) - z(t_1))^T W_T v(x,t_1), \tag{1084}$$

$$\frac{\mathrm{d}\tilde{G}_A(x,t)}{\mathrm{d}t} = u^T(x,t)W_A(t)u(x,t) + (y(x,t) - z(t))^T W_A(t)v(x,t), \qquad \tilde{G}_A(x,t_0) = 0, \tag{1085}$$

kde $v(x,t)$ je řešení soustavy rovnic (1079). Jestliže $F(x) \to 0$, pak také $y(x,t) - z(t) \to 0$ a poslední členy v (1084)–(1085) lze zanedbat. Tím odpadne i řešení soustavy rovnic (1079). Pro úlohy s nulovým reziduem tedy platí $G(x) \approx B(x)$, kde

$$B(x) = B_A(x,t_1) + u^T(x,t_1)W_T u(x,t_1), \tag{1086}$$

$$\frac{\mathrm{d}B_A(x,t)}{\mathrm{d}t} = u^T(x,t)W_A(t)u(x,t), \qquad B_A(x,t_0) = 0. \tag{1087}$$

Jelikož v (1086)–(1087) vystupuje matice $u(x,t)$ určená řešením soustavy rovnic (1070), používá se přímý výpočet gradientu účelové funkce podle vzorců (1069)–(1071). Celkem se řeší $(n_S+1)(n+1)+n^2$ diferenciálních rovnic v přímém směru.

13.4 Metody pro optimalizaci dynamických systémů

K minimalizaci funkce určené vztahy (1066)–(1068) lze použít libovolnou optimalizační metodu. Použijeme-li gradientní metodu, je třeba počítat gradient účelové funkce přímým řešením diferenciálních rovnic (1069)–(1071) nebo zpětným řešením diferenciálmích rovnic (1074)–(1076). K tomu, aby gradientní metoda byla globálně konvergentní, potřebujeme, aby funkce F splňovala předpoklady F1 a F3. Předpoklad F1 je důsledkem předpokladu D1. Ukážeme nyní, kdy je splněn předpoklad F3. Použijeme k tomu následující tvrzení, jehož důkaz lze nalézt například v [T4].

Tvrzení 9. *Uvažujme soustavu lineárních diferenciálních rovnic*

$$\frac{dy(t)}{dt} = A(t)y(t) + a(t), \qquad y(t_0) = y_0,$$

kde veličiny $A(t)$ a $a(t)$ jsou spojité na intervalu $[t_0,t_1]$. Pak

$$\|y(t)\| \le \left(\|y_0\| + \int_{t_0}^{t} \|a(\tau)\|\, d\tau\right) \exp\left(\int_{t_0}^{t} \|A(\tau)\|\, d\tau\right),$$

kdykoliv $t \in [t_0,t_1]$.

Věta 275. *Nechť jsou splněny předpoklady D1–D4. Pak funkce $F(x)$ určená vztahy (1066)–(1068) splňuje předpoklad F3.*

Důkaz Podle předpokladu D3 existuje konstanta $\overline{K} > 0$ taková, že $\|f_I(x)\| \le \overline{K}$, $\|df_I(x)/dx\| \le \overline{K}$ na $\mathcal{D}$ a $\|\partial f_S(x,y,t)/\partial x\| \le \overline{K}$, $\|\partial f_S(x,y,t)/\partial y\| \le \overline{K}$ na $\mathcal{D} \times \mathcal{D}_y \times [t_0,t_1]$ (ve smyslu poznámky 392 používáme společnou konstantu $\overline{K}$). Použijeme-li tvrzení 9 na systém (1067), dostaneme

$$\|u(x,t)\| \le (\overline{K} + \overline{K}(t_1 - t_0)) \exp(\overline{K}(t_1 - t_0)) \triangleq \overline{M}. \tag{1088}$$

Jelikož $u(x,t) = dy(x.t)/dx$, můžeme podle (1088) psát

$$\|y(x_2,t) - y(x_1,t)\| = \left\| \int_{t_0}^{t_1} u(x_1 + \lambda(x_2 - x_1))(x_2 - x_1)\mathrm{d}\lambda \right\| \le \overline{M}(x_2 - x_1).$$

Podle předpokladu D4 a poznámky 392 má zobrazení $f_S(x,y,t)$ lipschitzovské první derivace, takže

$$\begin{aligned}\left\|\frac{\partial f_S(x_2,y_2,t)}{\partial x}-\frac{\partial f_S(x_1,y_1,t)}{\partial x}\right\| &\leq \left\|\frac{\partial f_S(x_2,y_2,t)}{\partial x}-\frac{\partial f_S(x_2,y_1,t)}{\partial x}\right\| \\ &+ \left\|\frac{\partial f_S(x_2,y_1,t)}{\partial x}-\frac{\partial f_S(x_1,y_1,t)}{\partial x}\right\| \\ &\leq \overline{L}(\|y_2-y_1\|+\|x_2-x_1\|)\leq \overline{L}(\overline{M}+1)\|x_2-x_1\|\end{aligned}$$

a podobně

$$\left\|\frac{\partial f_S(x_2,y_2,t)}{\partial y}-\frac{\partial f_S(x_1,y_1,t)}{\partial y}\right\| \leq \overline{L}(\|y_2-y_1\|+\|x_2-x_1\|)\leq \overline{L}(\overline{M}+1)\|x_2-x_1\|.$$

Zobrazení $f_I(x)$ má též lipschitzovské první derivace, takže

$$\left\|\frac{df_I(x_2)}{dx}-\frac{df_I(x_1)}{dx}\right\|\leq \overline{L}\|x_2-x_1\|.$$

Použijeme-li (1070), dostaneme

$$\begin{aligned}\frac{d(u(x_2,t)-u(x_1,t))}{dt} &= \frac{\partial f_S(x_2,y_2,t)}{\partial y}(u(x_2,t)-u(x_1,t)) \\ &+\left(\frac{\partial f_S(x_2,y_2,t)}{\partial y}-\frac{\partial f_S(x_1,y_1,t)}{\partial y}\right)u(x_1,t) \\ &+\left(\frac{\partial f_S(x_2,y_2,t)}{\partial x}-\frac{\partial f_S(x_1,y_1,t)}{\partial x}\right)\end{aligned}$$

a

$$u(x_2,t_0)-u(x_1,t_0)=\frac{df_I(x_2)}{dx}-\frac{df_I(x_1)}{dx}\ ,$$

takže podle tvrzení 9 platí

$$\begin{aligned}\|u(x_2,t)-u(x_1,t)\| &\leq \left[\left\|\frac{df_I(x_2)}{dx}-\frac{df_I(x_1)}{dx}\right\|+\int_{t_0}^{t_1}\left(\left\|\frac{\partial f_S(x_2,y_2,\tau)}{\partial y}-\frac{\partial f_S(x_1,y_1,\tau)}{\partial y}\right\|\|u(x_1,\tau)\|\right.\right. \\ &+ \left.\left.\left\|\frac{\partial f_S(x_2,y_2,\tau)}{\partial x}-\frac{\partial f_S(x_1,y_1,\tau)}{\partial x}\right\|\right)\mathrm{d}\tau\right]\exp\left(\int_{t_0}^{t_1}\left\|\frac{\partial f_S(x_2,y_2,\tau)}{\partial y}\right\|\mathrm{d}\tau\right) \\ &\leq \overline{L}(1+(\overline{M}+1)^2(t_1-t_0))\exp(\overline{K}(t_1-t_0))\|x_2-x_1\|\stackrel{\Delta}{=}\overline{N}\|x_2-x_1\|. \qquad (1089)\end{aligned}$$

Jelikož rovnice (1071) má stejný tvar jako rovnice (1070) a funkce $F_A(x,y,t)$ má podle předpokladu D4 a poznámky 392 lipschitzovské první derivace, můžeme postupovat stejně jako v předchozím případě. Dostaneme tak odhad

$$\|g_A(x_2,t)-g_A(x_1,t)\|\leq \overline{L}(\overline{M}+1)^2(t_1-t_0)\exp(\overline{K}(t_1-t_0))\|x_2-x_1\|\leq \overline{N}\|x_2-x_1\|. \qquad (1090)$$

Použijeme-li (1071), dostaneme

$$\begin{aligned}g^T(x_2,t)-g^T(x_1,t) &= \tilde{g}_A^T(x_2,t)-\tilde{g}_A^T(x_1,t)+\frac{\partial f_T(x_2,y(x_2,t_1))}{\partial y}(u(x_2,t_1)-u(x_1,t_1)) \\ &+\left(\frac{\partial f_T(x_2,y(x_2,t_1))}{\partial y}-\frac{\partial f_T(x_1,y(x_1,t_1))}{\partial y}\right)u(x_1,t_1) \\ &+\left(\frac{\partial f_T(x_2,y(x_2,t_1))}{\partial x}-\frac{\partial f_T(x_1,y(x_1,t_1))}{\partial x}\right),\end{aligned}$$

takže podle (1088), (1089), (1090) a předpokladu D4 platí

$$\|g(x_2,t)-g(x_1,t))\| \le (\overline{N}+\overline{K}\,\overline{N}+\overline{L}\,\overline{M}+\overline{L})\|x_2-x_1\| \triangleq \overline{G}\|x_2-x_1\|.$$

□

Poznámka 395. Předpoklad D4 je zbytečně silný (spojitost a omezenost druhých derivací potřebujeme, používáme-li Hessovu matici). Věta 275 platí i tehdy, nahradíme-li spojitost a omezenost druhých derivací lipschitzovskostí prvních derivací.

Nyní se omezíme na případ, kdy funkce F_A a F_T mají tvar uvedený v oddílu 13.3. Budeme přitom používat nádující předpoklad.

Předpoklad D5. *Zobrazení $z(t)$ a $W_A(t)$ jsou spojitá a omezená na intervalu $[t_0,t_1]$ (takže $\|z(t)\| \le \overline{K}$ a $\|W_A(t)\| \le \overline{K}$, pokud $t_0 \le t \le t_1$) a platí $\|W_T\| \le \overline{K}$.*

Věta 276. *Nechť jsou splněny předpoklady D1–D4 (bez požadavků kladených na funkce F_A a F_T) a předpoklad D5. Pak metoda s lokálně omezeným krokem (definice 38) používající matici (1081) je globálně konvergentní.*

Důkaz Podle věty 118 je metoda s lokálně omezeným krokem globálně konvergentní, splňuje-li funkce F předpoklady F1 a F3 a platí-li $\|B(x_i)\| \le \overline{B}$, $i \in N$. Předpoklady F1 a F3 jsou splněny podle věty 275. Stačí tedy dokázat omezenost matice $B(x)$. Ta však plyne z předpokladu D5 z odhadu (1088) a z vyjádření (1081)–(1087), neboť

$$\begin{aligned}\|B(x)\| &\le \int_{t_0}^{t_1} \|u^T(x,t)W_A(t)u(x,t)\|\,dt + \|u^T(x,t_1)W_T u(x,t_1)\| \le \\ &\le \overline{K}\,\overline{M}^2(t_1-t_0) + \overline{K}\,\overline{M}^2\,.\end{aligned}$$

□

Věta 277. *Nechť x_i, $i \in N$, je posloupnost bodů generovaná metodou s lokálně omezeným krokem s maticí (1081) taková, že $x_i \to x^*$ pro $i \to \infty$, kde bod $x^* \in R^n$ splňuje postačující podmínky druhého řádu pro lokální minimum funkce $F(x)$. Nechť jsou splněny předpoklady D1–D4 (bez požadavků kladených na funkce F_A a F_T) a předpoklad D5. Pak pokud $F(x_i) \to 0$ pro $i \to \infty$, posloupnost x_i, $i \in N$, konverguje k bodu $x^* \in R^n$ superlineárně, neboli*

$$\lim_{i\to\infty} \frac{\|x_{i+1}-x^*\|}{\|x_i-x^*\|} = 0$$

Důkaz Dokážeme, že $B(x_i) \to G(x_i)$ pro $i \to \infty$, takže superlineární konvergence plyne z věty 122. Podle předpokladů D2–D4 jsou všechna zobrazení v rovnici (1079) spojitá a omezená, takže můžeme použít tvrzení 9, podle kterého je řešení $v(x,t) = du(x,t)/dx$ omezené. Existuje tedy konstanta $\overline{C} > 0$ taková, že $\|v(x,t)\| \le \overline{C}$, pokud $x \in \mathcal{D}$ a $t \in [t_0,t_1]$. Použijeme-li (1084) a (1081), můžeme psát

$$\begin{aligned}\|G(x_i)-B(x_i)\| &\le \int_{t_0}^{t_1} \|(y(x_i,t)-z(t))^T W_A(t)v(x_i,t)\|\,dt + \|(y(x_i,t_1)-z(t_1))^T W_T v(x_i,t_1)\| \\ &\le \overline{C}\,\overline{K}^{1/2}\int_{t_0}^{t_1} \|W_A^{1/2}(t)(y(x_i,t)-z(t))\|\,dt + \overline{C}\,\overline{K}^{1/2}\|W_T^{1/2}(y(x_i,t_1)-z(t_1))\|\end{aligned}$$

a podle (1065), kde funkce F_A a F_T mají tvar uvedený v oddílu 13.3, platí

$$\begin{aligned}2F(x_i) &= \int_{t_0}^{t_1} (y(x_i,t)-z(t))^T W_A(t)(y(x_i,t)-z(t))\,dt \\ &+ (y(x_i,t_1)-z(t_1))^T W_T(y(x_i,t_1)-z(t_1)) \\ &= \int_{t_0}^{t_1} \|W_A^{1/2}(t)(y(x_i,t)-z(t))\|^2\,dt + \|W_T^{1/2}(y(x_i,t_1)-z(t_1))\|^2\end{aligned}$$

Odtud plyne. že

$\|G(x_i)-B(x_i)\| \to 0$, pokud $F(x_i) \to 0$. □

14 Automatické a numerické derivování

K určování parciálních derivací funkcí více proměnných existují čtyři základní postupy:

(1) Analytické derivování, kdy uživatel zadává výrazy pro parciální derivace formou úseků zdrojového programu.
(2) Symbolické derivování, kdy se zadává pouze výraz pro výpočet funkční hodnoty a pomocí programu pro symbolickou analýzu (který je součástí systémů Mathematica, Maple a Derive) se z tohoto výrazu odvodí vzorce pro parciální derivace. Výsledek je v podstatě stejný jako v předchozím případě, ale vzorce pro parciální derivace nezadává uživatel nýbrž je vytváří program.
(3) Numerické derivování, kdy se derivace aproximují pomocí diferencí.
(4) Automatické derivování, kdy se používají pravidla pro výpočet derivací, do nichž se dosazují numerické hodnoty. Nesestavují se tedy žádné analytické výrazy, ale všechny operace se provádějí s čísly.

Podobně jako výpočet gradientu v případě optimalizace dynamických systémů, lze automatické derivování realizovat dvojím způsobem:

- Přímé automatické derivování, které je vhodné pro výpočet směrové derivace funkce $F : R^n \to R$ nebo pro výpočet součinu Jacobiovy matice zobrazení $f : R^n \to R^m$ a vektoru $d \in R^n$.
- Zpětné automatické derivování, které je vhodné pro výpočet gradientu funkce $F : R^n \to R$ nebo pro výpočet součinu transponované Jacobiovy matice zobrazení $f : R^n \to R^m$ a vektoru $w \in R^m$.

Abychom mohli provádět automatické derivování, je třeba vzorec definující hodnotu derivované funkce (nebo zobrazení) vyjádřit jako posloupnost elementárních kroků obsahujících elementární operace a elementární funkce. Za tímto účelem použijeme označení $v_1, \ldots, v_n, v_{n+1}, \ldots, v_l$, kde $v_i = x_i$, $1 \le i \le n$, $v_i = \varphi_i(v_j, j \in I_i)$, $n+1 \le i \le l$ a $v_l = F(x)$ (předpokládáme, že $F := R^n \to R$). Zápis $v_i = \varphi_i(v_j, j \in I_i)$ znamená, že hodnota v_i závisí na předchozích hodnotách $v_j, j \in I_i$ (předpokládáme, že $I_i \subset \{1, \ldots, i-1\}$), přičemž $\varphi_i(v_j, j \in I_i)$ je příslušná elementární funkce. Za elementární funkce považujeme součet, součin, lineární funkci a elementární funkce, které jsou součástí programovacího jazyka (obecná mocnina, exponenciála, logaritmus, goniometrické a hyperbolické funkce a funkce k nim inverzní), takže funkce $\varphi_i(v_j, j \in I_i)$, $n+1 \le i \le l$, závisejí nanejvýš na dvou proměnných (množiny I_i, $n+1 \le i \le l$, jsou nanejvýš dvouprvkové). Jako příklad ukážeme posloupnost elementárních kroků pro výpočet funkce $F(x) = \sin(x_1)\exp(x_2/x_3)$.

v_1	$=$	x_1			
v_2	$=$	x_2			
v_3	$=$	x_3			
v_4	$=$	$\varphi_4(v_2, v_3)$	$=$	v_2/v_3	x_2/x_3
v_5	$=$	$\varphi_5(v_4)$	$=$	$\exp(v_4)$	$\exp(x_2/x_3)$
v_6	$=$	$\varphi_6(v_1)$	$=$	$\sin(v_1)$	$\sin(x_1)$
v_7	$=$	$\varphi_7(v_5, v_6)$	$=$	$v_6 v_5$	$\sin(x_1)\exp(x_2/x_3)$
$F(x)$	$=$	v_7			$\sin(x_1)\exp(x_2/x_3)$

Tabulka 20: Vyčíslení funkce $F(x) = \sin(x_1)\exp(x_2/x_3)$.

Řádky nad první vodorovnou čarou budeme nazývat přípravnou fází algoritmu. Ostatní řádky nazveme výpočetní fází algoritmu. Poznamenejme, že vzorce zapsané napravo od svislé čáry, které ukazují skutečnou hodnotu dané proměnné, slouží pouze pro orientaci (tyto vzorce budeme uvádět i v dalších tabulkách i když jsou pro automatické derivování irelevantní). Funkční hodnota $F(x) = v_7$ se počítá podle jednoduchých vzorců vystupujících za posledním rovnítkem. Do těchto vzorců se dosazují numerické hodnoty (počínaje

numerickými hodnotami proměnných x_1, x_2, x_3, x_4). Při sestavování tabulky elementárních kroků vycházíme z analytického popisu funkce zadaného uživatelem. Nejprve se provede syntaktická analýza daného výrazu, což je proces, který používá většina kompilátorů, a na základě této analýzy se sestavuje posloupnost elementárních operací a elementárních funkcí.

14.1 Automatický výpočet prvních derivací

Nejjednodušším způsobem výpočtu prvních derivací je přímé automatické derivování. Při tomto způsobu se za každý elementární krok pro výpočet funkční hodnoty přidá řádek obsahující elementární krok pro výpočet zvolené derivace. Používá se přitom pravidlo řetězení. Pro $n+1 \leq i \leq l$ tedy platí

$$v_i = \varphi_i(v_j, j \in I_i), \tag{1091}$$

$$v_i' = \sum_{j \in I_i} \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j} v_j', \tag{1092}$$

přičemž $F'(x) = v_l'$ (čárka označuje zvolenou derivaci). Je-li funkce $\varphi_i(v_j, j \in I_i)$ elementární funkcí, použijí se pro výpočet hodnoty $\partial \varphi_i(v_j, j \in I_i)/\partial v_j$ pravidla uvedená v tabulce 21 (kde $a > 0$ a $c \neq 0$).

$F(x)$	$F'(x)$	$\mathcal{D}_F$	$F(x)$	$F'(x)$	$\mathcal{D}_F$
x^n	$nx^{(n-1)}$	R	x^c	$cx^{(c-1)}$	$x > 0$
e^x	e^x	R	a^x	$a^x \log a$	R
$\log x$	$\frac{1}{x}$	$x > 0$	$\log_a x$	$\frac{1}{x \log a}$	$x > 0$
$\sin x$	$\cos x$	R	$\cos x$	$-\sin x$	R
$\operatorname{tg} x$	$\frac{1}{\cos^2 x}$	$\cos x \neq 0$	$\operatorname{cotg} x$	$\frac{-1}{\sin^2 x}$	$\sin x \neq 0$
$\sinh x$	$\cosh x$	R	$\cosh x$	$\sinh x$	R
$\operatorname{tgh} x$	$\frac{1}{\cosh^2 x}$	R	$\operatorname{cotgh} x$	$\frac{-1}{\sinh^2 x}$	$x \neq 0$
$\arcsin x$	$\frac{1}{\sqrt{1-x^2}}$	$\lvert x\rvert < 1$	$\arccos x$	$\frac{-1}{\sqrt{1-x^2}}$	$\lvert x\rvert < 1$
$\operatorname{arctg} x$	$\frac{1}{1+x^2}$	R	$\operatorname{arccotg} x$	$\frac{-1}{1+x^2}$	R
$\operatorname{argsinh} x$	$\frac{1}{\sqrt{x^2+1}}$	R	$\operatorname{argcosh} x$	$\frac{1}{\sqrt{x^2-1}}$	$x > 1$
$\operatorname{argtgh} x$	$\frac{1}{1-x^2}$	$\lvert x\rvert < 1$	$\operatorname{argcotgh} x$	$\frac{1}{1-x^2}$	$\lvert x\rvert > 1$
$u+v$	$u'+v'$	R	$u-v$	$u'-v'$	R
uv	$u'v+uv'$	R	u/v	$(u'-v'u/v)/v$	$v \neq 0$

Tabulka 21: Derivování elementárních funkcí.

Pokud $x_i = \psi_i(t)$, $1 \leq i \leq n$, můžeme přímým automatickým derivováním vypočítat derivaci $\mathrm{d}F(x)/\mathrm{d}t$ tak, že v přípravné fázi položíme $v_i = \psi_i(t)$, $v_i' = \psi_i'(t)$, $1 \leq i \leq n$. Přímé automatické derivování je velmi

vhodné pro výpočet směrové derivace

$$F'(x,d) = \lim_{t\downarrow 0} \frac{F(x+td) - F(x)}{t} = (\nabla F(x))^T d = \sum_{i=1}^{n} d_i \frac{\partial F(x)}{\partial x_i}.$$

V tomto případě v přípravné fázi pokládáme $v_i = x_i$, $v_i' = d_i$, $1 \le i \le n$.

Příklad 8. Pomocí přímého automatického derivování vypočteme hodnotu $F(x(t))$ a derivaci $\mathrm{d}F(x)/\mathrm{d}t$ funkce uvedené v tabulce 20 za předpokladu, že $x_1 = 2t$, $x_2 = t^2$, $x_3 = -t$.

v_1	=	x_1	$2t$
v_1'	=	x_1'	2
v_2	=	x_2	t^2
v_2'	=	x_2'	$2t$
v_3	=	x_3	$-t$
v_3'	=	x_3'	-1
v_4	=	v_2/v_3	$-t$
v_4'	=	$(v_2' - v_4 v_3')/v_3$	-1
v_5	=	$\exp(v_4)$	$\exp(-t)$
v_5'	=	$\exp(v_4)v_4'$	$-\exp(-t)$
v_6	=	$\sin(v_1)$	$\sin(2t)$
v_6'	=	$\cos(v_1)v_1'$	$2\cos(2t)$
v_7	=	$v_6 v_5$	$\sin(2t)\exp(-t)$
v_7'	=	$v_6' v_5 + v_6 v_5'$	$(2\cos(2t) - \sin(2t))\exp(-t)$
$F(x)$	=	v_7	$\sin(2t)\exp(-t)$
$\mathrm{d}F(x)/\mathrm{d}t$	=	v_7'	$(2\cos(2t) - \sin(2t))\exp(-t)$

Tabulka 22: Vyčíslení derivace funkce závislé na parametru.

Příklad 9. Pomocí přímého automatického derivování vypočteme směrovou derivaci funkce uvedené v tabulce 20 ve směru $d = [2, 1, -1]^T$.

v_1	=	x_1	
v_1'	=	2	
v_2	=	x_2	
v_2'	=	1	
v_3	=	x_2	
v_3'	=	-1	
v_4	=	v_2/v_3	x_2/x_3
v_4'	=	$(v_2' - v_4 v_3')/v_3$	$(1 + x_2/x_3)/x_3$
v_5	=	$\exp(v_4)$	$\exp(x_4)$
v_5'	=	$\exp(v_4)v_4'$	$\exp(x_4)(1 + x_2/x_3)/x_3$
v_6	=	$\sin(v_1)$	$\sin(x_1)$
v_6'	=	$\cos(v_1)v_1'$	$2\cos(x_1)$
v_7	=	$v_6 v_5$	$\sin(x_1)\exp(x_2/x_3)$
v_7'	=	$v_6' v_5 + v_6 v_5'$	$(2\cos(x_1) + sin(x_1)\exp(x_4)(1 + x_2/x_3)/x_3$
$F(x)$	=	v_7	$\sin(x_1)\exp(x_2/x_3)$
$F'(x,h)$	=	v_7'	$(2\cos(x_1) + \sin(x_1)\exp(x_4)(1 + x_2/x_3)/x_3$

Tabulka 23: Výpočet směrové derivace funkce $F(x) = \sin(x_1)\exp(x_2/x_3)$.

Z tabulky 23 je patrné, že pro současné vyčíslení funkční hodnoty a jedné směrové derivace potřebujeme zhruba trojnásobek operací jako pro samotný výpočet funkční hodnoty. Teoreticky se dá dokázat, že $\mathrm{Op}(F(x))+\mathrm{Op}(F'(x,h)) \leq 3\,\mathrm{Op}(F(x))$ ($\mathrm{Op}(.)$ je počet aritmetických operací potřebných pro výpočet dané veličiny). Přímé automatické derivování není vhodné pro výpočet gradientu funkce, neboť výpočet každé parciální derivace vyžaduje alespoň $\mathrm{Op}(F(x))$ aritmetických operací, takže platí $\mathrm{Op}(F(x)) + \mathrm{Op}(g(x)) \geq (n+1)\,\mathrm{Op}(F(x))$.

Přímé automatické derivování můžeme snadno zobecnit pro výpočet směrových derivací prvků $f_k(x)$, $1 \leq k \leq m$, zobrazení $f : R^n \to R^m$, neboli součinu Jacobiovy matice $J(x)$ a vektoru d. Často se stává, že výrazy odpovídající jednotlivým prvkům zobrazení f spolu nijak nesouvisí. V tomto případě lze samostatně zpracovávat m posloupností elementárních kroků, což odpovídá výpočtu směrových derivací m nezávislých funkcí. Pokud některé prvky zobrazení f obsahují společné výrazy, lze určení všech derivací provést v jediném společném schematu. Pak platí $f_k(x) = v_{l-m+k}$, $f'_k(x) = v'_{l-m+k}$, $1 \leq k \leq m$.

Příklad 10. Uvažujme zobrazení $f : R^3 \to R^3$ s Jacobiovou maticí $J \in R^{3\times 3}$, kde

$$f(x) = \begin{bmatrix} \sin(x_1)\exp(x_2) \\ \cos(x_2)/x_3 \\ (x_1^2 - x_2)^2 \end{bmatrix}, \qquad J(x) = \begin{bmatrix} \cos(x_1)\exp(x_2) & \sin(x_1)\exp(x_2) & 0 \\ 0 & -\sin(x_2)/x_3 & -\cos(x_2)/x_3^2 \\ 4x_1(x_1^2 - x_2) & -2(x_1^2 - x_2) & 0 \end{bmatrix}.$$

Pomocí přímého automatického derivování vypočteme součin matice $J(x)$ a vektoru $d = [1, 2, 1]^T$.

v_1	=	x_1	
v'_1	=	1	
v_2	=	x_2	
v'_2	=	2	
v_3	=	x_3	
v'_3	=	1	
v_4	=	$\sin(v_1)$	$\sin(x_1)$
v'_4	=	$\cos(v_1)v'_1$	$\cos(x_1)$
v_5	=	$\exp(v_2)$	$\exp(x_2)$
v'_5	=	$\exp(v_2)v'_2$	$2\exp(x_2)$
v_6	=	$\cos(v_2)$	$\cos(x_2)$
v'_6	=	$-\sin(v_2)v'_2$	$-2\sin(x_2)$
v_7	=	$v_1^2 - v_2$	$x_1^2 - x_2$
v'_7	=	$2v_1v'_1 - v'_2$	$2x_1 - 2$
v_8	=	v_4v_5	$\sin(x_1)\exp(x_2)$
v'_8	=	$v'_4v_5 + v_4v'_5$	$(\cos(x_1) + 2\sin(x_1))\exp(x_2)$
v_9	=	v_6/v_3	$\cos(x_2)/x_3$
v'_9	=	$(v'_6 - v_9v'_3)/v_3$	$-(2\sin(x_2) + \cos(x_2)/x_3)/x_3$
v_{10}	=	v_7^2	$(x_1^2 - x_2)^2$
v'_{10}	=	$2v_7v'_7$	$2(x_1^2 - x_2)(2x_1 - 2)$
$f_1(x)$	=	v_8	$\sin(x_1)\exp(x_2)$
$f_2(x)$	=	v_9	$\cos(x_2)/x_3$
$f_3(x)$	=	v_{10}	$(x_1^2 - x_2)^2$
$e_1^T J(x)d$	=	v'_8	$\sin(x_1)\exp(x_2)$
$e_2^T J(x)d$	=	v'_9	$-(2\sin(x_2) + \cos(x_2)/x_3)/x_3$
$e_3^T J(x)d$	=	v'_{10}	$2(x_1^2 - x_2)(2x_1 - 2)$

Tabulka 24: Násobení Jacobiovy matice vektorem.

Přímé automatické derivování není vhodné pro výpočet prvků Jacobiovy matice zobrazení, neboť by bylo třeba počítat n součinů $J(x)e_i$ $1 \leq i \leq n$. Situace se však zjednoduší, je-li Jacobiova matice řídká (takže funkce $f_k(x)$, $1 \leq k \leq m$, jsou separovatelné). V tomto případě stačí pro $1 \leq k \leq m$ počítat

směrové derivace $f'_k(x, e_i)$, $i \in N_k$, kde N_k jsou množiny indexů proměnných definujících podprostory R^n_k, v nichž leží definiční obory funkcí $f_k(x)$, $1 \le k \le m$ (použité označení je zavedeno v oddílu 10.4). Celkem se tedy vyhodnocuje $m\,O(1) = O(m)$ směrových derivací (jednotlivé funkce f_k, $1 \le k \le m$, se zpracovávají samostatně).

Nyní ukážeme, jak lze spočítat gradient funkce více proměnných pomocí zpětného automatického derivování. Za tímto účelem použijeme pro $1 \le j \le l$ označení

$$\bar{v}_j = \frac{\partial F(x)}{\partial v_j}$$

(takže $\bar{v}_l = 1$) a píšeme

$$\bar{I}_j = \{i : j \in I_i\} \subset \{j+1, \dots l\},$$

neboli $i \in \bar{I}_j \;\Leftrightarrow\; j \in I_i$. Nejprve budeme předpokládat, že $F : R^n \to R$ (budeme uvažovat jedinou funkční hodnotu).

Věta 278. *Nechť pro $n+1 \le i \le l$ platí (1091). Označme $\bar{v}_i = \partial F(x)/\partial v_i$, $1 \le i \le l$. Pak platí $\bar{v}_l = 1$ a*

$$\bar{v}_j = \sum_{i \in \bar{I}_j} \bar{v}_i \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j}, \tag{1093}$$

kde $l - 1 \ge j \ge 1$.

Důkaz Zřejmě $\bar{v}_l = \partial F(x)/\partial v_l = \partial v_l/\partial v_l = 1$. Vypíšeme-li elementární kroky pro vyčíslení hodnoty funkce $F(x) = v_l$, vidíme, že v_l závisí (kromě jiného) na proměnných $v_i, i \in \bar{I}_j$, a ty zase na proměnné v_j. Můžeme tedy psát

$$v_l = v_l\left(v_i(v_j, \dots), i \in \bar{I}_j, \dots\right),$$

kde $v_l(.)$ je nějaká funkce, jejíž explicitní vyjádření nás nezajímá. Podle pravidla řetězení pak dostaneme

$$\bar{v}_j = \frac{\partial v_l}{\partial v_j} = \sum_{i \in \bar{I}_j} \frac{\partial v_l}{\partial v_i}\frac{\partial v_i}{\partial v_j} = \sum_{i \in \bar{I}_j} \bar{v}_i \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j}.$$

□

Při realizaci zpětného automatického derivování se nejprve sestaví posloupnost elementárních kroků (1091) pro výpočet funkční hodnoty, přičemž se určí množiny I_i, $n+1 \le i \le l$. Pak lze postupovat dvojím způsobem. První způsob spočívá v tom, že se na začátku zpětného výpočtu ze vztahu $i \in \bar{I}_j \;\Leftrightarrow\; j \in I_i$ určí množiny $\bar{I}_j$, $1 \le j \le l-1$. Pak se položí $\bar{v}_l = 1$ a pro $l - 1 \ge i \ge 1$ (ve zpětném směru) se počítají veličiny $\bar{v}_j$, $l - 1 \ge j \ge 1$, podle vzorců (1093). Nakonec se položí $g_i(x) = \bar{v}_i$, $1 \le i \le n$.

Druhý způsob je algoritmicky jednoduušší, neboť není třeba určovat množiny $\bar{I}_j$, $1 \le j \le l-1$. Tento způsob je založen na postupném načítání hodnot do proměnných $\bar{v}_j$, $1 \le j \le l-1$, které je třeba v přípravné fázi vynulovat. Pak se položí $\bar{v}_l = 1$ a pro $l - 1 \ge i \ge 1$ (ve zpětném směru) se aktualizují veličiny $\bar{v}_j$, $j \in I_i$, podle vzorců

$$\bar{v}_j := \bar{v}_j + \bar{v}_i \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j}, \qquad j \in I_i. \tag{1094}$$

Nakonec se položí $g_i(x) = \bar{v}_i$, $1 \le i \le n$.

Oba dva způsoby zpětného výpočtu gradientu jsou výpočetně ekvivalentní. Z tabulky 25 uvedené v příkladu 11 je patrné, že pro počet operací platí $\mathrm{Op}(F(x)) + \mathrm{Op}(g(x)) \approx 3\,\mathrm{Op}(F(x))$ (teoreticky lze dokázat že $\mathrm{Op}(F(x)) + \mathrm{Op}(g(x)) \le 3\,\mathrm{Op}(F(x))$).

Zpětné automatické derivování můžeme snadno zobecnit pro výpočet součinu transponoané Jacobiovy matice $J^T(x)$ a vektoru w. V tomto případě na začátku výpočetní fáze pokládáme $\bar{v}_i = 0$, $1 \le i \le l - m$, $\bar{v}_{l-m+i} = w_i$, $1 \le i \le m$, a pak pro $l - 1 \ge i \ge n + 1$ aktualizujeme veličiny $\bar{v}_j = \partial f^T w/\partial v_j$, $j \in I_i$, podle vzorců (1094).

Zpětné automatické derivování není opět vhodné pro výpočet prvků Jacobiovy matice zobrazení, neboť by bylo třeba počítat m součinů $J^T(x)e_k$ $1 \leq k \leq m$. Situace se však zjednoduší, je-li Jacobiova matice řídká (takže funkce $f_k(x)$, $1 \leq k \leq m$, jsou separovatelné). V tomto případě mají gradienty $\nabla f_k(x)$, $1 \leq k \leq m$, pouze $n_k = O(1)$ nenulových prvků (neboť $e_i^T \nabla f_k(x) = 0$, pokud $i \notin N_k$) a jejich výpočet vyžaduje vyhodnocení $m\,O(1) = O(m)$ výrazů (jednotlivé funkce f_k, $1 \leq k \leq m$, se zpracovávají samostatně).

Příklad 11. Pomocí zpětného automatického derivování vypočteme gradient funkce $F : R \to R^4$ zadané předpisem $F(x) = x_1 \sin(x_2 x_3 + x_1 \exp(x_4))$.

v_1	$=$	x_1	
v_2	$=$	x_2	
v_3	$=$	x_3	
v_4	$=$	x_4	
$\bar{v}$	$=$	0	
v_5	$=$	$v_2 v_3$	$x_2 x_3$
v_6	$=$	$\exp(v_4)$	$\exp(x_4)$
v_7	$=$	$v_1 v_6$	$x_1 \exp(x_4)$
v_8	$=$	$v_5 + v_7$	$x_2 x_3 + x_1 \exp(x_4)$
v_9	$=$	$\sin(v_8)$	$\sin(x_2 x_3 + x_1 \exp(x_4))$
v_{10}	$=$	$v_1 v_9$	$x_1 \sin(x_2 x_3 + x_1 \exp(x_4))$
$F(x)$	$=$	v_{10}	$x_1 \sin(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_{10}$	$:=$	1	1
$\bar{v}_9$	$:=$	$\bar{v}_9 + \bar{v}_{10} v_1$	x_1
$\bar{v}_1$	$:=$	$\bar{v}_1 + \bar{v}_{10} v_9$	$\sin(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_8$	$:=$	$\bar{v}_8 + \bar{v}_9 \cos(v_8)$	$x_1 \cos(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_7$	$:=$	$\bar{v}_7 + \bar{v}_8$	$x_1 \cos(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_5$	$:=$	$\bar{v}_5 + \bar{v}_8$	$x_1 \cos(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_6$	$:=$	$\bar{v}_6 + \bar{v}_7 v_1$	$x_1^2 \cos(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_1$	$:=$	$\bar{v}_1 + \bar{v}_7 v_6$	$x_1 \cos(x_2 x_3 + x_1 \exp(x_4)) \exp(x_4) + \sin(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_4$	$:=$	$\bar{v}_4 + \bar{v}_6 \exp(v_4)$	$x_1^2 \cos(x_2 x_3 + x_1 \exp(x_4)) \exp(x_4)$
$\bar{v}_3$	$:=$	$\bar{v}_3 + \bar{v}_5 v_2$	$x_1 x_2 \cos(x_2 x_3 + x_1 \exp(x_4))$
$\bar{v}_2$	$:=$	$\bar{v}_2 + \bar{v}_5 v_3$	$x_1 x_3 \cos(x_2 x_3 + x_1 \exp(x_4))$
$g_1(x)$	$=$	$\bar{v}_1$	$x_1 \cos(x_2 x_3 + x_1 \exp(x_4)) \exp(x_4) + \sin(x_2 x_3 + x_1 \exp(x_4))$
$g_2(x)$	$=$	$\bar{v}_2$	$x_1 x_3 \cos(x_2 x_3 + x_1 \exp(x_4))$
$g_3(x)$	$=$	$\bar{v}_3$	$x_1 x_2 \cos(x_2 x_3 + x_1 \exp(x_4))$
$g_4(x)$	$=$	$\bar{v}_4$	$x_1^2 \cos(x_2 x_3 + x_1 \exp(x_4)) \exp(x_4)$

Tabulka 25: Zpětný výpočet gradientu funkce $F(x) = x_1 \sin(x_2 x_3 + x_1 \exp(x_4))$.

Příklad 12. Uvažujme zobrazení uvedené v příkladu 10. Pomocí zpětného automatického derivování vypočteme součin matice $J^T(x)$ a vektoru $q = [1, 0, 1]^T$.

v_1	$=$	x_1	
v_2	$=$	x_2	
v_3	$=$	x_2	
$\bar{v}$	$=$	0	
v_4	$=$	$\sin(v_1)$	$\sin(x_1)$
v_5	$=$	$\exp(v_2)$	$\exp(x_2)$
v_6	$=$	$\cos(v_2)$	$\cos(x_2)$
v_7	$=$	$v_1^2 - v_2$	$x_1^2 - x_2$
v_8	$=$	$v_4 v_5$	$\sin(x_1)\exp(x_2)$
v_9	$=$	v_6 / v_3	$\cos(x_2)/x_3$
v_{10}	$=$	v_7^2	$(x_1^2 - x_2)^2$
$f_1(x)$	$=$	v_8	$\sin(x_1)\exp(x_2)$
$f_2(x)$	$=$	v_9	$\cos(x_2)/x_3$
$f_3(x)$	$=$	v_{10}	$(x_1^2 - x_2)^2$
$\bar{v}_{10}$	$:=$	1	1
$\bar{v}_9$	$:=$	0	0
$\bar{v}_8$	$:=$	1	1
$\bar{v}_7$	$:=$	$\bar{v}_7 + 2\bar{v}_{10} v_7$	$2(x_1^2 - x_2)$
$\bar{v}_6$	$:=$	$\bar{v}_6 + \bar{v}_9 / v_3$	0
$\bar{v}_3$	$:=$	$\bar{v}_3 + -\bar{v}_9 / v_3^2$	0
$\bar{v}_4$	$:=$	$\bar{v}_4 + \bar{v}_8 v_5$	$\exp(x_2)$
$\bar{v}_5$	$:=$	$\bar{v}_5 + \bar{v}_8 v_4$	$\sin(x_1)$
$\bar{v}_1$	$:=$	$\bar{v}_1 + 2\bar{v}_7 v_1$	$4x_1(x_1^2 - x_2)$
$\bar{v}_2$	$:=$	$\bar{v}_2 + -\bar{v}_7$	$-2(x_1^2 - x_2)$
$\bar{v}_2$	$:=$	$\bar{v}_2 + -\bar{v}_6 \sin(v_2)$	$-2(x_1^2 - x_2)$
$\bar{v}_2$	$:=$	$\bar{v}_2 + \bar{v}_5 \exp(v_2)$	$-2(x_1^2 - x_2) + \sin(x_1)\exp(x_2)$
$\bar{v}_1$	$:=$	$\bar{v}_1 + \bar{v}_4 \cos(v_1)$	$2x_1(x_1^2 - x_2) + \cos(x_1)\exp(x_2)$
$e_1^T J^T(x) q$	$=$	$\bar{v}_1$	$2x_1(x_1^2 - x_2) + \cos(x_1)\exp(x_2)$
$e_2^T J^T(x) q$	$=$	$\bar{v}_2$	$-2(x_1^2 - x_2) + \sin(x_1)\exp(x_2)$
$e_3^T J^T(x) q$	$=$	$\bar{v}_3$	0

Tabulka 26: Násobení transponované Jacobiovy matice vektorem.

14.2 Automatický výpočet druhých derivací

Druhé derivace můžeme získat aplikací zpětného automatického derivování na výrazy získané přímým automatickým derivováním. Tento způsob, který se používá k určení gradientu směrové derivace $\nabla F'(x, d)$, což odpovídá násobení Hessovy matice $G(x)$ vektorem d, je založen na následující větě.

Věta 279. *Nechť pro $n + 1 \leq i \leq l$ platí (1091)–(1092). Označme $\bar{v}_i = \partial F(x)/\partial v_i$, $\bar{v}_i' = (\partial F(x)/\partial v_i)'$, $1 \leq i \leq l$, kde čárka znamená zvolenou derivaci. Pak platí $\bar{v}_l = 1$, $\bar{v}_l' = 0$ a*

$$\bar{v}_j = \sum_{i \in \bar{I}_j} \bar{v}_i \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j}. \tag{1095}$$

$$\bar{v}_j' = \sum_{i \in \bar{I}_j} \left(\bar{v}_i' \frac{\partial \varphi_i(v_j, j \in I_i)}{\partial v_j} + \bar{v}_i \sum_{k \in I_i} \frac{\partial^2 \varphi_i(v_j, j \in I_i)}{\partial v_j \partial v_k} v_k' \right), \tag{1096}$$

kde $l - 1 \geq j \geq 1$.

Důkaz Zřejmě $\bar{v}_l = \partial F(x)/\partial v_l = \partial v_l/\partial v_l = 1$, takže $\bar{v}'_l = 0$. Aplikujeme-li pravidlo řetězení na (1095), pak pro $l-1 \leq j \leq 1$ dostaneme

$$\begin{aligned}\bar{v}'_j &= \left(\sum_{i\in\bar{I}_j} \bar{v}_i \frac{\partial\varphi_i(v_j, j\in I_i)}{\partial v_j}\right)' \\ &= \sum_{i\in\bar{I}_j}\left(\bar{v}'_i\frac{\partial\varphi_i(v_j, j\in I_i)}{\partial v_j} + \bar{v}_i\sum_{k\in I_i}\frac{\partial^2\varphi_i(v_j, j\in I_i)}{\partial v_j\partial v_k}v'_k\right).\end{aligned}$$

□

Při výpočtu gradientu směrove derivace se nejprve v přímém směru napočítají hodnoty v_i, v'_i, $1 \leq i \leq l$ a pak se ve zpětném směru kromě hodnot $\bar{v}_j$, $l-1 \geq j \geq 1$, počítají jejich derivace $\bar{v}'_j$, $l-1 \geq j \geq 1$ tak, že se řádek (1094), kde $j \in I_i$, nahradí dvěma řádky

$$\bar{v}_j := \bar{v}_j + \bar{v}_i\frac{\partial\varphi_i(v_j, j\in I_i)}{\partial v_j}, \tag{1097}$$

$$\bar{v}'_j := \bar{v}'_j + \bar{v}'_i\frac{\partial\varphi_i(v_j, j\in I_i)}{\partial v_j} + \bar{v}_i\sum_{k\in I_i}\frac{\partial^2\varphi_i(v_j, j\in I_i)}{\partial v_j\partial v_k}v'_k, \tag{1098}$$

kde $l-1 \geq j \geq 1$. Na začátku přímého výpočtu (v přípravné fázi) se pokládá $v_i = x_i$, $v'_i = d_i$, $1 \leq i \leq n$ a $\bar{v} = 0$, $\bar{v}' = 0$. Pro výpočet druhých derivací elementárních funkcí se používají pravidla uvedená v tabulce 27.

$F(x)$	$F''(x)$	$\mathcal{D}_F$	$F(x)$	$F''(x)$	$\mathcal{D}_F$
e^x	e^x	R	a^x	$a^x \log^2 a$	R
$\log x$	$\frac{-1}{x^2}$	$x > 0$	$\log_a x$	$\frac{-1}{x^2 \log a}$	$x > 0$
$\sin x$	$-\sin x$	R	$\cos x$	$-\cos x$	R
$\operatorname{tg} x$	$\frac{2\sin x}{\cos^3 x}$	$\cos x \neq 0$	$\operatorname{cotg} x$	$\frac{2\cos x}{\sin^3 x}$	$\sin x \neq 0$
$\sinh x$	$\sinh x$	R	$\cosh x$	$\cosh x$	R
$\operatorname{tgh} x$	$\frac{-2\sinh x}{\cosh^3 x}$	R	$\operatorname{cotgh} x$	$\frac{2\cosh x}{\sinh^3 x}$	$x \neq 0$
$\arcsin x$	$\frac{x}{\sqrt{(1-x^2)^3}}$	$\lvert x\rvert < 1$	$\arccos x$	$\frac{-x}{\sqrt{(1-x^2)^3}}$	$\lvert x\rvert < 1$
$\operatorname{arctg} x$	$\frac{2x}{(1+x^2)^2}$	R	$\operatorname{arccotg} x$	$\frac{-2x}{(1+x^2)^2}$	R
$\operatorname{argsinh} x$	$\frac{-x}{\sqrt{(x^2+1)^2}}$	R	$\operatorname{argcosh} x$	$\frac{-x}{\sqrt{(x^2-1)^2}}$	$x > 1$
$\operatorname{argtgh} x$	$\frac{2x}{(1-x^2)^2}$	$\lvert x\rvert < 1$	$\operatorname{argcotgh} x$	$\frac{2x}{(1-x^2)^2}$	$\lvert x\rvert > 1$

Tabulka 27: Derivování elementárních funkcí.

Příklad 13. Pomocí přímého a zpětného automatického derivování vypočteme součin Hessovy matice Rosenbrockovy funkce $F(x) = 100(x_1^2 - x_2)^2 + (x_1 - 1)^2$ a vektoru $d = [1, 1]^T$.

v_1	$=$	x_1	
v_1'	$=$	1	
v_2	$=$	x_2	
v_2'	$=$	1	
$\bar{v}$	$=$	0	
$\bar{v}'$	$=$	0	
v_3	$=$	v_1^2	x_1^2
v_3'	$=$	$2v_1v_1'$	$2x_1$
v_4	$=$	$v_3 - v_2$	$x_1^2 - x_2$
v_4'	$=$	$v_3' - v_2'$	$2x_1 - 1$
v_5	$=$	$100v_4^2$	$100(x_1^2 - x_2)^2$
v_5'	$=$	$200v_4v_4'$	$200(x_1^2 - x_2)(2x_1 - 1)$
v_6	$=$	$(v_1 - 1)^2$	$(x_1 - 1)^2$
v_6'	$=$	$2(v_1 - 1)v_1'$	$2(x_1 - 1)$
v_7	$=$	$v_5 + v_6$	$100(x_1^2 - x_2)^2 + (x_1 - 1)^2$
v_7'	$=$	$v_5' + v_6'$	$200(x_1^2 - x_2)(2x_1 - 1) + 2(x_1 - 1)$
$\bar{v}_7$	$=$	1	1
$\bar{v}_5$	$=$	$\bar{v}_5 + \bar{v}_7$	1
$\bar{v}_5'$	$=$	$\bar{v}_5' + \bar{v}_7'$	0
$\bar{v}_6$	$=$	$\bar{v}_6 + \bar{v}_7$	1
$\bar{v}_6'$	$=$	$\bar{v}_6' + \bar{v}_7'$	0
$\bar{v}_1$	$=$	$\bar{v}_1 + \bar{v}_6\varphi_6'$	$2(x_1 - 1)$
$\bar{v}_1'$	$=$	$\bar{v}_1' + \bar{v}_6'\varphi_6' + \bar{v}_6\varphi_6''v_1'$	2
$\bar{v}_4$	$=$	$\bar{v}_4 + \bar{v}_5\varphi_5'$	$200(x_1^2 - x_2)$
$\bar{v}_4'$	$=$	$\bar{v}_4' + \bar{v}_5'\varphi_5' + \bar{v}_5\varphi_5''v_4'$	$200(2x_1 - 1)$
$\bar{v}_2$	$=$	$\bar{v}_2 - \bar{v}_4$	$-200(x_1^2 - x_2)$
$\bar{v}_2'$	$=$	$\bar{v}_2' - \bar{v}_4'$	$-200(2x_1 - 1)$
$\bar{v}_3$	$=$	$\bar{v}_3 + \bar{v}_4$	$200(x_1^2 - x_2)$
$\bar{v}_3'$	$=$	$\bar{v}_3' + \bar{v}_4'$	$200(2x_1 - 1)$
$\bar{v}_1$	$=$	$\bar{v}_1 + \bar{v}_3\varphi_3'$	$2(x_1 - 1) - 400x_1(x_1^2 - x_2)$
$\bar{v}_1'$	$=$	$\bar{v}_1' + \bar{v}_3'\varphi_3' + \bar{v}_3\varphi_3''v_1'$	$2 + 400x_1(2x_1 - 1) + 400x_1(x_1^2 - x_2)$
$e_1^T G(x)d$	$=$	$\bar{v}_1'$	$2 + 400x_1(2x_1 - 1) + 400x_1(x_1^2 - x_2)$
$e_2^T G(x)d$	$=$	$\bar{v}_2'$	$-200(2x_1 - 1)$

Tabulka 28: Násobení Hessovy matice Rosenbrockovy funkce vektorem.

Chceme-li vypočítat všechny prvky Hessovy matice, je třeba počítat gradienty parciálních derivací $\partial F(x)/\partial x_i$, $1 \leq i \leq n$, což vyžaduje n krát více operací než výpočet gradientu jedné směrové derivace. Je-li Hessova matice řídká, můžeme použít postup uvedený v oddílu 10.2. Prvky Hessovy matice vypočteme zpětným automatickým derivováním směrových derivací $G(x)v_i$, $1 \leq i \leq k$, kde v_i, $1 \leq i \leq k$, jsou vektory obsahující pouze nuly a jednotky takové, že $(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$ (vzorec (887)).

14.3 Numerický výpočet gradientu

Numerický výpočet parciálních derivací $g_i(x) = \partial F(x)/\partial x_i$, $i \leq i \leq n$, lze provádět podle různých diferenčních vzorců, jejichž přesnost roste s rostoucím počtem použitých funkčních hodnot. Protože výpočet funkčních hodnot bývá často limitujícím faktorem optimalizačního procesu, používáme pouze jednoduché diferenční vzorce, z nichž nejjednodušší má tvar

$$g_i(x) \approx \frac{F(x + \delta_i e_i) - F(x)}{\delta_i}\,, \tag{1099}$$

kde $e_i \in R^n$ je i-tý sloupec jednotkové matice řádu n. Číslo δ_i je třeba volit dostatečně malé, aby diference (1099) dostatečně přesně aproximovala parciální derivaci $g_i(x)$. V počátečních fázích optimalizačního procesu můžeme položit

$$\delta_i = \sqrt{\varepsilon_M}\max(1, |x_i|)\,,$$

kde ε_M je strojová přesnost. Vzhledem k tomu, že podle věty 8 platí $\|g(x)\| \to 0$, pokud $x \to x^*$, klesá v konečných fázích optimalizačního procesu relativní přesnost diferenčního vzorce (1099). Je proto nutné věnovat velkou pozornost výběru čísla δ_i tak, aby tato přesnost byla co největší. V dalším výkladu budeme pro jednoduchost předpokládat, že $\delta_i > 0$ a $G_{ii}(x) > 0$ ($G_{ii}(x)$ je i-tý diagonální prvek Hessovy matice).

Přesnost diferenčního vzorce (1099) ovlivňují dva druhy chyb. Linearizační chyba $\tau(x, \delta_i)$ vzniká zanedbáním nelineární části Taylorova rozvoje a její velikost je určena vztahem

$$\tau(x,\delta_i) = \left|\frac{F(x+\delta_i e_i) - F(x)}{\delta_i} - g_i(x)\right| = \frac{\delta_i}{2}\,|G_{ii}(x + \hat{\lambda}\delta_i e_i)| \approx \frac{\delta_i}{2}\,G_{ii}(x)\,, \tag{1100}$$

kde $0 \le \hat{\lambda} \le 1$ (tvrzení 2). Zaokrouhlovací chyba $\varrho(x, \delta_i)$ vzniká nepřesným výpočtem funkčních hodnot a její velikost je určena vztahem

$$\begin{aligned}\varrho(x,\delta_i) &= \left|\frac{\overline{F}(x+\delta_i e_i) - \overline{F}(x)}{\delta_i} - \frac{F(x+\delta_i e_i) - F(x)}{\delta_i}\right| \le \\ &\le \frac{1}{\delta_i}\left(|\overline{F}(x+\delta_i e_i) - F(x+\delta_i e_i)| + |\overline{F}(x) - F(x)|\right) \approx \frac{2}{\delta_i}\,\varepsilon_A(x)\,,\end{aligned} \tag{1101}$$

kde $\varepsilon_A(x) = |\overline{F}(x) - F(x)|$ je absolutní chyba výpočtu funkční hodnoty ($\overline{F}(x)$ je vypočtená funkční hodnota a $F(x)$ je skutečná funkční hodnota). Z těchto úvah je zřejmé, že $\tau(x, \delta_i)$ roste a $\varrho(x, \delta_i)$ klesá se zvětšující se hodnotou čísla δ_i. Optimální hodnotu čísla δ_i volíme tak, aby platilo $\tau(x, \delta_i) = \varrho(x, \delta_i)$, což spolu s (1100) a (1101) dává

$$\delta_i = 2\sqrt{\frac{\varepsilon_A(x)}{G_{ii}(x)}}\,. \tag{1102}$$

Ve vzorci (1102) vystupují funkce $\varepsilon_A(x)$ a $G_{ii}(x)$, které obvykle neznáme. Proto je nutné tyto funkce nějakým způsobem aproximovat. Označme $\varepsilon_R(x) = \varepsilon_A(x)/|F(x)|$. Nedochází-li při výpočtu funkční hodnoty ke ztrátě přesnosti vlivem odčítání dvou přibližně stejných čísel, můžeme předpokládat, že platí $\varepsilon_R(x) = \varepsilon_R$, kde obvykle $\varepsilon_R \approx \varepsilon_M$. Pak můžeme psát

$$\varepsilon_A(x) = \varepsilon_R |F(x)|\,. \tag{1103}$$

Tento model však není obecně přijatelný, jak je ukázáno v následujícím příkladu.

Příklad 14. Uvažujme funkci $F : R^n \to R$ definovanou vztahem

$$F(x) = \sum_{k=1}^{m} |f_k(x) - y_k|\,,$$

kde $f_k : R^n \to R$, $1 \le k \le m$, jsou funkce pro něž lze použít model (1103) a y_k, $1 \le k \le m$, jsou hodnoty takové, že součet jejich absolutních hodnot je dostatečně veliký. Pak platí

$$\begin{aligned}\varepsilon_A(x) &= |\overline{F}(x) - F(x)| = \left|\sum_{k=1}^{m} |\overline{f}_k(x) - y_k| - \sum_{k=1}^{m} |f_k(x) - y_k|\right| \le \\ &\le \sum_{k=1}^{m} \left||\overline{f}_k(x) - y_k| - |f_k(x) - y_k|\right| \le \sum_{k=1}^{m} |\overline{f}_k(x) - f_k(x)| = \\ &= \varepsilon_R \sum_{k=1}^{m} |f_k(x)|\,,\end{aligned}$$

přičemž může nastat i rovnost. Předpokládejme, že pro $x \to x^*$ platí $f_k(x) \to y_k$, $1 \le k \le m$. Pak v dostatečné blízkosti bodu $x^* \in R^n$ platí

$$\varepsilon_A(x) \approx \varepsilon_R \sum_{k=1}^{m} |f_k(x)| \approx \varepsilon_R \sum_{k=1}^{m} |y_k| = \varepsilon_A$$

a funkce $\varepsilon_A(x) \approx \varepsilon_A$ je v podstatě konstantní, zatímco funkce $\varepsilon_R(x) \approx \varepsilon_A/|F(x)|$ roste nade všechny meze.

Abychom zahrnuli podobné případy, musíme model (1103) poněkud upravit. Můžeme použít například model

$$\varepsilon_A(x) = \varepsilon_R(\varepsilon_P + |F(x)|)\,, \tag{1104}$$

kde ε_P je mez, pod níž se přestává projevovat vliv hodnoty $|F(x)|$. Pokud neznáme nic bližšího o funkci $\varepsilon_A(x)$, můžeme zvolit $\varepsilon_P = \sqrt{\varepsilon_R}$.

Funkci $G_{ii}(x)$ (i-tý diagonální prvek Hessovy matice) můžeme aproximovat podle vzorce

$$G_{ii}(x) \approx B_{ii}\,, \tag{1105}$$

kde B_{ii} je i-tý diagonální prvek matice B, kterou používáme v inverzních metodách s proměnnou metrikou, jež jsou popsány v oddílu 4.1 (vzorec (306)).

Použijeme-li vzorce (1100), (1101) a (1102), můžeme určit relativní chybu výpočtu parciální derivace. Platí

$$\frac{\tau(x,\delta_i) + \varrho(x,\delta_i)}{|g_i(x)|} \approx \frac{2\sqrt{\varepsilon_A(x) G_{ii}(x)}}{|g_i(x)|}\,.$$

Je-li tato chyba větší než povolená mez, musíme použít přesnější diferenční vzorec

$$g_i(x) \approx \frac{F(x + \delta_i e_i) - F(x - \delta_i e_i)}{2\delta_i}\,. \tag{1106}$$

Přesnost tohoto diferenčního vzorce je ovlivněna linearizační chybou

$$\tau(x,\delta_i) \approx \frac{\delta_i}{3}\, T_{iii}(x)\,.$$

kde $T_{iii}(x)$ je i-tý diagonální prvek tenzoru třetích derivací (předpokládejme pro jednoduchost, že platí $T_{iii}(x) > 0$), a zaokrouhlovací chybou

$$\varrho(x,\delta_i) \approx \frac{1}{\delta_i}\, \varepsilon_A(x)\,.$$

Součet těchto chyb je minimální, volíme-li číslo δ_i podle vzorce

$$\delta_i = \sqrt[3]{\frac{3\varepsilon_A(x)}{T_{iii}(x)}}\,. \tag{1107}$$

Označme δ_i^A číslo určené podle vzorce (1102) a δ_i^B číslo určené podle vzorce (1107). Předpokládáme-li, že absolutní chyba výpočtu $\varepsilon_A(x)$ je v obou případech stejná, dostaneme

$$\frac{1}{4}\,(\delta_i^A)^2 G_{ii}(x) = \frac{1}{3}\,(\delta_i^B)^3 T_{iii}(x)\,.$$

Neliší-li se hodnoty $G_{ii}(x)$ a $T_{iii}(x)$ o mnoho řádů, můžeme psát

$$\delta_i^B \approx (\delta_i^A)^{2/3}\,. \tag{1108}$$

Nevyhovuje-li nám tedy diferenční vzorec (1099) s hodnotou $\delta_i = \delta_i^A$, můžeme použít diferenční vzorec (1101) s hodnotou $\delta_i = \delta_i^B$ určenou podle vzorce (1108).

Používáme-li diferenční vzorec (1106), máme k dispozici funkční hodnoty $F(x + \delta_i e_i)$, $F(x - \delta_i e_i)$ a $F(x)$. Tyto funkční hodnoty můžeme použít k aproximaci i-tého diagonálního prvku Hessovy matice. Platí

$$G_{ii}(x) \approx \frac{F(x + \delta_i e_i) + F(x - \delta_i e_i) - 2F(x)}{\delta_i^2} . \tag{1109}$$

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 29. Data $x \in R^n$, $\varepsilon_A > 0$, $k \in \{0, 1\}$, $l_i \in \{1, 2\}$, $1 \leq i \leq n$.

Krok 1 Vypočteme hodnotu ε_A podle vzorce (1103) nebo (1104). Položíme $i := 1$.

Krok 2 Jestliže $l_i = 1$, a $k = 0$, položíme $G_{ii} := 1$. Jestliže $l_i = 1$ a $k = 1$, položíme $G_{ii} := |B_{ii}|$. Jestliže $l_i = 2$, položíme $G_{ii} := |G_{ii}|$.

Krok 3 Položíme $l_i := 1$. Vypočteme číslo δ_i podle vzorce (1102) a prvek g_i podle vzorce (1099).

Krok 4 Jestliže $2\sqrt{\varepsilon_A G_{ii}}/|g_i| \leq \overline{\varepsilon}$, přejdeme na krok 6.

Krok 5 Položíme $l_i := 2$ a $\delta_i := \sqrt[3]{\delta_i^2}$. Vypočteme prvek g_i podle vzorce (1106) a číslo G_{ii} podle vzorce (1109).

Krok 6 Jestliže $i = n$, ukončíme výpočet. Jestliže $i < n$, položíme $i := i + 1$ a přejdeme na krok 2.

V algoritmu 29 používáme čísla $\overline{\varepsilon} > 0$, $k \in \{0, 1\}$ a vektor $l \in \{1, 2\}^n$. Číslo $\overline{\varepsilon} > 0$ udává maximální přípustnou relativní chybu prvku gradientu (obvykle se používá hodnota $\overline{\varepsilon} = 10^{-2}$). Číslo k udává, zda je ($k = 1$) nebo není ($k = 0$) k dispozici matice $B \approx G(x)$. Tuto matici používáme v inverzních metodách s proměnnou metrikou, které jsou popsány v oddílu 4.1. Vektor l obsahuje informace o tom, zda byl v předchozím výpočtu prvku g_i použit vzorec (1099) ($l_i = 1$), nebo vzorec (1106) ($l = 2$). Před prvním použitím algoritmu 29 je nutné položit $l_i = 1$, $1 \leq i \leq n$.

15 Základy nehladké analýzy

V této kapitole probereme základní pojmy konečněrozměrné nehladké analýzy a jejich aplikace na řešení systémů nehladkých rovnic a na nepodmíněnou minimalizaci nehladkých funkcí. V důkazech některých vět budeme používat dva klasické výsledky, jejichž konečněrozměrné verze zde uvedeme (připomeňme, že lineární forma $l(x)$ je lineární funkce taková, že $l(0) = 0$).

Věta 280. *(Hahn-Banach). Nechť funkce $F : R^n \to R$ je pozitivně homogenní (platí $F(\lambda x) = \lambda F(x)$ pokud $\lambda \geq 0$) a subaditivní (platí $F(x_1 + x_2) \leq F(x_1) + F(x_2)$). Nechť $X \subset R^n$ je podprostor a $l : X \to R$ je lineární forma taková, že $l(x) \leq F(x)\ \forall x \in X$. Pak existuje lineární forma $l_n : R^n \to R$ taková, že $l_n(x) = l(x)\ \forall x \in X$ a $l_n(x) \leq F(x)\ \forall x \in R^n$.*

Důkaz Předpokládejme, že podprostor X má dimenzi m, kde $1 \leq m < n$, a označme $X_m = X$. Důkaz provedeme indukcí. Nechť $X_m \subset X_i \subset X_{i+1} \subset X_n = R^n$, kde dimenze uvedených podprostorů jsou rovny jejich indexům, přičemž $X_{i+1} = X_i \oplus \mathcal{L}(y_i)$, kde $y_i \in X_i^\perp$. Předpokládejme, že existuje lineární forma $l_i(x)$ taková, že $l_i(x) = l(x)\ \forall x \in X$ a $l_i(x) \leq F(x)\ \forall x \in X_i$ (platí to zřejmě pro $i = m$). Zvolme libovolně vektor $x_{i+1} \in X_{i+1}$, který lze jednoznačně vyjádřit ve tvaru

$$x_{i+1} = x_i + \lambda_i y_i,$$

kde $x_i \in X_i$, $y_i \in X_i^\perp$ a $\lambda_i \in R$. Nechť $l_{i+1} : X^{i+1} \to R$ je lineární forma taková, že $l_{i+1}(x) = l_i(x)$ $\forall x \in X_i$. Pak

$$l_{i+1}(x_{i+1}) = l_i(x_i) + \lambda_i l_{i+1}(y_i) = l_i(x_i) + \lambda_i c_i,$$

kde číslo $c_i = l_{i+1}(y_i)$, určující jednoznačně funkci l_{i+1}, nezávisí na volbě x_i a λ_i. Toto číslo je třeba určit tak, aby platilo $l_{i+1}(x_{i+1}) \leq F(x_{i+1})$, neboli $l_i(x_i) + \lambda_i c_i \leq F(x_i + \lambda_i y_i)$. Pokud $\lambda_i = 0$, je tato nerovnost splněna triviálně. V opačném případě, vydělíme-li tuto nerovnost číslem $|\lambda_i|$ a přihlédneme-li k pozitivní homogenitě funkce F, můžeme psát

$$l_i(\tilde{x}_i) + \sigma_i c_i \leq F(\tilde{x}_i + \sigma_i y_i),$$

kde $\tilde{x}_i = x_i/|\lambda_i| \in X_i$ a $\sigma_i = \text{sign}(\lambda_i)$. Pokud $\lambda_i > 0$ (takže $\sigma_i = 1$), dostaneme $c_i \leq F(\tilde{x}_i + y_i) - l_i(\tilde{x}_i)$. Pokud $\lambda_i < 0$ (takže $\sigma_i = -1$), dostaneme $c_i \geq l_i(\tilde{x}_i) - F(\tilde{x}_i - y_i)$. Tyto nerovnosti jsou splněny pokud

$$l(\tilde{x}_i) - F(\tilde{x}_i - y_i) \leq c_i \leq F(\tilde{x}_i + y_i) - l(\tilde{x}_i)$$

a číslo c_i vyhovující těmto nerovnostem existuje, platí-li $l(\tilde{x}_i) - F(\tilde{x}_i - y_i) \leq F(\tilde{x}_i + y_i) - l(\tilde{x}_i)$. Tato nerovnost je splněna, neboť $\tilde{x}_i \in X_i$, takže $l_i(\tilde{x}_i) \leq F(\tilde{x}_i)$, a použitím pozitivní homogenity a subaditivity funkce F dostaneme

$$F(\tilde{x}_i + y_i) - 2l(\tilde{x}_i) + F(\tilde{x}_i - y_i) \geq F(\tilde{x}_i + y_i + \tilde{x}_i - y_i) - 2l(\tilde{x}_i) = 2(F(\tilde{x}_i) - l(\tilde{x}_i)) \geq 0.$$

Tím jsme provedli indukční krok a dokázali tak indukcí existenci lineární formy $l_n : R^n \to R$ takové, že $l_n(x) = l(x)\ \forall x \in X$ a $l_n(x) \leq F(x)\ \forall x \in R^n$. □

Důsledek 32. *Nechť funkce $F : R^n \to R$ je pozitivně homogenní a subaditivní. Pak existuje vektor $g \in R^n$ takový, že $g^T x \leq F(x)\ \forall x \in R^n$.*

Důkaz Zvolme $X = \{0\}$ (podprostor nulové dimenze) a $l : X \to R$, kde $l(0) = 0$. Jelikož pro každou pozitivně homogenní funkci platí $F(0) = 0$, je splněna nerovnost $l(0) \leq F(0)$. Podle Hahn-Banachovy věty existuje lineární forma $l_n : R^n \to R$ taková, že $l_n(x) = l(x)\ \forall x \in X$ a $l_n(x) \leq F(x)\ \forall x \in R^n$. Zbytek tvrzení plyne z toho, že lineární formu v R^n lze vyjádřit skalárním součinem $l_n(x) = g^T x$, kde $g \in R^n$. □

Tvrzení 10. *(Rademacher). Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská (definice 116) v oblasti $\Omega \in R^n$. Pak F je diferencovatelná skoro všude v Ω (množina $\{x \in \Omega : \nabla F(x)$ neexistuje$\}$ má Lebesgueovu míru nula).*

Poznámka 396. V R^n, kde $n > 1$, nemusí být množina míry nula nikterak exotická nebo zanedbatelná. Lebesgueovu míru nula mají například podprostory a lineární variety dimenze nižší než n a jejich části, například stěny a hrany konvexních těles. Důležitou vlastností množin míry nula je, že neobsahují podmnožiny otevřené v R^n.

15.1 Konvexní množiny

Definice 94. *Řekněme, že množina $\mathcal{C} \in R^n$ je konvexní, jestliže z $x \in \mathcal{C}$, $y \in \mathcal{C}$ plyne*

$$\lambda x + (1-\lambda)y \in \mathcal{C}, \tag{1110}$$

pokud $0 \leq \lambda \leq 1$, což můžeme zapsat ve tvaru

$$x + (1-\lambda)(y-x) \in \mathcal{C} \quad nebo \quad y + \lambda(x-y) \in \mathcal{C}. \tag{1111}$$

kde $0 \leq \lambda \leq 1$. To znamená, že konvexní množina $\mathcal{C}$ obsahuje úsečku spojující body $x \in \mathcal{C}$ a $y \in \mathcal{C}$.

Definice 95. *Nechť $m \geq 1$, $x_i \in R^n$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \cdots + \lambda_m = 1$. Pak bod*

$$x = \sum_{i=1}^{m} \lambda_i x_i,$$

nazveme konvexní kombinací bodů $x_i \in R^n$, $1 \leq i \leq m$.

Věta 281. *Množina $\mathcal{C} \subset R^n$ je konvexní právě tehdy, obsahuje-li všechny konvexní kombinace svých bodů.*

Důkaz Obsahuje-li množina $\mathcal{C}$ všechny konvexní kombinace svých bodů, obsahuje též konvexní kombinace tvaru (1110), takže je konvexní. Opačnou implikaci dokážeme indukcí. Předpokládejme, že konvexní množina $\mathcal{C}$ obsahuje všechny konvexní kombinace svých m bodů, kde $m \geq 1$ (pro $m = 1$ je to zřejmé, neboť z $x_1 \in \mathcal{C}$ a $\lambda_1 = 1$ plyne $\lambda_1 x_1 = x_1 \in \mathcal{C}$). Pak pro $x_i \in \mathcal{C}$, $\lambda_i \geq 0$, $1 \leq i \leq m+1$, $\lambda_1 + \cdots + \lambda_{m+1} = 1$ můžeme psát

$$\sum_{i=1}^{m+1} \lambda_i x_i = \sum_{i=1}^{m} \lambda_i x_i + \lambda_{m+1} x_{m+1} = (1-\lambda_{m+1}) x'_{m+1} + \lambda_{m+1} x_{m+1} \in \mathcal{C},$$

kde

$$x'_{m+1} = \sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} x_i \triangleq \sum_{i=1}^{m} \lambda'_i x_i \in \mathcal{C},$$

neboť $x_i \in \mathcal{C}$, $\lambda'_i \geq 0$, $1 \leq i \leq m$, $\lambda'_1 + \cdots + \lambda'_m = 1$. □

Poznámka 397. Podobným způsobem, jaký jsme použili v důkazu věty 281, lze ukázat, že konvexní kombinace konvexních kombinací je opět konvexní kombinací.

Poznámka 398. Nechť $x_i \in R^n$, $1 \leq i \leq m$, a $x = \sum_{i=1}^{m} \lambda_i x_i$. Pak bod x nazveme:

(a) Lineární kombinací bodů $x_i \in R^n$, jsou-li koeficienty $\lambda_i \in R$ libovolné.
(b) Nezápornou lineární kombinací bodů $x_i \in R^n$, platí-li $\lambda_i \geq 0$, $1 \leq i \leq m$.
(c) Afinní kombinací bodů $x_i \in R^n$, platí-li $\sum_{i=1}^{m} \lambda_i = 1$.
(d) Konvexní kombinací bodů $x_i \in R^n$, platí-li $\sum_{i=1}^{m} \lambda_i = 1$ a $\lambda_i \geq 0$, $1 \leq i \leq m$.

Tyto kombinace definují po řadě lineární podprostory, konvexní kužely, afinní množiny (lineární variety) a konvexní množiny. Lineárními podprostory a afinními množinami se podrobně zabývat nebudeme (probírají se v kurzech lineární algebry). Připomeňme, že všechny uvedené množiny jsou konvexní a afinní množina je posunutým lineárním podprostorem. To znamená, že je-li množina $\mathcal{C}$ afinní a $x \in \mathcal{C}$, je množina $\mathcal{C} - x$ lineárním podprostorem. Lze tedy definovat dimenzi afinní množiny jako dimenzi odpovídajícího lineárního podprostoru a jelikož konvexní množinu lze vnořit do afinní množiny (kterou dostaneme vynecháním omezení $\lambda_i \geq 0$, $1 \leq i \leq m$) i dimenzi konvexní množiny.

Věta 282. *Průniky a lineární kombinace konvexních (nebo afinních) množin jsou konvexními (nebo afinními) množinami.*

Důkaz Důkaz provedeme pouze pro konvexní množiny. V případě afinních množin stačí v důkazu vypustit podmínku $0 \leq \lambda \leq 1$.

(a) Nechť $\mathcal{C} = \cap_\alpha \mathcal{C}_\alpha$, kde $\mathcal{C}_\alpha \subset R^n$ jsou konvexní množiny. Nechť $x \in \mathcal{C}$, $y \in \mathcal{C}$. Pak platí $x \in \mathcal{C}_\alpha$ a $y \in \mathcal{C}_\alpha$ $\forall \alpha$ a tedy $\lambda x + (1-\lambda)y \in \mathcal{C}_\alpha$ $\forall \alpha$, pokud $0 \leq \lambda \leq 1$. Odtud plyne, že $\lambda x + (1-\lambda)y \in \mathcal{C}$.

(b) Nechť $\mathcal{C} = \sum_{i=1}^m \lambda_i \mathcal{C}_i$, kde $\mathcal{C}_i \subset R^n$ jsou konvexní množiny a $\lambda_i \in R$. Nechť $x \in \mathcal{C}$, $y \in \mathcal{C}$ a $0 \leq \lambda \leq 1$. Pak existují body $x_i \in \mathcal{C}_i$, $y_i \in \mathcal{C}_i$, $1 \leq i \leq m$, takové, že

$$\lambda x + (1-\lambda)y = \lambda \sum_{i=1}^m \lambda_i x_i + (1-\lambda) \sum_{i=1}^m \lambda_i y_i = \sum_{i=1}^m \lambda_i (\lambda x_i + (1-\lambda) y_i) \stackrel{\Delta}{=} \sum_{i=1}^m \lambda_i z_i.$$

Jelikož $x_i \in \mathcal{C}_i$, $y_i \in \mathcal{C}_i$, platí $z_i = \lambda x_i + (1-\lambda) y_i \in \mathcal{C}_i$, $1 \leq i \leq m$, takže $\lambda x + (1-\lambda)y \in \mathcal{C}$. □

Definice 96. *Konvexním (nebo afinním) obalem množiny $\mathcal{C} \subset R^n$ nazveme průnik všech konvexních (nebo afinních) množin obsahujících $\mathcal{C}$.*

Poznámka 399. K označení konvexního a afinního obalu množiny $\mathcal{C}$ budeme používat symboly $\operatorname{conv} \mathcal{C}$ a $\operatorname{aff} \mathcal{C}$. Zřejmě $\mathcal{C} \subset \operatorname{conv} \mathcal{C} \subset \operatorname{aff} \mathcal{C}$.

Věta 283. *Konvexní (nebo afinní) obal množiny $\mathcal{C} \subset R^n$ je množina všech konvexních (nebo afinních) kombinací bodů z $\mathcal{C}$.*

Důkaz Důkaz provedeme pouze pro konvexní množiny. V případě afinních množin stačí v důkazu vypustit podmínku $\lambda_i \geq 0$, $1 \leq i \leq m$. Nechť $\tilde{\mathcal{C}}$ je množina všech konvexních kombinací bodů z $\mathcal{C}$. Jelikož $\tilde{\mathcal{C}}$ je konvexní a $\mathcal{C} \subset \tilde{\mathcal{C}}$, platí $\operatorname{conv} \mathcal{C} \subset \tilde{\mathcal{C}}$. Nechť $y \in \tilde{\mathcal{C}}$, takže $y = \lambda_1 x_1 + \cdots + \lambda_m x_m$, kde $x_i \in \mathcal{C}$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \cdots + \lambda_m = 1$. Jelikož $x_i \in \mathcal{C}_\alpha$, $1 \leq i \leq m$, pro každou konvexní množinu $\mathcal{C}_\alpha \subset R^n$ obsahující $\mathcal{C}$, platí

$$y \in \bigcap_{\mathcal{C} \subset \mathcal{C}_\alpha} \mathcal{C}_\alpha = \operatorname{conv} \mathcal{C},$$

což dává $\tilde{\mathcal{C}} \subset \operatorname{conv} \mathcal{C}$ □

Věta 284. *(Caratheodory) Nechť $y \in \operatorname{conv} \mathcal{C}$ (nebo $y \in \operatorname{aff} \mathcal{C}$), kde $\mathcal{C} \subset R^n$. Pak existuje nejvýše $n+1$ bodů $x_i \in \mathcal{C}$, $1 \leq i \leq n+1$, takových, že y je jejich konvexní (nebo afinní) kombinací.*

Důkaz Důkaz provedeme pouze pro konvexní množiny. V případě afinních množin stačí v důkazu vypustit podmínku $\lambda_i \geq 0$, $1 \leq i \leq m$. Dokážeme, že pokud platí

$$y = \sum_{i=1}^m \lambda_i x_i, \tag{1112}$$

kde $m > n+1$, $x_i \in \mathcal{C}$, $\lambda_i \geq 0$, $1 \leq i \leq m$ a $\lambda_1 + \cdots + \lambda_m = 1$, lze vždy snížit počet bodů v konvexní kombinaci. Jelikož m je přirozené číslo (konečné), dostaneme po konečném počtu takových snížení konvexní kombinaci s nejvýše $n+1$ body. Označme

$$\hat{y} = \begin{bmatrix} y \\ 1 \end{bmatrix}, \quad \hat{x}_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix}, \quad 1 \leq i \leq m.$$

Pak $\hat{y} \in R^{n+1}$ je lineární kombinací vektorů $\hat{x}_i \in R^{n+1}$, $1 \leq i \leq m$ (s kladnými koeficienty). Jelikož $m > n+1$, jsou vektory $\hat{x}_i \in R^{n+1}$, $1 \leq i \leq m$, lineárně závislé. Existují tedy koeficienty α_i, $1 \leq i \leq m$, z nichž alespoň jeden je nenulový tak, že

$$\sum_{i=1}^m \alpha_i \hat{x}_i = 0. \tag{1113}$$

Protože poslední složky vektorů $\hat{x}_i$ jsou jednotkové, musí platit

$$\sum_{i=1}^{m} \alpha_i = 0,$$

takže alespoň jeden z těchto koeficientů je záporný. Použijeme-li (1112) a (1113) dostaneme

$$\hat{y} = \sum_{i=1}^{m} \lambda_i \hat{x}_i = \sum_{i=1}^{m} \lambda_i \hat{x}_i + \lambda \sum_{i=1}^{m} \alpha_i \hat{x}_i = \sum_{i=1}^{m} (\lambda_i + \lambda\alpha_i)\hat{x}_i \triangleq \sum_{i=1}^{m} \lambda'_i \hat{x}_i$$

pro libovolné číslo $\lambda > 0$. Nechť

$$\lambda = -\frac{\lambda_j}{\alpha_j} = \min_{\alpha_i < 0} \left(-\frac{\lambda_i}{\alpha_i} \right).$$

Pak platí $\lambda'_i \geq 0$, $1 \leq i \leq m$, $\lambda'_j = 0$, $\lambda'_1 + \cdots + \lambda'_m = 1$, takže bod y je konvexní kombinací bodů $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_m$, kterých je $m-1$. □

Konvexní množiny mají výhodné topologické vlastnosti týkající se jejich vnitřku a uzávěru (vnitřek a uzávěr množiny $\mathcal{C}$ budeme označovat symboly $\mathcal{C}^\circ$ a $\overline{\mathcal{C}}$).

Věta 285. *Konvexní množina $\mathcal{C} \in R^n$ má neprázdný vnitřek $\mathcal{C}^\circ \subset \mathcal{C}$ právě tehdy, když* $\dim \mathcal{C} = n$.

Důkaz Má-li konvexní množina $\mathcal{C} \in R^n$ dimenzi n, obsahuje $n+1$ bodů $x_i \in \mathcal{C}$, $1 \leq i \leq n+1$, které neleží v lineární varietě dimenze nižší než n. Konvexní obal těchto bodů, který je obsažen v $\mathcal{C}$, je nedegerovaným simplexem a jeho vnitřní body jsou vnitřními body množiny $\mathcal{C}$, takže platí $\mathcal{C}^\circ \neq \emptyset$. Je-li naopak $\mathcal{C}^\circ \neq \emptyset$, obsahuje $\mathcal{C}$ nějakou otevřenou množinu $\mathcal{B}(x, \varepsilon) \subset R^n$, takže $\dim \mathcal{C} = n$. □

Lemma 121. *Nechť $\mathcal{C} \in R^n$ je konvexní množina dimenze n, $x \in \mathcal{C}^\circ$, $\overline{x} \in \overline{\mathcal{C}}$ a $y = \lambda\overline{x} + (1-\lambda)x$, kde $0 \leq \lambda < 1$. Pak $y \in \mathcal{C}^\circ$.*

Důkaz Máme dokázat existenci čísla $\varepsilon > 0$ takového, že $y + \mathcal{B}(0, \varepsilon) \in \mathcal{C}$. Jelikož $\overline{x} \in \overline{\mathcal{C}}$, platí $\overline{x} \in \mathcal{C} + \mathcal{B}(0, \varepsilon)$ pro libovolné číslo $\varepsilon > 0$. To znamená, že

$$y + \mathcal{B}(0,\varepsilon) \subset \lambda(\mathcal{C} + \mathcal{B}(0,\varepsilon)) + (1-\lambda)x + \mathcal{B}(0,\varepsilon) = \lambda\mathcal{C} + (1-\lambda)\left(x + \frac{1+\lambda}{1-\lambda}\mathcal{B}(0,\varepsilon) \right).$$

Nechť $0 \leq \lambda < 1$. Jelikož $x \in \mathcal{C}^\circ$, existuje číslo $\varepsilon > 0$ takové, že $x + ((1+\lambda)/(1-\lambda))\mathcal{B}(0,\varepsilon) \in \mathcal{C}$. Pro toto číslo platí $y + \mathcal{B}(0,\varepsilon) \subset \lambda\mathcal{C} + (1-\lambda)\mathcal{C} \subset \mathcal{C}$, neboť

$$\lambda\mathcal{C} + (1-\lambda)\mathcal{C} = \{ z \in R^n : z = \lambda z_1 + (1-\lambda)z_2,\ z_1 \in \mathcal{C},\ z_2 \in \mathcal{C} \} \subset \mathcal{C}.$$

□

Konvexní a afinní množiny obsažené v R^n mají často dimenzi nižší než n. Takovéto množiny neobsahují vnitřní body a mají tedy prázdný vnitřek. Konvexní a afinní množiny jsou však natolik speciální, že lze zavést obecnější pojem relativních vnitřních bodů.

Definice 97. *Nechť $\mathcal{C} \subset R^n$ je konvexní (nebo afinní) množina. Bod $x \in \mathcal{C}$ nazveme relativním vnitřním bodem množiny $\mathcal{C}$, existuje-li číslo $\varepsilon > 0$ takové, že $\mathcal{B}(x,\varepsilon) \cap \operatorname{aff} \mathcal{C} \subset \mathcal{C}$. Množinu $\mathcal{C}^\diamond$ všech relativních vnitřních bodů množiny $\mathcal{C}$ nazveme jejím relativním vnitřkem.*

Konvexní množiny dimenzí nižších než n mají podobné topologické vlastnosti jako konvexní množiny dimenze n, nahradíme-li klasický vnitřek relativním vnitřkem. Má-li konvexní množina $\mathcal{C} \subset R^n$ dimenzi $m < n$ a $z \in \mathcal{C}$, leží konvexní množina $\mathcal{C} - z$ v lineárním podprostoru $\mathcal{L}_m \subset R^n$ dimenze m. Nechť $a_1, \ldots, a_m$ je ortonormální báze v $\mathcal{L}_m$ a $A = [a_1, \ldots, a_m]$. Pak existuje spojité vzájemně jednoznačné zobrazení $x = Av + z$ mezi body $x \in \mathcal{L}_m + z$ a $v \in R^m$. Toto zobrazení převede $\mathcal{C} \subset \mathcal{L}_m + z$ na $\mathcal{C}_m \subset R^m$ a také relativní vnitřek $\mathcal{C}^\diamond$ na klasický vnitřek $\mathcal{C}_m^\circ$. Z těchto úvah plyne bezprostřední důsledek lemmatu 121.

Důsledek 33. *Nechť $x \in \mathcal{C}^\diamond$, $\overline{x} \in \overline{\mathcal{C}}$ a $y = \lambda\overline{x} + (1-\lambda)x$, kde $0 \le \lambda < 1$. Pak $y \in \mathcal{C}^\diamond$.*

Nyní dokážeme důležitou větu týkající se průniku konvexních množin $\mathcal{C}_1, \ldots, \mathcal{C}_m$, které mohou mít libovolné dimenze.

Věta 286. *Nechť $\mathcal{C}_1, \ldots, \mathcal{C}_m$ jsou konvexní množiny v R^n takové, že $\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond \neq \emptyset$. Pak platí*

$$\overline{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m} = \overline{\mathcal{C}}_1 \cap \cdots \cap \overline{\mathcal{C}}_m.$$

Důkaz Inkluze $\overline{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m} \subset \overline{\mathcal{C}}_1 \cap \cdots \cap \overline{\mathcal{C}}_m$ je zřejmá. Nechť

$$\overline{x} \in \overline{\mathcal{C}}_1 \cap \cdots \cap \overline{\mathcal{C}}_m, \qquad x \in \mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond$$

a $x_i = \lambda_i\overline{x} + (1-\lambda_i)x$, kde $0 \le \lambda_i < 1$. Podle důsledku 33 je bod x_i relativním vnitřním bodem množin $\mathcal{C}_1, \ldots, \mathcal{C}_m$ a leží tedy v průniku $\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond$. Pokud $\lambda_i \to 1$, platí $x_i \to \overline{x}$, takže $\overline{x} \in \overline{\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond} \subset \overline{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_2}$. □

Poznámka 400. Požadavek $\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond \neq \emptyset$, použitý ve větě 286, je podstatný. Nechť

$$\begin{aligned}\mathcal{C}_1 &= \{\lambda s : \lambda \ge 0,\ s = [s_1, s_2] \in R^2,\ s_2 > 0\} = \{x = [x_1, x_2] \in R^2,\ x_2 > 0\} \cup [0,0] \\ \mathcal{C}_2 &= \{\lambda s : \lambda \ge 0,\ s = [s_1, s_2] \in R^2,\ s_2 < 0\} = \{x = [x_1, x_2] \in R^2,\ x_2 < 0\} \cup [0,0]\end{aligned}$$

($\mathcal{C}_1$ je otevřená horní polorovina s přidaným bodem $[0,0]$ a $\mathcal{C}_2$ je otevřená dolní polorovina s přidaným bodem $[0,0]$). Pak $\mathcal{C}_1 \cap \mathcal{C}_2 = \{[0,0]\}$ a $\mathcal{C}_1^\diamond \cap \mathcal{C}_2^\diamond = \emptyset$. Jelikož $\overline{\mathcal{C}}_1$ je uzavřená horní polorovina a $\overline{\mathcal{C}}_2$ je uzavřená dolní polorovina, platí $\overline{\mathcal{C}_1 \cap \mathcal{C}_2} = \{[0,0]\}$ a $\overline{\mathcal{C}}_1 \cap \overline{\mathcal{C}}_2 = \{[x_1, 0] : x_1 \in R\}$ (takže $\overline{\mathcal{C}_1 \cap \mathcal{C}_2}$ je bod a $\overline{\mathcal{C}}_1 \cap \overline{\mathcal{C}}_2$ je přímka).

V dalších úvahách se zaměříme především na uzavřené konvexní množiny. Uzavřená a omezená množina v R^n je kompaktní, což znamená, že z každé posloupnosti jejích bodů lze vybrat konvergentní podposloupnost.

Věta 287. *Je-li množina $\mathcal{C}$ kompaktní, je i množina* $\operatorname{conv}\mathcal{C}$ *kompaktní.*

Důkaz (a) Nechť $y \in \operatorname{conv}\mathcal{C}$. Pak podle věty 284 existují vektory $x_i \in \mathcal{C}$ a čísla $\lambda_i \ge 0$, $1 \le i \le n+1$, $\lambda_1 + \cdots + \lambda_{n+1} = 1$ tak, že $y = \lambda_1 x_1 + \cdots + \lambda_{n+1} x_{n+1}$. Jelikož množina $\mathcal{C}$ je omezená, existuje číslo $M > 0$ takové, že $\|x_i\| \le M$, $1 \le i \le n+1$. Pak ale $\|y\| = \|\lambda_1 x_1 + \cdots + \lambda_{n+1} x_{n+1}\| \le (\lambda_1 + \cdots + \lambda_{n+1})M = M$, takže množina $\operatorname{conv}\mathcal{C}$ je omezená

(b) Nechť $\{y_i\} \subset \operatorname{conv}\mathcal{C}$ je posloupnost taková, že $y_i \to y \in R^n$. Máme dokázat, že $y \in \operatorname{conv}\mathcal{C}$. Jelikož $y_i \in \operatorname{conv}\mathcal{C}$, existují podle věty 284 vektory $x_i^k \in \mathcal{C}$ a čísla $\lambda_i^k \ge 0$, $1 \le k \le n+1$, $\lambda_i^1 + \cdots + \lambda_i^{n+1} = 1$ takové, že $y_i = \lambda_i^1 x_i^1 + \cdots + \lambda_i^{n+1} x_i^{n+1}$. Protože množina $\mathcal{C}$ je kompaktní a číslo $n+1$ je konečné, lze vybrat podposloupnost $\{\tilde{y}_i\} \subset \{y_i\}$ takovou, že odpovídající podposloupnosti $\{\tilde{x}_i^k\} \subset \{x_i^k\}$, $\{\tilde{\lambda}_i^k\} \subset \{\lambda_i^k\}$, $1 \le k \le n+1$, jsou konvergentní, čili $\tilde{x}_i^k \to \tilde{x}^k \in \mathcal{C}$, $\tilde{\lambda}_i^k \to \tilde{\lambda}^k \ge 0$, $1 \le k \le n+1$, $\tilde{\lambda}^1 + \cdots + \tilde{\lambda}^{n+1} = 1$. Jelikož vybraná posloupnost má stejnou limitu jako původní konvergentní posloupnost, lze psát

$$\begin{aligned} y &= \lim_{i\to\infty} y_i = \lim_{i\to\infty} \tilde{y}_i = \lim_{i\to\infty} \left(\sum_{k=1}^{n+1} \tilde{\lambda}_i^k \tilde{x}_i^k \right) \\ &= \sum_{k=1}^{n+1} \left(\lim_{i\to\infty} \tilde{\lambda}_i^k \right) \left(\lim_{i\to\infty} \tilde{x}_i^k \right) = \sum_{k=1}^{n+1} \tilde{\lambda}^k \tilde{x}^k \in \operatorname{conv}\mathcal{C}. \end{aligned}$$

□

Poznámka 401. Předpoklad omezenosti je ve větě 287 podstatný. Je-li $\mathcal{C}$ uzavřená ale neomezená, nemusí být $\operatorname{conv}\mathcal{C}$ uzavřená. Nechť $\mathcal{C} \subset R^2$ a $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$, kde $\mathcal{C}_1$ je úsečka spojující body $x_1 = [-1, 0]$, $x_2 = [1, 0]$

a $\mathcal{C}_2$ je polopřímka $[0,t]$, $t \geq 0$. Nechť $y_i = [1/i - 1, 1]$ a $z_i = [0, i]$. Protože $x_1 \in \mathcal{C}$ a $z_i \in \mathcal{C}$, podle (1111) platí

$$y_i = [1/i - 1, 1] = [-1, 0] + \frac{1}{i}\left([0, i] - [-1, 0]\right) = x_1 + \frac{1}{i}(z_i - x_1) \in \operatorname{conv}\mathcal{C}.$$

Ale $y_i \to y = [-1, 1]$, pokud $i \to \infty$, a bod y nelze vyjádřit jako lineární kombinaci bodů z $\mathcal{C}$, takže $\operatorname{conv}\mathcal{C}$ není uzavřená.

Definice 98. *Nechť $\mathcal{C} \subset R^n$. Pak funkci*

$$d_{\mathcal{C}}(x) = \inf_{y \in \mathcal{C}} \|y - x\|$$

nazveme vzdáleností bodu x od množiny $\mathcal{C}$ (nebo vzdálenostní funkcí množiny $\mathcal{C}$).

Poznámka 402. Je-li množina $\mathcal{C} \subset R^n$ uzavřená, platí

$$d_{\mathcal{C}}(x) = \min_{y \in \mathcal{C}} \|y - x\|.$$

Existuje tedy bod $y \in \mathcal{C}$ takový, že $d_{\mathcal{C}}(x) = \|y - x\|$. V dalším výkladu se omezíme na uzavřené množiny i když většina tvrzení má obecnější charakter.

Věta 288. *Nechť množina $\mathcal{C} \subset R^n$ je uzavřená. Pak vzdálenostní funkce $d_{\mathcal{C}}$ je lipschitzovská v R^n s koeficientem $L = 1$. Je-li $\mathcal{C}$ konvexní, je $d_{\mathcal{C}}$ konvexní v R^n a ke každému bodu $x \in R^n$ existuje právě jeden bod $y \in \mathcal{C}$ takový, že*

$$\|y - x\| = d_{\mathcal{C}}(x).$$

Důkaz Nechť $x_1 \in R^n$, $x_2 \in R^n$. Jelikož množina $\mathcal{C}$ je uzavřená, existuje podle poznámky 402 bod $y \in \mathcal{C}$ takový, že

$$\|y - x_1\| = d_{\mathcal{C}}(x_1).$$

Platí tedy

$$d_{\mathcal{C}}(x_2) \leq \|y - x_2\| \leq \|y - x_1\| + \|x_1 - x_2\| = d_{\mathcal{C}}(x_1) + \|x_2 - x_1\|,$$

neboli

$$d_{\mathcal{C}}(x_2) - d_{\mathcal{C}}(x_1) \leq \|x_2 - x_1\|.$$

Protože nezáleží na pořadí bodů x_1, x_2, platí

$$|d_{\mathcal{C}}(x_2) - d_{\mathcal{C}}(x_1)| \leq \|x_2 - x_1\|,$$

takže funkce $d_{\mathcal{C}}$ je lipschitzovská v R^n s koeficientem $L = 1$ Nechť $\mathcal{C}$ je konvexní a $x_1 \in R^n$, $x_2 \in R^n$. Podle poznámky 402 existují body $y_1 \in \mathcal{C}$, $y_2 \in \mathcal{C}$ takové, že

$$\begin{aligned} \|y_1 - x_1\| &= d_{\mathcal{C}}(x_1), \\ \|y_2 - x_2\| &= d_{\mathcal{C}}(x_2). \end{aligned}$$

Položme $y = \lambda_1 y_1 + \lambda_2 y_2$, kde $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ a $\lambda_1 + \lambda_2 = 1$. Zřejmě $y \in \mathcal{C}$, takže platí

$$\begin{aligned} d_{\mathcal{C}}(\lambda_1 x_1 + \lambda_2 x_2) &\leq \|y - \lambda_1 x_1 - \lambda_2 x_2\| \leq \lambda_1 \|y_1 - x_1\| + \lambda_2 \|y_2 - x_2\| \\ &= \lambda_1 d_{\mathcal{C}}(x_1) + \lambda_2 d_{\mathcal{C}}(x_2) \end{aligned}$$

a $d_{\mathcal{C}}$ je konvexní v R^n Nechť $\mathcal{C}$ je konvexní, $x \in R^n$ a $y_1 \in \mathcal{C}$, $y_2 \in \mathcal{C}$ jsou dva různé body takové, že $\|y_1 - x\| = d_{\mathcal{C}}(x)$, $\|y_2 - x\| = d_{\mathcal{C}}(x)$. Pak

$$\|y_2 - y_1\|^2 = \|(y_2 - x) - (y_1 - x)\|^2 = \|y_2 - x\|^2 + \|y_1 - x\|^2 - 2(y_2 - x)^T(y_1 - x) > 0$$

takže

$$(y_2 - x)^T(y_1 - x) < d_{\mathcal{C}}^2(x). \tag{1114}$$

Položme nyní $y = \frac{1}{2}(y_2 + y_1)$. Jelikož $\mathcal{C}$ je konvexní, platí $y \in \mathcal{C}$. Dále podle (1114) platí

$$\|y - x\|^2 = \frac{1}{4}\|(y_2 - x) + (y_1 - x)\|^2 = \frac{1}{4}\left(\|y_2 - x\|^2 + \|y_1 - x\|^2 + 2(y_2 - x)^T(y_1 - x)\right) < d_{\mathcal{C}}^2(x),$$

což je spor, neboť $y \in \mathcal{C}$, takže podle poznámky 402 $d_{\mathcal{C}}(x) \leq \|y - x\|$. □

Definice 99. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina, $x \in R^n$ a $y \in \mathcal{C}$ je bod takový, že $\|y - x\| = d_{\mathcal{C}}(x)$. Pak řekneme, že y je projekcí bodu x do množiny C a píšeme $y = P_{\mathcal{C}}(x)$.*

Věta 289. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \notin \mathcal{C}$. Pak bod $y = P_{\mathcal{C}}(x)$ je hraničním bodem množiny $\mathcal{C}$.*

Důkaz Připomeňme, že bod $y \in R^n$ je hraničním bodem uzavřené množiny $\mathcal{C} \subset R^n$, jestliže $y \in \mathcal{C}$ a existuje posloupnost $\{x_i\} \subset R^n \setminus \mathcal{C}$ taková, že $x_i \to y$. Nechť $x_i = x + t_i(y - x)$, $0 < t_i < 1$, $i \in N$. Pak platí $\|x_i - x\| < \|y - x\|$, $i \in N$, a pokud $t_i \to 1$, máme posloupnost bodů $x_i \notin \mathcal{C}$ takovou, že $x_i \to y$. □

Lemma 122. . *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina, $x \in R^n$ a $y = P_{\mathcal{C}}(x)$. Pak platí*

$$(x - y)^T(z - y) \leq 0 \quad \forall z \in \mathcal{C}$$

Důkaz Jelikož $y \in \mathcal{C}$, $z \in \mathcal{C}$ a $\mathcal{C}$ je konvexní, platí $y + \lambda(z - y) = \lambda z + (1 - \lambda)y \in \mathcal{C}$ $\forall 0 \leq \lambda \leq 1$. Označme

$$\varphi(\lambda) = \|y + \lambda(z - y) - x\|^2 = \|y - x\|^2 - 2\lambda(x - y)^T(z - y) + \lambda^2\|z - y\|^2.$$

Pak zřejmě $\varphi(0) = d_{\mathcal{C}}(x)$ a $\varphi'(0) = -2(x - y)^T(z - y)$. Pokud by platilo $(x - y)^T(z - y) > 0$, neboli $\varphi'(0) < 0$, existovala by hodnota $0 < \lambda \leq 1$ taková, že $\varphi(\lambda) < \varphi(0)$, neboli $\|y + \lambda(z - y) - x\|^2 < d_{\mathcal{C}}(x)$, což není možné, neboť $y + \lambda(z - y) \in \mathcal{C}$ $\forall 0 \leq \lambda \leq 1$. □

Věta 290. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina. Pak*

$$\|P_{\mathcal{C}}(x_2) - P_{\mathcal{C}}(x_1)\| \leq \|x_2 - x_1\| \quad \forall x_1, x_2 \in R^n.$$

Důkaz Nechť $y_1 = P_{\mathcal{C}}(x_1)$ a $y_2 = P_{\mathcal{C}}(x_2)$. Podle lemmatu 122 platí

$$\begin{aligned}(x_1 - y_1)^T(z_1 - y_1) &\leq 0 \quad \forall z_1 \in \mathcal{C},\\ (x_2 - y_2)^T(z_2 - y_2) &\leq 0 \quad \forall z_2 \in \mathcal{C}.\end{aligned}$$

Dosadíme-li $z_1 = y_2$, $z_2 = y_1$ a sečteme-li obě nerovnosti, dostaneme

$$((y_2 - y_1) - (x_2 - x_1))^T(y_2 - y_1) \leq 0,$$

neboli

$$\|y_2 - y_1\|^2 \leq (x_2 - x_1)^T(y_2 - y_1) \leq \|x_2 - x_1\|\|y_2 - y_1\|,$$

což dává $\|y_2 - y_1\| \leq \|x_2 - x_1\|$. □

Definice 100. *Nechť $a \in R^n$ a $\alpha \in R$. Pak množinu*

$$\mathcal{H}(a, \alpha) = \{y \in R^n : a^T y \leq \alpha\}$$

nazveme poloprostorem určeným normálovým vektorem a a číslem α.

Poznámka 403. Hranicí poloprostoru $\mathcal{H}(a, \alpha)$ je nadrovina

$$\mathcal{L}(a, \alpha) = \mathcal{H}(a, \alpha) \cap \mathcal{H}(-a, -\alpha) = \{y \in R^n : a^T y = \alpha\}.$$

˘Číslo α určuje vzdálenost nadroviny $\mathcal{L}(a, \alpha)$ od počátku. Tato vzdálenost se rovná podílu $|\alpha|/\|a\|$. Odtud plyne, že bod $y = 0$ je hraničním bodem poloprostoru $\mathcal{H}(a, \alpha)$ (leží v hraniční nadrovině $\mathcal{L}(a, \alpha)$) právě tehdy, když $\alpha = 0$.

Věta 291. *Poloprostor $\mathcal{H}(a,\alpha)$ je uzavřenou konvexní množinou.*

Důkaz (a) Nechť $\{y_i\} \subset \mathcal{H}(a,\alpha)$ je posloupnost taková, že $y_i \to y$. Jelikož $a^T y_i \leq \alpha \ \forall i \in N$ a neostrá nerovnost je invariantní vůči limitnímu přechodu, platí $a^T y \leq \alpha$, takže $y \in \mathcal{H}(a,\alpha)$. Poloprostor $\mathcal{H}(a,\alpha)$ je tedy uzavřený.

(b) Nechť $y_1 \in \mathcal{H}(a,\alpha)$, $y_2 \in \mathcal{H}(a,\alpha)$, takže $a^T y_1 \leq \alpha$, $a^T y_2 \leq \alpha$, a nechť $y = \lambda y_1 + (1-\lambda) y_2$, kde $0 \leq \lambda \leq 1$. Pak platí

$$a^T y = a^T(\lambda y_1 + (1-\lambda)y_2) = \lambda a^T y_1 + (1-\lambda) a^T y_2 \leq \lambda\alpha + (1-\lambda)\alpha = \alpha,$$

takže $y \in \mathcal{H}(a,\alpha)$. Poloprostor $\mathcal{H}(a,\alpha)$ je tedy konvexní. □

Věta 292. *Nechť $\mathcal{C}$ je uzavřená konvexní množina a nechť $x \notin \mathcal{C}$. Pak existuje poloprostor $\mathcal{H}(a,\alpha)$ takový, že $\mathcal{C} \subset \mathcal{H}(a,\alpha)$ a $x \notin \mathcal{H}(a,\alpha)$. Tento poloprostor lze volit tak, že $a = x - P_{\mathcal{C}}(x)$ a $\alpha = a^T P_{\mathcal{C}}(x)$. Pak $P_{\mathcal{C}}(x) \in \mathcal{L}(a,\alpha)$, takže $\mathcal{C} \cap \mathcal{L}(a,\alpha) \neq \emptyset$.*

Důkaz Máme dokázat, že existuje vektor $a \in R^n$ a číslo $\alpha \in R$ tak, že

$$a^T x > \alpha \geq a^T z \quad \forall z \in \mathcal{C}.$$

Nechť $y = P_{\mathcal{C}}(x)$, takže $\|y - x\| = d_{\mathcal{C}}(x)$. Položme $a = x - y$ a $\alpha = a^T y$. Pak platí

$$a^T x = (x-y)^T x = (x-y)^T(x-y) + (x-y)^T y = \|x-y\|^2 + a^T y > \alpha,$$

neboť $x \neq \mathcal{C}$, takže $\|x - y\| \neq 0$. Použijeme-li lemma 122, dostaneme

$$a^T z - \alpha = (x-y)^T z - (x-y)^T y = (x-y)^T(z-y) \leq 0 \quad \forall z \in \mathcal{C}.$$

□

Důsledek 34. *Nechť $\mathcal{C}_1$, $\mathcal{C}_2$ jsou uzavřené konvexní množiny takové, že $\mathcal{C}_1 \cap \mathcal{C}_2 = \emptyset$. Pak existuje poloprostor $\mathcal{H}(a,\alpha)$ takový, že $\mathcal{C}_1 \subset \mathcal{H}(a,\alpha)$ a $\mathcal{C}_2 \cap \mathcal{H}(a,\alpha) = \emptyset$*

Důkaz Jelikož množiny $\mathcal{C}_1$, $\mathcal{C}_2$ jsou uzavřené, existují body $x_1 \in \mathcal{C}_1$, $x_2 \in \mathcal{C}_2$ takové, že

$$d(\mathcal{C}_1, \mathcal{C}_2) \triangleq \inf_{y_1 \in \mathcal{C}_1, y_2 \in \mathcal{C}_2} \|y_2 - y_1\| = \min_{y_1 \in \mathcal{C}_1, y_2 \in \mathcal{C}_2} \|y_2 - y_1\| = \|x_2 - x_1\|$$

(argumentace je stejná jako v poznámce 402, dvojice $x_1 \in \mathcal{C}_1$, $x_2 \in \mathcal{C}_2$ nemusí být určena jednoznačně). Protože $x_2 \notin \mathcal{C}_1$ plyne z věty 292 (a jejího důkazu), že $\mathcal{C}_1 \subset \mathcal{H}(a_1,\alpha_1)$ a $x_2 \notin \mathcal{H}(a_1,\alpha_1)$, kde $a_1 = x_2 - x_1$ a $\alpha_1 = a_1^T x_1$. Podobně $\mathcal{C}_2 \subset \mathcal{H}(a_2,\alpha_2)$ a $x_1 \notin \mathcal{H}(a_2,\alpha_2)$, kde $a_2 = x_1 - x_2$ a $\alpha_2 = a_2^T x_2$. Zbývá dokázat, že $\mathcal{H}(a_1,\alpha_1) \cap \mathcal{H}(a_2,\alpha_2) = \emptyset$ (pak lze volit $a = a_1$, $\alpha = \alpha_1$). Nechť $y \in \mathcal{H}(a_1,\alpha_1) \cap \mathcal{H}(a_2,\alpha_2)$. Pak platí

$$(x_2 - x_1)^T y = a_1^T y \leq \alpha_1 = (x_2 - x_1)^T x_1,$$
$$(x_1 - x_2)^T y = a_2^T y \leq \alpha_2 = (x_1 - x_2)^T x_2,$$

jejichž sečtením dostaneme

$$0 \leq (x_2 - x_1)^T (x_1 - x_2) = -\|x_2 - x_1\|^2 < 0$$

(neboť $x_2 \neq x_1$), což je spor. □

Věta 293. *Uzavřená konvexní množina $\mathcal{C} \subset R^n$ je průnikem všech poloprostorů obsahujících $\mathcal{C}$.*

Důkaz Nechť $\tilde{\mathcal{C}}$ je průnikem všech poloprostorů obsahujících uzavřenou konvexní množinu $\mathcal{C}$. Jelikož kazdý poloprostor je podle věty 291 uzavřený a konvexní, je množina $\tilde{\mathcal{C}}$ uzavřená a konvexní a platí. $\mathcal{C} \subset \tilde{\mathcal{C}}$. Stačí tedy dokázat, že $\tilde{\mathcal{C}} \subset \mathcal{C}$. Předpokládejme naopak, že existuje bod $x \in \tilde{\mathcal{C}}$ takový, že $x \notin \mathcal{C}$. Pak podle věty 292 existuje poloprostor $\mathcal{H}$ takový, že $\mathcal{C} \subset \mathcal{H}$ a $x \notin \mathcal{H}$. Jelikož $\mathcal{C} \subset \mathcal{H}$, platí $\mathcal{C} \subset \tilde{\mathcal{C}} \subset \mathcal{H}$, což je spor, neboť $x \in \tilde{\mathcal{C}}$ a $x \notin \mathcal{H}$. □

Definice 101. *Konvexní množina, která je průnikem konečného počtu poloprostorů, se nazývá polyedrální množinou.*

Definice 102. *Nechť $\mathcal{C}$ je uzavřená konvexní množina a $\mathcal{H}(a,\alpha)$ je poloprostor s hranicí $\mathcal{L}(a,\alpha)$ takový, že $\mathcal{C} \subset \mathcal{H}(a,\alpha)$ a $\mathcal{C} \cap \mathcal{L}(a,\alpha) \neq \emptyset$. Pak řekneme, že $\mathcal{H}(a,\alpha)$ je tečným poloprostorem a $\mathcal{L}(a,\alpha)$ tečnou nadrovinou množiny $\mathcal{C}$.*

Poznámka 404. Ve větě 293 se můžeme omezit na tečné poloprostory (uzavřená konvexní množina je průnikem svých tečných poloprostorů). Obsahuje-li poloprostor $\mathcal{H}(a,\alpha)$ konvexní množinu $\mathcal{C}$, přižemž $\mathcal{C} \cap \mathcal{L}(a,\alpha) = \emptyset$, lze volbou $\alpha' = \max_{y\in\mathcal{C}} a^T y$ docílit toho, že $\mathcal{C} \subset \mathcal{H}(a,\alpha') \subset \mathcal{H}(a,\alpha)$ a $\mathcal{C} \cap \mathcal{L}(a,\alpha') \neq \emptyset$.

Věta 294. *Nechť bod $y \in R^n$ je hraničním bodem uzavřené konvexní množiny $\mathcal{C}$. Pak existuje tečná nadrovina $\mathcal{L}(a,\alpha)$ taková, že $y \in \mathcal{L}(a,\alpha)$.*

Důkaz Jelikož $y \in \mathcal{C}$ je hraničním bodem uzavřené konvexní množiny $\mathcal{C}$, existuje posloupnost $\{x_i\} \subset R^n \setminus \mathcal{C}$ taková, že $x_i \to y$. Pro každý bod $x_i \notin \mathcal{C}$, $i \in N$, lze podle věty 292 sestrojit poloprostor $\mathcal{H}(a_i,\alpha_i)$ takový, že $\mathcal{C} \subset \mathcal{H}(a_i,\alpha_i)$ a $P_{\mathcal{C}}(x_i) \in \mathcal{L}(a_i,\alpha_i)$, přičemž $a_i = x_i - P_{\mathcal{C}}(x_i)$ a $\alpha_i = a_i^T P_{\mathcal{C}}(x_i)$. Jelikož $x_i \to y$, platí podle věty 290 $P_{\mathcal{C}}(x_i) \to P_{\mathcal{C}}(y)$, takže vektory a_i a čísla α_i jsou omezené a můžeme tudíž bez újmy na obecnosti předpokládat, že $a_i \to a$ a $\alpha_i \to \alpha$ (v opačném případě vybereme vhodné podposloupnosti). Jelikož se rovnost i neostrá nerovnost zachovávají při limitním přechodu, platí $\mathcal{C} \subset \mathcal{H}(a,\alpha)$ a $y \in \mathcal{L}(a,\alpha)$. □

Definice 103. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Není-li bod x konvexní kombinací žádných bodů z $\mathcal{C}$ různých od x, řekneme, že x je krajním bodem nebo vrcholem množiny $\mathcal{C}$.*

Poznámka 405. Z důkazu věty 281 plyne, že se v definici krajních bodů můžeme omezit na konvexní kombinace dvou bodů z $\mathcal{C}$ různých od x. Dále se můžeme omezit na průměry dvou bodů z $\mathcal{C}$ různých od x. Nechť $x = \lambda_1 x_1 + \lambda_2 x_2$, $\lambda_1 + \lambda_2 = 1$, $\lambda_1 \geq \lambda_2 \geq 0$. Položíme-li $x_3 = \lambda_1' x_1 + \lambda_2' x_2$, kde $\lambda_1' = 2\lambda_1 - 1$, $\lambda_2' = 2\lambda_2$, takže $\lambda_1' + \lambda_2' = 1$, $\lambda_1' \geq 0$, $\lambda_2' \geq 0$, platí $x_3 \in \mathcal{C}$ a $x = (x_1 + x_3)/2$. Bod $x \in \mathcal{C}$ je tedy krajním bodem konvexní množiny $\mathcal{C}$, neexistují-li dva body $x_1 \in \mathcal{C}$, $x_3 \in \mathcal{C}$ takové, že $x = (x_1 + x_3)/2$.

Věta 295. *Kompaktní konvexní množina je konvexním obalem svých krajních bodů.*

Důkaz větu dokážeme indukcí. V R je tvrzení zřejmé, neboť v tomto případě je každá kompaktní konvexní množina uzavřeným intervalem, který je konvexním obalem svých krajních bodů. Předpokládejme, že tvrzení platí v R^k, kde k probíhá indexy $1 \leq k \leq n-1$. Nechť $\mathcal{C} \subset R^n$ je kompaktní konvexní množina a $x \in \mathcal{C}$ není jejím krajním bodem.

(a) Předpokládejme nejprve, že x je hraničním bodem množiny $\mathcal{C}$. Pak podle věty 294 existuje tečná nadrovina $\mathcal{L}(a,\alpha)$ taková, že $x \in \mathcal{L}(a,\alpha)$. Označme $\tilde{\mathcal{C}} = \mathcal{C} \cap \mathcal{L}(a,\alpha)$. Jelikož $\mathcal{C}$ a $\mathcal{L}(a,\alpha)$ jsou uzavřené konvexní množiny, $\mathcal{C}$ je kompaktní a $\mathcal{L}(a,\alpha)$ má dimenzi nižší než n, je i množina $\tilde{\mathcal{C}}$ kompaktní, konvexní a má dimenzi nižší než n. Podle indukčního předpokladu je tedy bod x konvexní kombinací krajních bodů množiny $\tilde{\mathcal{C}}$. Zbývá dokázat, že krajní body množiny $\tilde{\mathcal{C}}$ jsou také krajní body množiny $\mathcal{C}$. Předpokládejme naopak, že bod y je krajním bodem množiny $\tilde{\mathcal{C}}$, ale není krajním bodem množiny $\mathcal{C}$. Pak podle poznámky 405 existují body $y_1 \in \mathcal{C} \setminus \mathcal{L}(a,\alpha)$, $y_2 \in \mathcal{C} \setminus \mathcal{L}(a,\alpha)$, takové, že $y = (y_1 + y_2)/2$. Jelikož $y \in \mathcal{L}(a,\alpha)$, platí $\alpha = a^T y = (a^T y_1 + a^T y_2)/2$ a pokud $a^T y_1 < \alpha$, musí být $a^T y_2 > \alpha$, což je spor, neboť $y_2 \in \mathcal{C}$ a $\mathcal{C} \subset \mathcal{H}(a,\alpha)$, takže nutně $a^T y_2 \leq \alpha$.

(b) Je-li x vnitřním bodem množiny $\mathcal{C}$, která je kompaktní, lze tímto bodem vést přímku, která protne hranici množiny $\mathcal{C}$ ve dvou různých bodech $x_1 \neq x$ a $x_2 \neq x$. Zřejmě x je konvexní kombinací bodů x_1 a x_2. Jelikož v (a) bylo dokázáno, že body x_1 a x_2 jsou konvexními kombinacemi krajních bodů množiny $\mathcal{C}$, je i bod x konvexní kombinací krajních bodů množiny $\mathcal{C}$. □

Definice 104. *Nechť $\mathcal{C} \subset R^n$. Pak funkci*

$$\delta_{\mathcal{C}}(x) = \sup_{y\in\mathcal{C}} y^T x$$

nazveme opěrnou funkcí množiny $\mathcal{C}$.

Poznámka 406. Nechť množina $\mathcal{C} \subset R^n$ je kompaktní. Pak platí

$$\delta_{\mathcal{C}}(x) = \max_{y \in \mathcal{C}} y^T x.$$

Existuje tedy bod $y \in \mathcal{C}$ takový, že $\delta_{\mathcal{C}}(x) = y^T x$. V dalším výkladu se omezíme na kompaktní množiny i když většina tvrzení má obecnější charakter.

Věta 296. *Nechť množina $\mathcal{C} \subset R^n$ je kompaktní. Pak opěrná funkce $\delta_{\mathcal{C}}$ je pozitivně homogenní, subaditivní a lipschitzovská v R_n.*

Důkaz Podle poznámky 406 pro $x \in R^n$ a $\lambda \geq 0$ platí

$$\delta_{\mathcal{C}}(\lambda x) = \max_{y \in \mathcal{C}} y^T(\lambda x) = \lambda \max_{y \in \mathcal{C}} y^T x = \lambda \delta_{\mathcal{C}}(x),$$

takže funkce $\delta_{\mathcal{C}}$ je pozitivně homogenní. Podobně pro $x_1 \in R^n$ a $x_2 \in R^n$ platí

$$\delta_{\mathcal{C}}(x_1 + x_2) = \max_{y \in \mathcal{C}} y^T(x_1 + x_2) \leq \max_{y \in \mathcal{C}} y^T x_1 + \max_{y \in \mathcal{C}} y^T x_2 = \delta_{\mathcal{C}}(x_1) + \delta_{\mathcal{C}}(x_2),$$

takže funkce $\delta_{\mathcal{C}}$ je subaditivní. Ze subaditivity plyne nerovnost

$$\delta_{\mathcal{C}}(x_2) = \delta_{\mathcal{C}}(x_1 + (x_2 - x_1)) \leq \delta_{\mathcal{C}}(x_1) + \delta_{\mathcal{C}}(x_2 - x_1)$$

a jelikož $\mathcal{C}$ je kompaktní existuje konstanta L taková, že $\|y\| \leq L \ \forall y \in \mathcal{C}$. Můžeme tedy psát

$$\delta_{\mathcal{C}}(x_2) - \delta_{\mathcal{C}}(x_1) \leq \max_{y \in \mathcal{C}} y^T(x_2 - x_1) \leq L\|x_2 - x_1\|.$$

Protože nezáleží na pořadí bodů x_1, x_2, platí

$$|\delta_{\mathcal{C}}(x_2) - \delta_{\mathcal{C}}(x_1)| \leq L\|x_2 - x_1\|,$$

takže funkce $\delta_{\mathcal{C}}$ je lipschitzovská v R^n. □

Věta 297. *Nechť množina $\mathcal{C} \subset R^n$ je kompaktní. Pak*

$$\delta_{\mathcal{C}}(x) = \delta_{\operatorname{conv}\mathcal{C}}(x) \quad \forall x \in R^n.$$

Důkaz Protože $\mathcal{C} \subset \operatorname{conv}\mathcal{C}$, platí podle poznámky 406 $\delta_{\mathcal{C}}(x) \leq \delta_{\operatorname{conv}\mathcal{C}}(x) \ \forall x \in R^n$. Nechť $x \in R^n$. Podle věty 284 lze každý vektor $y \in \operatorname{conv}\mathcal{C}$ vyjádřit jako konvexní kombinaci nejvýše $n+1$ vektorů $y_i \in \mathcal{C}$, $1 \leq i \leq n+1$. Můžeme tedy psát

$$\begin{aligned} \delta_{\operatorname{conv}\mathcal{C}}(x) &= \max_{y \in \operatorname{conv}\mathcal{C}} y^T x = \max\left\{ \sum_{i=1}^{n+1} \lambda_i y_i^T x : y_i \in \mathcal{C}, \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\} \\ &\leq \max_{y \in \mathcal{C}} y^T x = \delta_{\mathcal{C}}(x). \end{aligned}$$

□

Věta 298. *Nechť množiny $\mathcal{C}_1 \subset R^n$, $\mathcal{C}_2 \subset R^n$ jsou konvexní a kompaktní. Pak $\mathcal{C}_1 \subset \mathcal{C}_2$ platí právě tehdy, když*

$$\delta_{\mathcal{C}_1}(x) \leq \delta_{\mathcal{C}_2}(x) \quad \forall x \in R^n.$$

Důkaz Jestliže $\mathcal{C}_1 \subset \mathcal{C}_2$, pak podle poznámky 406 platí $\delta_{\mathcal{C}_1}(x) \leq \delta_{\mathcal{C}_2}(x) \ \forall x \in R^n$. Předpokládejme, že $\delta_{\mathcal{C}_1}(x) \leq \delta_{\mathcal{C}_2}(x) \ \forall x \in R^n$ a existuje bod $\overline{y} \in \mathcal{C}_1$ takový, že $\overline{y} \notin \mathcal{C}_2$. Pak podle věty 292 existuje vektor $a \in R^n$ a číslo $\alpha \in R$ tak, že

$$a^T \overline{y} > \alpha \geq a^T y \quad \forall y \in \mathcal{C}_2.$$

Platí tedy

$$\delta_{\mathcal{C}_1}(a) \geq a^T \overline{y} > \delta_{\mathcal{C}_2}(a),$$

což je ve sporu s předpokladem. □

Poznámka 407. Ve větě 298 se můžeme omezit na vektory jednotkové délky. Inkluze $\mathcal{C}_1 \subset \mathcal{C}_2$ platí právě tehdy, když

$$\delta_{\mathcal{C}_1}(x) \leq \delta_{\mathcal{C}_2}(x) \quad \forall\, (x \in R^n,\ \|x\| = 1).$$

Plyne to z pozitivní homogenity opěrné funkce (věta 296).

Důsledek 35. *Nechť množina $\mathcal{C} \subset R^n$ je konvexní a kompaktní. Pak $y \in \mathcal{C}$ právě tehdy, když*

$$y^T x \leq \delta_{\mathcal{C}}(x) \quad \forall x \in R^n.$$

Věta 299. *Nechť množiny $\mathcal{C}_1 \subset R^n$, $\mathcal{C}_2 \subset R^n$ jsou kompaktní. Pak*

$$\delta_{\mathcal{C}_1+\mathcal{C}_2}(x) = \delta_{\mathcal{C}_1}(x) + \delta_{\mathcal{C}_2}(x).$$

Důkaz Platí

$$\begin{aligned} \delta_{\mathcal{C}_1+\mathcal{C}_2}(x) &= \max_{y \in \mathcal{C}_1+\mathcal{C}_2} y^T x = \max_{\substack{y_1 \in \mathcal{C}_1 \\ y_2 \in \mathcal{C}_2}} (y_1+y_2)^T x = \max_{y_1 \in \mathcal{C}_1} y_1^T x + \max_{y_2 \in \mathcal{C}_2} y_2^T x \\ &= \delta_{\mathcal{C}_1}(x) + \delta_{\mathcal{C}_2}(x). \end{aligned}$$

□

Opěrná funkce množiny $\mathcal{C} \subset R^n$ má bezprostřední vztah k poloprostorům obsahujícím tuto množinu.

Věta 300. *Množina $\mathcal{C} \subset R^n$ leží v poloprostoru $\mathcal{H}(a,\alpha)$ právě tehdy, když $\alpha \geq \delta_{\mathcal{C}}(a)$, přičemž $\mathcal{H}(a,\alpha)$ je tečným poloprostorem množiny $\mathcal{C}$ právě tehdy, když $\alpha = \delta_{\mathcal{C}}(a)$.*

Důkaz Tvrzení plyne z definice 104 a z toho, že $\mathcal{C} \subset \mathcal{H}(a,\alpha)$ právě tehdy, když $\delta_{\mathcal{C}}(a) = \sup_{y \in \mathcal{C}} a^T y \leq \alpha$ a $\mathcal{C} \cap \mathcal{L}(a,\alpha) = \emptyset$, pokud $\delta_{\mathcal{C}}(a) = \sup_{y \in \mathcal{C}} a^T y < \alpha$. □

15.2 Konvexní kužely

Definice 105. *Řekneme, že množina $\mathcal{K} \subset R^n$ je kuželem, jestliže z $x \in \mathcal{K}$ a $\lambda \geq 0$ plyne $\lambda x \in \mathcal{K}$.*

Věta 301. *Průniky a lineární kombinace kuželů jsou kužely.*

Důkaz (a) Nechť $\mathcal{K} = \cap_\alpha \mathcal{K}_\alpha$, kde $\mathcal{K}_\alpha \subset R^n$ jsou kužely. Nechť $x \in \mathcal{K}$ a $\lambda \geq 0$. Pak platí $x \in \mathcal{K}_\alpha$ a tedy $\lambda x \in \mathcal{K}_\alpha$ $\forall \alpha$. Odtud plyne, že $\lambda x \in \mathcal{K}$.

(b) Nechť $\mathcal{K} = \sum_{i=1}^m \lambda_i \mathcal{K}_i$, kde $\mathcal{K}_i \subset R^n$ jsou kužely a $\lambda_i \in R$. Nechť $x \in \mathcal{K}$ a $\lambda \geq 0$. Pak existují body $x_i \in \mathcal{K}_i$, $1 \leq i \leq m$, takové, že

$$\lambda x = \lambda \sum_{i=1}^m \lambda_i x_i = \sum_{i=1}^m \lambda_i (\lambda x_i).$$

Jelikož $x_i \in \mathcal{K}_i$ a $\lambda \geq 0$, platí $\lambda x_i \in \mathcal{K}_i$, $1 \leq i \leq m$, takže $\lambda x \in \mathcal{K}$. □

Definice 106. *Kuželovým obalem množiny $\mathcal{C} \subset R^n$ nazveme průnik*

$$\text{cone}\ \mathcal{C} = \bigcap_{\mathcal{C} \subset \mathcal{K}_\alpha} \mathcal{K}_\alpha$$

všech kuželů $\mathcal{K}_\alpha \subset R^n$ obsahujících $\mathcal{C}$.

Věta 302. *Nechť $\mathcal{C} \subset R^n$. Pak platí*

$$\text{cone}\ \mathcal{C} = \bigcup_{\lambda \geq 0} \lambda \mathcal{C} = \{x \in R^n :\ x = \lambda y,\ \ y \in \mathcal{C},\ \ \lambda \geq 0\}$$

Kužel cone $\mathcal{C}$ *je tedy množinou všech nezáporných násobků bodů z $\mathcal{C}$.*

Důkaz Nechť $\tilde{\mathcal{K}} = \bigcup_{\lambda \geq 0} \lambda \mathcal{C}$. Jelikož $\tilde{\mathcal{K}}$ je kužel obsahující množinu $\mathcal{C}$, platí cone $\mathcal{C} \subset \tilde{\mathcal{K}}$. Nechť naopak $x \in \tilde{\mathcal{K}}$, takže $x = \lambda y$, kde $y \in \mathcal{C}$ a $\lambda \geq 0$. Nechť $\mathcal{K}_\alpha$ je libovolný kužel obsahující množinu $\mathcal{C}$. Jelikož $y \in \mathcal{C}$, platí $y \in \mathcal{K}_\alpha$ a jelikož $\lambda \geq 0$, platí $x = \lambda y \in \mathcal{K}_\alpha$. Tudíž $x \in$ cone $\mathcal{C}$. □

Věta 303. *Množina $\mathcal{K} \subset R^n$ je konvexním kuželem právě tehdy, obsahuje-li všechny nezáporné lineární kombinace svých bodů.*

Důkaz Obsahuje-li množina $\mathcal{K}$ všechny nezáporné lineární kombinace svých bodů, obsahuje též konvexní kombinace tvaru (1110) a nezáporné násobky svých bodů, takže je konvexním kuželem. Nechť $\mathcal{K}$ je konvexním kuželem a $x_i \in \mathcal{K}$, $\lambda_i \geq 0$, $1 \leq i \leq m$. Položme $\lambda = \lambda_1 + \cdots + \lambda_m$. Jestliže $\lambda = 0$, platí $\lambda_1 x_1 + \ldots \lambda_m x_m = 0 \in \mathcal{K}$. Jestliže $\lambda > 0$, položíme

$$x' = \sum_{i=1}^{m} \frac{\lambda_i}{\lambda} x_i \triangleq \sum_{i=1}^{m} \lambda_i' x_i,$$

kde $\lambda_1' + \cdots + \lambda_m' = 1$. Jelikož množina $\mathcal{K}$ je konvexní, platí $x' \in \mathcal{K}$, takže

$$x = \sum_{i=1}^{m} \lambda_i x_i = \lambda x' \in \mathcal{K}.$$

□

Důsledek 36. *Množina $\mathcal{K} \subset R^n$ je konvexním kuželem právě tehdy, obsahuje-li nezáporné násobky a součty svých bodů.*

Důkaz Obsahuje-li množina $\mathcal{K}$ nezáporné násobky a součty svých bodů, je kuželem podle definice 105 a z $x_1 \in \mathcal{K}$, $x_2 \in \mathcal{K}$ a $0 \leq \lambda \leq 1$ plyne $\lambda x_1 \in \mathcal{K}$, $(1-\lambda)x_2 \in \mathcal{K}$, takže $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{K}$ a $\mathcal{K}$ je konvexní množinou. Opačná implikace plyne bezprostředně z věty 303. □

Věta 304. *Konvexní kužel* cone (conv $\mathcal{C}$) = conv (cone $\mathcal{C}$) *je množinou všech nezáporných lineárních kombinací bodů z $\mathcal{C}$.*

Důkaz (a) Nejprve ukážeme, že cone (conv $\mathcal{C}$) = conv (cone $\mathcal{C}$), takže cone (conv $\mathcal{C}$) je konvexním kuželem. Nechť $x \in$ cone (conv $\mathcal{C}$). Pak podle věty 284 platí

$$x = \lambda \sum_{i=1}^{n+1} \lambda_i x_i = \sum_{i=1}^{n+1} \lambda_i (\lambda x_i), \quad \sum_{i=1}^{n+1} \lambda_i = 1,$$

kde $x_i \in \mathcal{C}$, $\lambda_i \geq 0$, $1 \leq i \leq n+1$, takže $x \in$ conv (cone $\mathcal{C}$). Nechť naopak $x \in$ conv (cone $\mathcal{C}$). Pak podle věty 284 platí

$$x = \sum_{i=1}^{n+1} \lambda_i (\mu_i x_i) = \lambda' \sum_{i=1}^{n+1} \lambda_i' x_i, \quad \sum_{i=1}^{n+1} \lambda_i' = \sum_{i=1}^{n+1} \lambda_i = 1,$$

kde $x_i \in \mathcal{C}$, $\lambda_i \geq 0$, $\mu_i \geq 0$, $\lambda_i' = \lambda_i \mu_i / \lambda' \geq 0$, $1 \leq i \leq n+1$, a $\lambda' = \lambda_1 \mu_1 + \cdots + \lambda_{n+1} \mu_{n+1}$, takže $x \in$ cone (conv $\mathcal{C}$).

(b) Nechť $\tilde{\mathcal{K}}$ je množina všech nezáporných lineárních kombinací bodů z $\mathcal{C}$. Jelikož nezáporná lineární kombinace nezáporných lineárních kombinací je opět nezápornou lineární kombinací je podle věty 303 $\tilde{\mathcal{K}}$ konvexním kuželem. Podle definice 106 tedy platí cone (conv $\mathcal{C}$) $\subset \tilde{\mathcal{K}}$. Nechť $x \in \tilde{\mathcal{K}}$ a $x \notin$ cone (conv $\mathcal{C}$). Pak podle věty 292 existuje poloprostor $\mathcal{H}(a, 0)$ takový, že cone (conv $\mathcal{C}$) $\subset \mathcal{H}(a, 0)$ a $x \notin \mathcal{H}(a, 0)$. Jelikož $\mathcal{H}(a, 0)$ je konvexním kuželem, je podle vět 282 a 301 i $\mathcal{H}(a, 0) \cap$ cone (conv $\mathcal{C}$) konvexním kuželem, což je spor s minimalitou $\tilde{\mathcal{K}}$ (definice 106). □

Jelikož uzavřený konvexní kužel je uzavřenou konvexní množinou, můžeme studovat tečné poloprostory uzavřených konvexních kuželů.

Věta 305. *Tečný poloprostor uzavřeného konvexního kuželu je uzavřeným konvexním kuželem (takže obsahuje počátek souřadnic). Jestliže $\mathcal{K} \subset \mathcal{H}(a,0)$, je $\mathcal{H}(a,0)$ tečným poloprostorem uzavřeného konvexního kuželu $\mathcal{K}$.*

Důkaz Nechť $\mathcal{H}(a,\alpha)$ je tečným poloprostorem uzavřeného konvexního kuželu $\mathcal{K}$. Jelikož $\mathcal{L}(a,\alpha) \cap \mathcal{K} \neq \emptyset$, existuje bod $y \in \mathcal{K}$ takový, že $a^T y = \alpha$. Protože $\mathcal{K}$ je kuželem, musí platit $\lambda y \in \mathcal{K} \subset \mathcal{H}(a,\alpha)\ \forall \lambda \geq 0$, neboli

$$\lambda\alpha = a^T(\lambda y) \leq \alpha \quad \forall \lambda \geq 0,$$

což lze zajistit pouze tehdy, když $\alpha = 0$. V tomto případě $0 \in \mathcal{H}(a,0)$ a pokud $x \in \mathcal{H}(a,0)$, pak také $\lambda x \in \mathcal{H}(a,0)\ \forall \lambda \geq 0$. Uzavřenost a konvexita $\mathcal{H}(a,0)$ plynou z věty 291. Zbytek tvrzení plyne z toho, že $0 \in \mathcal{K} \cup \mathcal{L}(a,0)$. □

Definice 107. *Nechť $\mathcal{C} \in R^n$. Množinu*

$$\mathcal{C}^* = \{x \in R^n : y^T x \leq 0 \quad \forall y \in \mathcal{C}\}$$

nazveme polárním kuželem množiny $\mathcal{C}$.

Poznámka 408. Z definice 107 lze snadno usoudit, že z $\mathcal{C}_1 \subset \mathcal{C}_2$ plyne $\mathcal{C}_2^* \subset \mathcal{C}_1^*$.

Věta 306. *Nechť $\mathcal{C} \subset R^n$. Pak množina $\mathcal{C}^*$ je uzavřeným konvexním kuželem.*

Důkaz (a) Nechť $\{x_i\} \subset \mathcal{C}^*$ je posloupnost taková, že $x_i \to x$. Jelikož $y^T x_i \leq 0\ \forall i \in N\ \forall y \in \mathcal{C}$ a neostrá nerovnost je invariantní vůči limitnímu přechodu, platí

$$y^T x = \lim_{i\to\infty} y^T x_i \leq 0 \quad \forall y \in \mathcal{C},$$

takže $x \in \mathcal{C}^*$.
(b) Nechť $x_1 \in \mathcal{C}^*$, $x_2 \in \mathcal{C}^*$. Pak platí $y^T x_1 \leq 0$, $y^T x_2 \leq 0\ \forall y \in \mathcal{C}$. Nechť $0 \leq \lambda \leq 1$ a $x = \lambda x_1 + (1-\lambda)x_2$. Pak

$$y^T x = y^T(\lambda x_1 + (1-\lambda)x_2) = \lambda y^T x_1 + (1-\lambda) y^T x_2 \leq 0 \quad \forall y \in \mathcal{C},$$

takže $x \in \mathcal{C}^*$.
(c) Nechť $x \in \mathcal{C}^*$ a $\lambda \geq 0$. Pak platí

$$y^T(\lambda x) = \lambda y^T x \leq 0 \quad \forall y \in \mathcal{C},$$

takže $\lambda x \in \mathcal{C}^*$. □

Věta 307. *Je-li $\mathcal{K} \subset R^n$ uzavřeným konvexním kuželem, platí $(\mathcal{K}^*)^* = \mathcal{K}$.*

Důkaz Nechť $y \in (\mathcal{K}^*)^*$. Pak podle definice 107 platí $y^T x \leq 0 \quad \forall x \in \mathcal{K}^*$, takže $y \in \mathcal{K}$. Nechť naopak $z \notin \mathcal{K}$. Jelikož $\mathcal{K}$ je uzavřeným konvexním kuželem, existuje podle věty 292 a věty 303 poloprostor $\mathcal{H}(x,0)$ takový, že $\mathcal{K} \subset \mathcal{H}(x,0)$ a $z \notin \mathcal{H}(x,0)$, neboli $x^T y \leq 0\ \forall y \in \mathcal{K}$ (takže $x \in \mathcal{K}^*$) a $x^T z > 0$. Protože $x \in \mathcal{K}^*$ a $x^T z > 0$, musí platit $z \notin (\mathcal{K}^*)^*$. □

Věta 308. *Nechť $\mathcal{K} \subset R^n$ je uzavřený konvexní kužel. Pak*

$$\mathcal{K}^* = \{x \in R^n : P_{\mathcal{K}}(x) = 0\}.$$

Důkaz Nechť $P_{\mathcal{K}}(x) = 0$. Pak podle věty 292 existuje tečný poloprostor množiny $\mathcal{K}$ s normálovým vektorem $a = x - P_{\mathcal{K}}(x) = x$ a číslem $\alpha = a^T P_{\mathcal{K}}(x) = 0$, takže $x^T y \leq 0\ \forall y \in \mathcal{K}$, neboli $x \in \mathcal{K}^*$. Nechť naopak $x \in \mathcal{K}^*$. Pak $x^T y \leq 0\ \forall y \in \mathcal{K}$, takže $\mathcal{K} \subset \mathcal{H}(x,0)$, a jelikož $P_{\mathcal{H}(x,0)}(x) = 0$, platí též $P_{\mathcal{K}}(x) = 0$. □

Věta 309. *Nechť $\mathcal{K} \subset R^n$ je uzavřený konvexní kužel. Pak $\mathcal{K}^*$ je sjednocením normálových vektorů tečných poloprostorů kuželu $\mathcal{K}$, neboli*

$$\mathcal{K}^* = \bigcup_{\mathcal{K} \subset \mathcal{H}(a,0)} a.$$

Důkaz Označme $\tilde{\mathcal{K}}^*$ množinu na pravé straně dokazované rovnosti. Nechť $a \in \tilde{\mathcal{K}}^*$. Pak $\mathcal{H}(a,0)$ je tečným poloprostorem kuželu $\mathcal{K}$, takže $a^T y \leq 0\ \forall y \in \mathcal{K}$, neboli $a \in \mathcal{K}^*$. Nechť naopak $a \in \mathcal{K}^*$. Pak $a^T y \leq 0$ $\forall y \in \mathcal{K}$, takže $\mathcal{K} \subset \mathcal{H}(a,0)$. Jelikož $\mathcal{H}(a,0)$ je podle věty 303 tečným poloprostorem množiny $\mathcal{K}$, platí $a \in \tilde{\mathcal{K}}^*$. □

Definice 108. *Kužel $\mathcal{K} \in R^n$, který je průnikem konečného počtu tečných poloprostorů, se nazvývá polyedrálním kuželem*

Věta 310. *Nechť $\mathcal{K} \in R^n$ je polyedrální kužel takový, že*

$$\mathcal{K} = \bigcap_{i=1}^{m} \mathcal{H}(a_i, 0)$$

Pak

$$\mathcal{K}^* = \text{cone}\,(\text{conv}\,\{a_i : 1 \leq i \leq m\}).$$

Důkaz Označme $\tilde{\mathcal{K}}^*$ množinu na pravé straně dokazované rovnosti. Nechť $a \in \tilde{\mathcal{K}}^*$. Pak podle věty 304 existují čísla $\lambda_i \geq 0$, $1 \leq i \leq m$, taková, že

$$a = \sum_{i=1}^{m} \lambda_i a_i.$$

Jelikož $a_i^T y \leq 0\ \forall y \in \mathcal{K}$ a $\lambda_i \geq 0$, platí $a^T y \leq 0\ \forall y \in \mathcal{K}$, takže $a \in \mathcal{K}^*$. Nechť naopak $a \notin \tilde{\mathcal{K}}^*$. Pak podle věty 292 existuje vektor $x \in R^n$ takový, že $x^T a_i \leq 0$, $1 \leq i \leq m$, a $x^T a > 0$. To znamená, že $x \in \mathcal{H}(a_i, 0)$, $1 \leq i \leq m$, neboli $x \in \mathcal{K}$, a jelikož $x^T a > 0$, musí platit $a \notin \mathcal{K}^*$. □

Příklad 15. Uvažujme polyedrální kužel

$$\mathcal{K} = \mathcal{H}(e_1,0) \cap \mathcal{H}(e_2,0) \cap \mathcal{L}(e_3,0) = \mathcal{H}(e_1,0) \cap \mathcal{H}(e_2,0) \cap \mathcal{H}(e_3,0) \cap \mathcal{H}(-e_3,0),$$

kde e_1,e_2, e_3 jsou sloupce jednotkové matice řádu 3, neboli $\mathcal{K} = \{x : x_1 \leq 0, x_2 \leq 0, x_3 = 0\}$, takže $\mathcal{K}$ má dimenzi 2 (je to záporný kvadrant roviny $\mathcal{L}(e_3,0)$). Pak podle věty 310 platí

$$\begin{aligned} \mathcal{K}^* &= \{a \in R^3 : a = \lambda_1 e_1 + \lambda_2 e_2 + \lambda_3 e_3,\ \lambda_1 \geq 0,\ \lambda_2 \geq 0,\ \lambda_3 \in R\} \\ &= \{a \in R^3 : a_1 \geq 0,\ a_2 \geq 0\} = \mathcal{H}(-e_1,0) \cap \mathcal{H}(-e_2,0), \end{aligned}$$

takže $\mathcal{K}^*$ je sjednocením dvou oktantů prostoru R^3 (vektory $a \in \mathcal{K}^*$ mají nezáporné první dvě složky).

Definice 109. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Kuželem přípustných směrů množiny $\mathcal{C}$ v bodě x nazveme množinu*

$$\mathcal{F}_{\mathcal{C}}(x) = \{y \in R^n :\ \text{existuje číslo}\ \bar{t} > 0\ \text{takové, že}\ x + ty \in \mathcal{C}\ \text{pro}\ 0 \leq t \leq \bar{t}\}$$

Věta 311. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Pak $\mathcal{F}_{\mathcal{C}}(x)$ je kuželem. Je-li $\mathcal{C}$ konvexní, je i $\mathcal{F}_{\mathcal{C}}(x)$ konvexní a platí $\mathcal{C} - x \subset \mathcal{F}_{\mathcal{C}}(x)$.*

Důkaz (a) Zřejmě $0 \in \mathcal{F}_{\mathcal{C}}(x)$. Nechť $y \in \mathcal{F}_{\mathcal{C}}(x)$ a $\lambda > 0$. Podle definice 109 existuje číslo $\bar{t} > 0$ takové, že $x + ty \in \mathcal{C}$ pro $0 \leq t \leq \bar{t}$. Pak ale $x + t\lambda y \in \mathcal{C}$ pro $0 \leq t \leq \bar{t}/\lambda$, takže $\lambda y \in \mathcal{F}_{\mathcal{C}}(x)$ a množina $\mathcal{F}_{\mathcal{C}}(x)$ je kuželem.

(b) Nechť $y_1 \in \mathcal{F}_{\mathcal{C}}(x)$ a $y_2 \in \mathcal{F}_{\mathcal{C}}(x)$. Podle definice 109 existují čísla $\bar{t}_1$ a $\bar{t}_1$ taková, že $x + ty_1 \in \mathcal{C}$ pro $0 \leq t \leq \bar{t}_1$ a $x + ty_2 \in \mathcal{C}$ pro $0 \leq t \leq \bar{t}_2$. Nechť $0 \leq \lambda \leq 1$ a $y = \lambda y_1 + (1-\lambda)y_2$. Pak, je-li $\mathcal{C}$ konvexní, platí

$$x + ty = x + t(\lambda y_1 + (1-\lambda)y_2) = \lambda(x + ty_1) + (1-\lambda)(x + ty_2) \in \mathcal{C}$$

pro $0 \leq t \leq \min(\bar{t}_1, \bar{t}_2)$, takže $y \in \mathcal{F}_{\mathcal{C}}(x)$. Zbytek tvrzení plyne z toho, že pokud $y \in \mathcal{C} - x$, platí podle (1111) $x + ty \in \mathcal{C}$ pro $0 \leq t \leq 1$, takže $y \in \mathcal{F}_{\mathcal{C}}(x)$. □

Definice 110. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Tečným kuželem množiny $\mathcal{C}$ v bodě x nazveme množinu*

$$\mathcal{T}_{\mathcal{C}}(x) = \{y \in R^n : \text{ existují posloupnosti } y_i \to y, \ t_i \downarrow 0 \text{ takové, že } x + t_i y_i \in \mathcal{C}\}$$

Poznámka 409. Tečný kužel obsahuje kužel přípustných směrů. Z definice 109 plyne, že množina $\mathcal{F}_{\mathcal{C}}(x)$ je kuželem přípustných směrů množiny $\mathcal{C}$ v bodě x právě tehdy, když

$$\mathcal{F}_{\mathcal{C}}(x) = \{y \in R^n : \text{ existuje posloupnost } t_i \downarrow 0 \text{ taková, že } x + t_i y \in \mathcal{C}\}.$$

Zvolíme-li v definici 110 $y_i = y$, $i \in \mathcal{N}$, dostaneme $\mathcal{F}_{\mathcal{C}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$.

Věta 312. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Pak $\mathcal{T}_{\mathcal{C}}(x)$ je uzavřeným kuželem. Je-li $\mathcal{C}$ konvexní, je i $\mathcal{T}_{\mathcal{C}}(x)$ konvexní a platí $\mathcal{C} - x \subset \mathcal{T}_{\mathcal{C}}(x)$.*

Důkaz (a) Nechť $y^k \in \mathcal{T}_{\mathcal{C}}(x)$, $y^k \to y$ a $\varepsilon > 0$. Pak existuje index $\overline{k} \in N$ takový, že $\|y^k - y\| < \varepsilon/2 \ \forall k \geq \overline{k}$. Jelikož $y^k \in \mathcal{T}_{\mathcal{C}}(x)$, existují posloupnosti

$$y_i^k \to y^k, \quad t_i^k \downarrow 0$$

takové, že $x + t_i^k y_i^k \in \mathcal{C}$. Pro každé $k \in N$ tedy existuje index $\overline{i}_k \in N$ takový, že

$$\|y_i^k - y^k\| < \varepsilon/2, \quad t_i^k < 1/k, \ \text{ a } \ x + t_i^k y_i^k \in \mathcal{C}$$

$\forall i \geq \overline{i}_k$. Zkonstruujeme-li posloupnost indexů $\{i_k\} \subset N$ tak, že $i_1 = \overline{i}_1$ a $i_{k+1} = \max(i_k + 1, \overline{i}_{k+1})$, můžeme psát

$$\|y_{i_k}^k - y^k\| < \varepsilon/2, \quad t_{i_k}^k < 1/k \ \text{ a } \ x + t_{i_k}^k y_{i_k}^k \in \mathcal{C}$$

pro libovolný index $k \in N$ a

$$\|y_{i_k}^k - y\| \leq \|y_{i_k}^k - y^k\| + \|y^k - y\| < \varepsilon$$

pro $k \geq \overline{k}$. Platí tedy $y_{i_k}^k \to y$, $t_{i_k}^k \downarrow 0$ a $x + t_{i_k}^k y_{i_k}^k \in \mathcal{C}$, což implikuje $y \in \mathcal{T}_{\mathcal{C}}(x)$, takže množina $\mathcal{T}_{\mathcal{C}}(x)$ je uzavřená.

(b) Nechť $y \in \mathcal{T}_{\mathcal{C}}(x)$ a $\lambda \geq 0$. Podle definice 110 existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in \mathcal{C}$. Pak ale

$$\lambda y_i \to \lambda y, \quad t_i/\lambda \downarrow 0 \ \text{ a } \ x + (t_i/\lambda)\lambda y_i = x + t_i y_i \in \mathcal{C},$$

takže $\lambda y \in \mathcal{T}_{\mathcal{C}}(x)$ a množina $\mathcal{T}_{\mathcal{C}}(x)$ je kuželem.

(c) Nechť $y^1 \in \mathcal{T}_{\mathcal{C}}(x)$ a $y^2 \in \mathcal{T}_{\mathcal{C}}(x)$. Podle definice 110 existují posloupnosti

$$y_i^1 \to y^1, \quad t_i^1 \downarrow 0, \quad y_i^2 \to y^2, \quad t_i^2 \downarrow 0$$

takové, že $x + t_i^1 y_i^1 \in \mathcal{C}$, $x + t_i^2 y_i^2 \in \mathcal{C}$. Je-li $\mathcal{C}$ konvexní, platí podle (1111) $x + t_i y_i^1 \in \mathcal{C}$, $x + t_i y_i^2 \in \mathcal{C}$ pro $t_i = \min(t_i^1, t_i^2)$. Nechť $0 \leq \lambda \leq 1$. Označme $y = \lambda y^1 + (1-\lambda)y^2$ a $y_i = \lambda y_i^1 + (1-\lambda)y_i^2$, $i \in N$. Pak

$$y_i = \lambda y_i^1 + (1-\lambda)y_i^2 \to \lambda y^1 + (1-\lambda)y^2 = y,$$

$t_i \downarrow 0$ a

$$x + t_i y_i = \lambda(x + t_i y_i^1) + (1-\lambda)(x + t_i y_i^2) \in \mathcal{C},$$

takže $y \in \mathcal{T}_{\mathcal{C}}(x)$ a množina $\mathcal{T}_{\mathcal{C}}(x)$ je konvexní. Zbytek tvrzení plyne z toho, že pokud $y \in \mathcal{C} - x$, platí podle věty 311 a poznámky 409 $y \in \mathcal{F}_{\mathcal{C}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$. □

Poznámka 410. Nechť $y \in \mathcal{T}_{\mathcal{C}}(x)$, takže existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $z_i = x + t_i y_i \in \mathcal{C}$. Pak platí

$$z_i - x = t_i y_i = t_i(y + o(1)) = t_i y + o(t_i) \quad \Rightarrow \quad y = \lim_{i \to \infty} \frac{z_i - x}{t_i}.$$

Věta 313. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Pak*

$$\mathcal{F}_{\mathcal{C}}(x) = \text{cone}\,(\mathcal{C} - x) = \bigcup_{\lambda \geq 0} \lambda(\mathcal{C} - x), \qquad \mathcal{T}_{\mathcal{C}}(x) = \overline{\mathcal{F}_{\mathcal{C}}(x)}.$$

Důkaz (a) Nechť $y \in \text{cone}\,(\mathcal{C} - x)$. Pokud $y = 0$, platí $y \in \mathcal{F}_{\mathcal{C}}(x)$, neboť $\mathcal{F}_{\mathcal{C}}(x)$ je kužel. Pokud $y \neq 0$, existuje bod $z \in \mathcal{C}$ a číslo $\lambda > 0$ tak, že $y = \lambda(z - x)$ neboli $x + y/\lambda \in \mathcal{C}$. Jelikož množina $\mathcal{C}$ je konvexní, platí $x + ty \in \mathcal{C}$ pro $0 \leq t \leq 1/\lambda$, takže $y \in \mathcal{F}_{\mathcal{C}}(x)$. Nechť naopak $y \in \mathcal{F}_{\mathcal{C}}(x)$. Pak $x + \bar{t}y \in \mathcal{C}$ a tedy $\bar{t}y \in \mathcal{C} - x$ pro nějaké $\bar{t} > 0$, což dává $y \in \text{cone}\,(\mathcal{C} - x)$.

(b) Nechť $y \in \mathcal{F}_{\mathcal{C}}(x)$. Pak existuje $\bar{t} > 0$ tak, že $x + ty \in \mathcal{C}$, pokud $0 \leq t \leq \bar{t}$. Zvolme posloupnosti $y_i = y$, $t_i = \bar{t}/i$, $i \in N$. Pak $y_i \to y$, $t_i \downarrow 0$ a $x + y_i t_i \in \mathcal{C}$, takže $y \in \mathcal{T}_{\mathcal{C}}(x)$. Platí tedy $\mathcal{F}_{\mathcal{C}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$ a jelikož $\mathcal{T}_{\mathcal{C}}(x)$ je uzavřená množina, též $\overline{\mathcal{F}_{\mathcal{C}}(x)} \subset \mathcal{T}_{\mathcal{C}}(x)$. Nechť naopak $y \in \mathcal{T}_{\mathcal{C}}(x)$. Pak existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in \mathcal{C}$. Označme $z_i = x + t_i y_i \in \mathcal{C}$. Pak $y_i = (z_i - x)/t_i$, takže $y_i \in \text{cone}\,(\mathcal{C} - x)$. Jelikož $y_i \to y$, platí $y \in \overline{\text{cone}\,(\mathcal{C} - x)}$, takže $\mathcal{T}_{\mathcal{C}}(x) \subset \overline{\text{cone}\,(\mathcal{C} - x)} = \overline{\mathcal{F}_{\mathcal{C}}(x)}$. □

Věta 314. *Nechť uzavřené množiny $\mathcal{C}_1, \ldots \mathcal{C}_m$ jsou konvexní a $x \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m$. Pak platí*

$$\mathcal{F}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x) = \mathcal{F}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{F}_{\mathcal{C}_m}(x) \tag{1115}$$

a pokud $\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond \neq \emptyset$, platí

$$\mathcal{T}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x) = \mathcal{T}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{T}_{\mathcal{C}_m}(x). \tag{1116}$$

Důkaz (a) Nechť $y \in \mathcal{F}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x)$. pak existuje číslo $\bar{t} > 0$ takové, že $x + ty \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m$ pro $0 \leq t \leq \bar{t}$. Pak ale $x + ty \in \mathcal{C}_j$, $1 \leq j \leq m$, pro $0 \leq t \leq \bar{t}$, takže $y \in \mathcal{F}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{F}_{\mathcal{C}_m}(x)$. Nechť naopak $y \in \mathcal{F}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{F}_{\mathcal{C}_m}(x)$. Pak existují čísla $\bar{t}_j > 0$, $1 \leq j \leq m$, taková, že $x + ty \in \mathcal{C}_j$ pro $0 \leq t \leq \bar{t}_j$. Platí tedy $x + ty \in \mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m$ pro $0 \leq t \leq \min(\bar{t}_1, \ldots, \bar{t}_m)$, neboli $y \in \mathcal{F}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x)$.

(b) Pokud $\mathcal{C}_1^\diamond \cap \cdots \cap \mathcal{C}_m^\diamond \neq \emptyset$, je též $(\mathcal{F}_{\mathcal{C}_1}(x))^\diamond \cap \cdots \cap (\mathcal{F}_{\mathcal{C}_m}(x))^\diamond \neq \emptyset$, neboť podle věty 313 platí $\mathcal{F}_{\mathcal{C}_j}(x) \supset \mathcal{C}_j - x$ a $\dim \mathcal{F}_{\mathcal{C}_j}(x) = \dim \mathcal{C}_j$, $1 \leq j \leq m$. Použijeme-li větu 313, vztah (1115) a větu 286, dostaneme

$$\mathcal{T}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x) = \overline{\mathcal{F}_{\mathcal{C}_1 \cap \cdots \cap \mathcal{C}_m}(x)} = \overline{\mathcal{F}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{F}_{\mathcal{C}_m}(x)} = \overline{\mathcal{F}_{\mathcal{C}_1}(x)} \cap \cdots \cap \overline{\mathcal{F}_{\mathcal{C}_m}(x)} = \mathcal{T}_{\mathcal{C}_1}(x) \cap \cdots \cap \mathcal{T}_{\mathcal{C}_m}(x).$$

□

Příklad 16. *Nechť $\mathcal{C}_1 = \{y \in R^2 : (y_1 - 1)^2 + y_2^2 \leq 4\}$ a $\mathcal{C}_2 = \{y \in R^2 : (y_1 + 1)^2 + y_2^2 \leq 4\}$ a $x = [0, -\sqrt{3}]$ ($\mathcal{C}_1$ je kruh se středem v bodě $[1, 0]$ a poloměrem 2 a $\mathcal{C}_2$ je kruh se středem v bodě $[-1, 0]$ a poloměrem 2). Pak $\mathcal{C}_1^\diamond \cap \mathcal{C}_2^\diamond \neq \emptyset$ (tento průnik obsahuje bod $[0, 0]$). Zřejmě*

$$\begin{aligned} \mathcal{T}_{\mathcal{C}_1}(x) &= \{y = [y_1, y_2] \in R^2,\ y_1 \geq -\sqrt{3}\, y_2\} \\ \mathcal{T}_{\mathcal{C}_2}(x) &= \{y = [y_1, y_2] \in R^2,\ y_1 \leq \ \sqrt{3}\, y_2\}, \end{aligned}$$

takže platí

$$\mathcal{T}_{\mathcal{C}_1 \cap \mathcal{C}_2}(x) = \mathcal{T}_{\mathcal{C}_1}(x) \cap \mathcal{T}_{\mathcal{C}_2}(x) = \{y = [y_1, y_2] \in R^2 :\ -\sqrt{3} y_2 \leq y_1 \leq \sqrt{3} y_2\}.$$

Příklad 17. *Nechť $\mathcal{C}_1 = \{y \in R^2 : y_1^2 + (y_2 - 1)^2 \leq 1\}$ a $\mathcal{C}_2 = \{y \in R^2 : y_1^2 + (y_2 + 1)^2 \leq 1\}$ a $x = [0, 0]$ ($\mathcal{C}_1$ je kruh se středem v bodě $[0, 1]$ a poloměrem 1 a $\mathcal{C}_2$ je kruh se středem v bodě $[0, -1]$ a poloměrem 1). Pak $\mathcal{C}_1^\diamond \cap \mathcal{C}_2^\diamond = \emptyset$. Zřejmě*

$$\begin{aligned} \mathcal{F}_{\mathcal{C}_1}(x) &= \{\lambda s : \lambda \geq 0,\ s = [s_1, s_2] \in R^2,\ s_2 > 0\} = \{y = [y_1, y_2] \in R^2,\ y_2 > 0\} \cup [0, 0] \\ \mathcal{F}_{\mathcal{C}_2}(x) &= \{\lambda s : \lambda \geq 0,\ s = [s_1, s_2] \in R^2,\ s_2 < 0\} = \{y = [y_1, y_2] \in R^2,\ y_2 < 0\} \cup [0, 0], \end{aligned}$$

takže tak jako v poznámce 400 dostaneme $\mathcal{F}_{\mathcal{C}_1}(x) \cap \mathcal{F}_{\mathcal{C}_2}(x) = \{[0, 0]\}$ a $\mathcal{F}_{\mathcal{C}_1}^\diamond(x) \cap \mathcal{F}_{\mathcal{C}_2}^\diamond(x) = \emptyset$. Jelikož $\overline{\mathcal{F}_{\mathcal{C}_1}(x)}$ je uzavřená horní polorovina a $\overline{\mathcal{F}_{\mathcal{C}_1}(x)}$ je uzavřená dolní polorovina, platí

$$\mathcal{T}_{\mathcal{C}_1 \cap \mathcal{C}_2}(x) = \overline{\mathcal{F}_{\mathcal{C}_1 \cap \mathcal{C}_2}(x)} = \{[0, 0]\} \neq \{[y_1, 0] : y_1 \in R\} = \overline{\mathcal{F}_{\mathcal{C}_1}(x)} \cap \overline{\mathcal{F}_{\mathcal{C}_2}(x)} = \mathcal{T}_{\mathcal{C}_1}(x) \cap \mathcal{T}_{\mathcal{C}_2}(x).$$

Věta 315. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$ je jejím hraničním bodem. Pak $\mathcal{T}_\mathcal{C}(x)$ je průnikem všech tečných poloprostorů množiny $\mathcal{C} - x$ obsahujících počátek souřadnic, neboli*

$$\mathcal{T}_\mathcal{C}(x) = \bigcap_{\mathcal{C}-x \subset \mathcal{H}(a,0)} \mathcal{H}(a,0).$$

Je-li množina $\mathcal{C} \in R^n$ polyedrální, je i tečný kužel $\mathcal{T}_\mathcal{C}(x)$ polyedrální a existují tečné poloprostory $\mathcal{H}(a_i,0)$, $1 \le i \le m$, takové, že

$$\mathcal{T}_\mathcal{C}(x) = \bigcap_{i=1}^{m} \mathcal{H}(a_i,0).$$

Důkaz (a) Označme $\overline{\mathcal{K}}$ průnik všech tečných poloprostorů množiny $\mathcal{C} - x$ obsahujících počátek souřadnic. Nechť $y \in \text{cone}\,(\mathcal{C} - x)$. Pak existuje bod $z \in \mathcal{C} - x$ takový, že $y = \lambda z$, $\lambda \ge 0$, a pro libovolný tečný poloprostor $\mathcal{H}(a,0)$ množiny $\mathcal{C} - x$ platí $a^T y = \lambda a^T z \le 0$, neboli $y \in \mathcal{H}(a,0)$. Platí tedy cone $(\mathcal{C} - x) \subset \overline{\mathcal{K}}$ a jelikož $\overline{\mathcal{K}}$ je uzavřeným kuželem, též $\mathcal{T}_\mathcal{C}(x) \subset \overline{\mathcal{K}}$.

(b) Nechť $y \in \overline{\mathcal{K}}$ a $y \notin \mathcal{T}_\mathcal{C}(x)$. Jelikož $\mathcal{T}_\mathcal{C}(x)$ je uzavřená konvexní množina, existuje podle věty 292 tečný poloprostor této množiny takový, že $\mathcal{T}_\mathcal{C}(x) \subset \mathcal{H}$ a $y \notin \mathcal{H}$. Podle věty 303 je $0 \in \mathcal{H}$, takže $\mathcal{H}$ je tečným poloprostorem množiny $\mathcal{C} - x$ obsahujícím počátek souřadnic. Platí tedy $\overline{\mathcal{K}} \subset \mathcal{H}$ a jelikož $y \notin \mathcal{H}$, musí být $y \notin \overline{\mathcal{K}}$, což je ve sporu s předpokladem, že $y \in \overline{\mathcal{K}}$.

(c) Je-li množina $\mathcal{C} \in R^n$ polyedrální, má tuto vlastnost i množina $\mathcal{C} - x$. Jelikož $\mathcal{C} - x$ je průnikem konečného počtu poloprostorů, lze i v průniku definujícím $\mathcal{T}_\mathcal{C}(x)$ vybrat konečný počet poloprostorů. □

Definice 111. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Normálovým kuželem množiny $\mathcal{C}$ v bodě x nazveme množinu*

$$\mathcal{N}_\mathcal{C}(x) = \mathcal{T}_\mathcal{C}^*(x),$$

kde $\mathcal{T}_\mathcal{C}^*(x)$ je polární kužel tečného kuželu $\mathcal{T}_\mathcal{C}(x)$.

Poznámka 411. Podle věty 306 je množina $\mathcal{N}_\mathcal{C}(x)$ uzavřeným konvexním kuželem.

Věta 316. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Pak*

$$\mathcal{N}_\mathcal{C}(x) = \{z \in R^n : (y - x)^T z \le 0 \quad \forall y \in \mathcal{C}\}.$$

Důkaz Platí

$$\begin{aligned}
\mathcal{N}_\mathcal{C}(x) &= \{z \in R^n : (y - x)^T z \le 0 \quad \forall (y - x) \in \mathcal{T}_\mathcal{C}(x)\} \\
&= \{z \in R^n : (y - x)^T z \le 0 \quad \forall (y - x) \in \overline{\bigcup_{\lambda \ge 0} \lambda(\mathcal{C} - x)}\} \\
&= \{z \in R^n : (y - x)^T z \le 0 \quad \forall (y - x) \in \bigcup_{\lambda \ge 0} \lambda(\mathcal{C} - x)\} \\
&= \{z \in R^n : (y - x)^T z \le 0 \quad \forall y \in \mathcal{C}\}.
\end{aligned}$$

První rovnost plyne z definic 107 a 111, druhá z věty 313, třetí z invariance neostré nerovnosti vůči limitnímu přechodu a poslední z invariance neostré nerovnosti vůči násobení nezáporným číslem λ. □

Věta 317. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$ je jejím hraničním bodem. Pak $\mathcal{N}_\mathcal{C}(x)$ je sjednocením normálových vektorů tečných poloprostorů množiny $\mathcal{C} - x$ obsahujících počátek souřadnic, neboli*

$$\mathcal{N}_\mathcal{C}(x) = \bigcup_{\mathcal{C}-x \subset \mathcal{H}(a,0)} a.$$

Je-li množina $\mathcal{C} \in R^n$ polyedrální, je i normálový kužel $\mathcal{T}_\mathcal{C}(x)$ polyedrální a existují tečné poloprostory $\mathcal{H}(a_i,0)$, $1 \le i \le m$, takové, že

$$\mathcal{N}_\mathcal{C}(x) = \text{cone}\,(\text{conv}\,\{a_i : 1 \le i \le m\}).$$

Důkaz Toto tvrzení je důsledkem věty 309, věty 310 a věty 315. □

Význam tečného kuželu $\mathcal{T}_{\mathcal{C}}(x)$ a normálového kuželu $\mathcal{N}_{\mathcal{C}}(x)$ pro charakterizaci lokálního minima spojitě diferencovatelné funkce F na množně $\mathcal{C} \in R^n$ dokládá tato věta.

Věta 318. *Nechť $\mathcal{C} \in R^n$ je uzavřená množina a $x \in \mathcal{C}$. Je-li bod x lokálním minimem funkce $F : R^n \to R$, která je spojitě diferencovatelná v nějakém okolí $\mathcal{B}(x, \varepsilon)$ bodu x, platí $-g(x) \in \mathcal{N}_{\mathcal{C}}(x)$.*

Důkaz Nechť $-g(x) \notin \mathcal{N}_{\mathcal{C}}(x)$. Pak podle definice 111 existuje vektor $y \in \mathcal{T}_{\mathcal{C}}(x)$ takový, že $-g(x)^T y > 0$. Jelikož $y \in \mathcal{T}_{\mathcal{C}}(x)$, existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in \mathcal{C}$. Podle věty o střední hodnotě platí

$$\begin{aligned} F(x + t_i y_i) &= F(x) + t_i g^T(x) y_i + o(t_i y_i) = F(x) + t_i g^T(x) y + t_i g^T(x)(y_i - y) + o(t_i y_i) \\ &\leq F(x) + t_i (g^T(x) y + \|g(x)\| \|y_i - y\| + o(1)). \end{aligned}$$

Jelikož $\|y_i - y\| \to 0$ a $o(1) \to 0$ pro $i \to \infty$, existuje index $\bar{i} \in N$ takový, že

$$\|y_i - y\| \leq -\frac{g^T(x) y}{3\|g(x)\|}, \qquad o(1) \leq -\frac{g^T(x) y}{3}$$

pro $i \geq \bar{i}$. Můžeme tedy psát

$$F(x + t_i y_i) \leq F(x) + t_i \left(g^T(x) y - \frac{g^T(x) y}{3} - \frac{g^T(x) y}{3} \right) = F(x) + t_i \frac{g^T(x) y}{3} < F(x)$$

pro $i \geq \bar{i}$, což je ve sporu s definicí lokálního minima, neboť $x + t_i y_i \in \mathcal{C}$ a $x + t_i y_i \to x$. □

Poznámka 412. Podmínku $-g(x) \in \mathcal{N}_{\mathcal{C}}(x)$ lze zapsat ve tvaru $0 \in g(x) + \mathcal{N}_{\mathcal{C}}(x)$.

Poznámka 413. Je-li x vnitřním bodem množiny $\mathcal{C}$, platí $\mathcal{T}_{\mathcal{C}}(x) = R^n$ a $\mathcal{N}_{\mathcal{C}}(x) = \{0\}$, takže tvrzení věty 318 je ekvivalentní první části tvrzení věty 3 (platí $g(x) = 0$).

Věta 319. *Nechť $\mathcal{C} \in R^n$ je polyedrální množina a bod $x \in \mathcal{C}$ je jejím hraničním bodem (takže podle věty 315 platí $\mathcal{T}_{\mathcal{C}}(x) = \bigcap_{i=1}^m \mathcal{H}(a_i, 0)$). Je-li bod x lokálním minimem funkce $F : R^n \to R$, která je spojitě diferencovatelná v nějakém okolí $\mathcal{B}(x, \varepsilon)$ bodu x, existují čísla $\lambda_i \geq 0$, $1 \leq i \leq m$ taková, že*

$$-g(x) = \sum_{i=1}^{m} \lambda_i a_i.$$

Důkaz Dokazované tvrzení je důsledkem věty 310 a věty 318. Poznamenejme, že polyedrální množina je vždy konvexní □

15.3 Konvexní funkce

Definice 112. *Řekneme, že funkce $F : R^n \to R$ je konvexní na konvexní množině $\mathcal{C} \subset R^n$, platí-li*

$$F(\lambda x_1 + (1 - \lambda) x_2) \leq \lambda F(x_1) + (1 - \lambda) F(x_2) \quad \forall\, 0 \leq \lambda \leq 1, \tag{1117}$$

kdykoliv $x_1 \in \mathcal{C}$ a $x_2 \in \mathcal{C}$. Řekneme, že funkce $F : R^n \to R$ je ryze konvexní na konvexní množině $\mathcal{C} \subset R^n$, platí-li

$$F(\lambda x_1 + (1 - \lambda) x_2) < \lambda F(x_1) + (1 - \lambda) F(x_2) \quad \forall\, 0 < \lambda < 1, \tag{1118}$$

kdykoliv $x_1 \in \mathcal{C}$, $x_2 \in \mathcal{C}$ a $x_1 \neq x_2$. Poznamenejme, že ryze konvexní funkce je vždy konvexní, neboť pro $\lambda = 0$, $\lambda = 1$ nebo $x_1 = x_2$ platí v (1117) rovnost pro každou funkci $F : R^n \to R$.

Poznámka 414. Vztahy (1117) a (1118) můžeme zapsat v ekvivalentním tvaru

$$F(x_1 + \lambda'(x_2 - x_1)) \le F(x_1) + \lambda'(F(x_2) - F(x_1)) \quad \forall\, 0 \le \lambda' \le 1,$$
$$F(x_1 + \lambda'(x_2 - x_1)) < F(x_1) + \lambda'(F(x_2) - F(x_1)) \quad \forall\, 0 < \lambda' < 1,$$

kde $\lambda' = 1 - \lambda$.

Poznámka 415. Indukcí snadno dokážeme, že z $x_i \in \mathcal{C}$, $\lambda_i \ge 0$ a $\sum_{i=1}^m \lambda_i = 1$ plyne

$$F\left(\sum_{i=1}^m \lambda_i x_i\right) \le \sum_{i=1}^m \lambda_i F(x_i),$$

pokud F je konvexní na $\mathcal{C}$ (princip důkazu je shodný s postupem uvedeným v důkazu věty 281).

Důležitou vlastnost konvexních funkcí, které nejsou ryze konvexní udává tato věta.

Věta 320. *Je-li funkce F konvexní, ale není ryze konvexní na $\mathcal{C}$, existují body $x \in \mathcal{C}$, $y \in \mathcal{C}$, $y \ne x$ takové, že funkce F je lineární na úsečce $[x, y]$.*

Důkaz Je li funkce F konvexní, ale není ryze konvexní na $\mathcal{C}$ existují body $x \in \mathcal{C}$, $y \in \mathcal{C}$, $y \ne x$ a číslo $0 < \lambda' < 1$ tak, že

$$x' = x + \lambda'(y - x) \in \mathcal{C}, \qquad F(x') = F(x) + \lambda'(F(y) - F(x)) \tag{1119}$$

(používáme nerovnosti uvedené v poznámce 414, kde $x_1 = y$ a $x_2 = x$). Předpokládejme, že pro nějaké číslo $\lambda' < \lambda'' < 1$ platí

$$x'' = x + \lambda''(y - x) \in \mathcal{C}, \qquad F(x'') < F(x) + \lambda''(F(y) - F(x)). \tag{1120}$$

Bod x' lze vyjádřit ve tvaru

$$x' = x + \lambda(x'' - x) = x + \lambda(x + \lambda''(y - x) - x) = x + \lambda\lambda''(y - x),$$

takže $\lambda' = \lambda\lambda''$. Protože funkce F je konvexní na $\mathcal{C}$ a platí (1120), dostaneme

$$\begin{aligned} F(x') &\le F(x) + \lambda(F(x'') - F(x)) < F(x) + \lambda(F(x) + \lambda''(F(y) - F(x)) - F(x)) \\ &= F(x) + \lambda\lambda''(F(y) - F(x)) = F(x) + \lambda'(F(y) - F(x)) \end{aligned}$$

což je ve sporu s (1119). Musí tedy platit $F(x + \lambda''(y - x)) - F(x) = \lambda''(F(y) - F(x))$ pro $\lambda' \le \lambda'' \le 1$, což znamená, že funkce F je lineární na úsečce $[x', y]$. Abychom dokázali, že funkce F je lineární na úsečce $[x, x']$, stačí obrátit pořadí bodů x a y. □

Definice 113. *Řekneme, že funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$, existuje-li číslo $\varepsilon > 0$ takové, že F je definovaná a konvexní v $\mathcal{B}(x, \varepsilon) = \{y : \|y - x\| < \varepsilon\}$. Řekneme, že funkce $F : R^n \to R$ je ryze konvexní v okolí bodu $x \in R^n$, je-li ryze konvexní v $\mathcal{B}(x, \varepsilon)$.*

Poznámka 416. Položíme-li $\mathcal{C} = \mathcal{B}(x, \varepsilon)$, lze definici 113 nahradit definicí 112. Musíme však mít na paměti, že množina $\mathcal{C}$ je v tomto případě otevřená, takže má dimenzi n.

Věta 321. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak F je lipschitzovská v okolí bodu x.*

Důkaz Jelikož F je konvexní v okolí bodu x, existuje číslo $\varepsilon > 0$ takové, že F je definovaná a konvexní v $\mathcal{B}(x, \varepsilon\sqrt{n+1})$ a tudíž i v nadkrychli

$$N(x, \varepsilon) = \{y \in R^n : x_i - \varepsilon \le y_i \le x_i + \varepsilon,\ 1 \le i \le n\} \subset \mathcal{B}(x, \varepsilon\sqrt{n+1}).$$

Nechť $y^{(k)}$, $1 \le k \le 2^n$, jsou vrcholy této nadkrychle. Označme

$$M = \max_{1\le k\le 2^n} F(y^{(k)}).$$

Jelikož každý bod $N(x,\varepsilon)$ lze podle věty 295 vyjádřit jako konvexní kombinaci vrcholů $y^{(k)}$, $1 \le k \le 2^n$, platí to i o bodech okolí $\mathcal{B}(x,\varepsilon) \subset N(x,\varepsilon)$. Nechť tedy $y \in \mathcal{B}(x,\varepsilon)$. Pak platí

$$y = \sum_{k=1}^{2^n} \lambda_k y^{(k)}, \quad \sum_{k=1}^{2^n} \lambda_k = 1,$$

kde $\lambda_k \ge 0$, $1 \le k \le 2^n$, takže

$$F(y) = F\left(\sum_{k=1}^{2^n} \lambda_k y^{(k)}\right) \le \sum_{k=1}^{2^n} \lambda_k F(y^{(k)}) \le M \sum_{k=1}^{2^n} \lambda_k = M.$$

Funkce F je tedy omezená shora na $\mathcal{B}(x,\varepsilon)$. Zvolme nyní $y \in \mathcal{B}(x,\varepsilon)$ a položme $y' = 2x - y$. Pak $\|y' - x\| = \|x - y\| < \varepsilon$ takže $y' \in \mathcal{B}(x,\varepsilon)$. Z konvexity plyne

$$F(x) = F\left(\frac{y+y'}{2}\right) \le \frac{1}{2}(F(y) + F(y')),$$

takže

$$F(y) \ge 2F(x) - F(y') \ge 2F(x) - M$$

a funkce F je omezená zdola na $\mathcal{B}(x,\varepsilon)$. Položme $\delta = \varepsilon/2$ a $m = 2F(x) - M$. Nechť $z \in \mathcal{B}(x,\delta)$, $z' \in \mathcal{B}(x,\delta)$ a $z \ne z'$. Položme

$$z'' = z' + \delta\frac{z' - z}{\|z' - z\|} \in \mathcal{B}(x,\varepsilon).$$

Přímým výpočtem dostaneme

$$z' = \frac{\|z' - z\|}{\delta + \|z' - z\|} z'' + \frac{\delta}{\delta + \|z' - z\|} z$$

a z konvexity plyne

$$\begin{aligned} F(z') - F(z) &\le \frac{\|z' - z\|}{\delta + \|z' - z\|} F(z'') + \frac{\delta}{\delta + \|z' - z\|} F(z) - F(z) \\ &= \frac{\|z' - z\|}{\delta + \|z' - z\|} (F(z'') - F(z)) \le \frac{1}{\delta}\|z' - z\|(M - m). \end{aligned}$$

Jelikož nezáleží na pořadí bodů z a z', dostaneme

$$|F(z') - F(z)| \le \frac{M - m}{\delta}\|z' - z\|,$$

takže F je lipschitzovská s konstantou $L = (M - m)/\delta$ na $\mathcal{B}(x,\delta)$. □

Lemma 123. *Nechť funkce $\varphi : R \to R$ je konvexní na intervalu $[a,b]$ a nechť $a \le t_1 < t_2 < t_3 \le b$. Pak platí*

$$\frac{\varphi(t_2) - \varphi(t_1)}{t_2 - t_1} \le \frac{\varphi(t_3) - \varphi(t_1)}{t_3 - t_1} \le \frac{\varphi(t_3) - \varphi(t_2)}{t_3 - t_2}.$$

Je-li funkce φ ryze konvexní, platí

$$\frac{\varphi(t_2) - \varphi(t_1)}{t_2 - t_1} < \frac{\varphi(t_3) - \varphi(t_1)}{t_3 - t_1} < \frac{\varphi(t_3) - \varphi(t_2)}{t_3 - t_2}.$$

Důkaz Zřejmě

$$t_2 = t_1 + \frac{t_2 - t_1}{t_3 - t_1}(t_3 - t_1),$$

kde $0 < (t_2 - t_1)/(t_3 - t_1) < 1$. Z konvexity funkce φ (poznámka 414) pak dostaneme

$$\varphi(t_2) \leq \varphi(t_1) + \frac{t_2 - t_1}{t_3 - t_1}(\varphi(t_3) - \varphi(t_1)),$$

což dokazuje levou nerovnost. Pravá nerovnost se dokazuje analogicky pomocí vztahu

$$t_3 = t_2 + \frac{t_3 - t_2}{t_3 - t_1}(t_3 - t_1).$$

Z konvexity funkce φ pak plyne

$$\varphi(t_3) \leq \varphi(t_2) + \frac{t_3 - t_2}{t_3 - t_1}(\varphi(t_3) - \varphi(t_1)),$$

což dokazuje pravou nerovnost. Je-li funkce φ ryze konvexní, lze použít ostré nerovnosti.

Definice 114. *Řekneme, že funkce $F : R^n \to R$ má v bodě $x \in R^n$ směrovou derivaci ve směru $h \in R^n$, existuje-li konečná limita*

$$F'(x, h) = \lim_{t \downarrow 0} \frac{F(x + th) - F(x)}{t} \tag{1121}$$

(symbol $t \downarrow 0$ značí, že $t > 0$ a $t \to 0$).

Věta 322. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$ (takže je konvexní a lipschitzovská s nějakou konstantou L v nějaké nadkouli $\mathcal{B}(x, \varepsilon)$, $\varepsilon > 0$). Pak:*

(a) Směrová derivace $F'(x, h)$ existuje pro každé $h \in R^n$ a platí

$$F'(x, h) = \inf_{0 < t < \varepsilon/\|h\|} \frac{F(x + th) - F(x)}{t}$$

(diferenční podíl v tomto vyjádření je neklesající funkcí proměnné t, $0 < t < \varepsilon/\|h\|$).

(b) Funkce $F'(x, \cdot) : R^n \to R$ je pozitivně homogenní, subaditivní a lipschitzovská s konstantou L.

(c) Funkce $F'(\cdot, \cdot) : R^n \times R^n \to R$ je shora polospojitá, neboli

$$\limsup_{i \to \infty} F'(x_i, h_i) \leq F'(x, h),$$

kdykoliv $x_i \to x$ a $h_i \to h$.

Důkaz (a) Nechť funkce F je konvexní v $\mathcal{B}(x, \varepsilon)$. Podle lemmatu 123 je funkce

$$\varphi(t) = \frac{F(x + th) - F(x)}{t}$$

neklesající (levá nerovnost) a zdola omezená pro $0 < t < \varepsilon/\|h\|$ (kdy $x + th \in \mathcal{B}(x, \varepsilon)$). Existuje tedy limita (1121). Zbytek tvrzení (a) plyne z toho, že $\varphi(t)$ je neklesající pro $0 < t < \varepsilon/\|h\|$.

(b) Nechť $\lambda > 0$. Pak platí

$$F'(x, \lambda h) = \lim_{t \downarrow 0} \frac{F(x + t\lambda h) - F(x)}{t} = \lambda \lim_{t \downarrow 0} \frac{F(x + t\lambda h) - F(x)}{\lambda t} = \lambda F'(x, h),$$

takže $F'(x,\cdot)$ je pozitivně homogenní. Dále platí

$$\begin{aligned} F'(x,h_1+h_2) &= \lim_{t\downarrow 0}\frac{F(x+t(h_1+h_2))-F(x)}{t} \\ &= \lim_{t\downarrow 0}\frac{F\left(\frac{1}{2}(x+2th_1)+\frac{1}{2}(x+2th_2)\right)-F(x)}{t} \\ &\le \lim_{t\downarrow 0}\frac{F(x+2th_1)-F(x)}{2t}+\lim_{t\downarrow 0}\frac{F(x+2th_2)-F(x)}{2t} \\ &= F'(x,h_1)+F'(x,h_2), \end{aligned}$$

takže $F'(x,\cdot)$ je subaditivní. Dále platí

$$F(x+th_2)-F(x+th_1)\le Lt\|h_2-h_1\|$$

pro $t>0$. Můžeme tedy psát

$$\lim_{t\downarrow 0}\frac{F(x+th_2)-F(x)}{t}\le\lim_{t\downarrow 0}\frac{F(x+th_1)-F(x)}{t}+L\|h_2-h_1\|,$$

takže

$$F'(x,h_2)-F'(x,h_1)\le L\|h_2-h_1\|.$$

Protože nezáleží na pořadí vektorů h_1 a h_2, platí

$$|F'(x,h_2)-F'(x,h_1)|\le L\|h_2-h_1\|,$$

což dokazuje lipschitzovskost $F(x,\cdot)$.

(c) Nechť $x_i\to x$ a $h_i\to h$. Položme $t_i=\sqrt{\|x_i-x\|}+1/i$ (člen $1/i$ je tam proto, aby platilo $t_i>0$ i když $x_i=x$) a předpokládejme bez újmy na obecnosti, že všechny body $x+t_ih$, $x+t_ih_i$ a $x_i+t_ih_i$ leží v $\mathcal{B}(x,\varepsilon)$. Pak podle (a) platí

$$F'(x_i,h_i)\le\frac{F(x_i+t_ih_i)-F(x_i)}{t_i}=\frac{F(x+t_ih)-F(x)}{t_i}+\frac{F(x_i+t_ih_i)-F(x+t_ih)}{t_i}+\frac{F(x)-F(x_i)}{t_i}.$$

Ale

$$\frac{|F(x_i+t_ih_i)-F(x+t_ih)|}{t_i}\le\frac{L(\|x_i-x\|+t_i\|h_i-h\|)}{t_i}\le L(\sqrt{\|x_i-x\|}+\|h_i-h\|)\to 0$$

a

$$\frac{|F(x_i)-F(x)|}{t_i}\le\frac{L\|x_i-x\|}{t_i}\le L\sqrt{\|x_i-x\|}\to 0.$$

Můžeme tedy psát

$$\limsup_{i\to\infty}F'(x_i,h_i)\le\limsup_{i\to\infty}\frac{F(x+t_ih)-F(x)}{t_i}=\lim_{t\downarrow 0}\frac{F(x+th)-F(x)}{t}=F'(x,h)$$

□

Poznámka 417. Z části (b) důkazu věty 322 vyplývá, že pokud směrová derivace $F'(x,\cdot)$ existuje, je pozitivně homogenní a je-li funkce $F:R^n\to R$ lipschitzovská v okolí bodu $x\in R^n$, je $F'(x,\cdot)$ lipschitzovská a tudíž spojitá (tato dvě dílčí tvrzení nevyžadují konvexitu).

Poznámka 418. Podle definice 114 platí $F'(x,0)=0$, takže podle věty 322 (b) dostaneme

$$|F'(x,h)| = |F'(x,h) - F'(x,0)| \leq L\|h\|.$$

Definice 115. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak množinu*

$$\partial F(x) = \left\{ g \in R^n : F'(x,h) \geq g^T h \ \ \forall h \in R^n \right\}$$

nazveme subdiferenciálem funkce F v bodě x. Elementy $g \in \partial F(x)$ budeme nazývat subgradienty funkce F v bodě x.

Věta 323. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$ (takže je konvexní a lipschitzovská s nějakou konstantou L v nějaké nadkouli $\mathcal{B}(x,\varepsilon)$, $\varepsilon > 0$). Pak:*

(a) Subdiferenciál $\partial F(x)$ je neprázdná konvexní kompaktní množina taková, že $\|g\| \leq L \ \forall g \in \partial F(x)$.
(b) Pro libovolný vektor $h \in R^n$ platí

$$F'(x,h) = \max\left\{ g^T h : g \in \partial F(x) \right\}.$$

(c) Jestliže $x_i \to x$, $g_i \in \partial F(x_i)$ a $g_i \to g$, pak $g \in \partial F(x)$ (polospojitost shora).
(d) Vztah $g \in \partial F(x)$ platí právě tehdy, když

$$F(x+h) - F(x) \geq g^T h \quad \forall h \in \mathcal{B}(0,\varepsilon). \tag{1122}$$

Důkaz (a) Podle věty 322 (b) je funkce $F'(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (důsledek 32) existuje vektor $g \in R^n$ takový, že

$$g^T h \leq F'(x,h) \quad \forall h \in R^n,$$

takže subdiferenciál $\partial F(x)$ je neprázdný. Nechť $g_1 \in \partial F(x)$, $g_2 \in \partial F(x)$ a $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$. Pak pro libovolný vektor $h \in R^n$ platí

$$(\lambda_1 g_1 + \lambda_2 g_2)^T h = \lambda_1 g_1^T h + \lambda_2 g_2^T h \leq \lambda_1 F'(x,h) + \lambda_2 F'(x,h) = F'(x,h),$$

takže subdiferenciál $\partial F(x)$ je konvexní. Nechť $g \in \partial F(x)$. Podle definice 115 a poznámky 418 platí

$$\|g\|^2 = g^T g \leq F'(x,g) \leq L\|g\|,$$

čili $\|g\| \leq L$, takže subdiferenciál $\partial F(x)$ je omezený. Nechť $g_i \in \partial F(x)$ a $g_i \to g$. Pak platí

$$g^T h = \lim_{i\to\infty} g_i^T h \leq F'(x,h),$$

takže $g \in \partial F(x)$ a subdiferenciál $\partial F(x)$ je uzavřený.

(b) Podle definice 115 platí

$$F'(x,h) \geq \max\left\{ g^T h : g \in \partial F(x) \right\}.$$

Předpokládejme, že pro nějaký vektor $\overline{h} \in R^n$ platí

$$F'(x,\overline{h}) > \max\left\{ g^T \overline{h} : g \in \partial F(x) \right\}. \tag{1123}$$

Uvažujme lineární funkci $l(\lambda\overline{h}) = \lambda F'(x,\overline{h})$ definovanou na jednorozměrném podprostoru $\{\lambda\overline{h} : \lambda \in R\} \subset R^n$. Jelikož $F'(x,\cdot)$ je pozitivně homogenní a subaditivní, existuje podle Hahn-Banachovy věty vektor

$\overline{g} \in R^n$ takový, že $F'(x,h) \geq \overline{g}^T h \ \forall h \in R^n$ a $\overline{g}^T(\lambda\overline{h}) = l(\lambda\overline{h}) = \lambda F'(x,\overline{h})$. Tedy $\overline{g} \in \partial F(x)$ a pro $\lambda = 1$ dostaneme $F'(x,\overline{h}) = \overline{g}^T\overline{h}$, což je ve sporu s předpokladem (1123).

(c) Nechť $x_i \to x$ a $g_i \to g$, kde $g_i \in \partial F(x_i)$. Pak pro libovolný vektor $h \in R^n$ platí

$$g^T h = \lim_{i\to\infty} g_i^T h \leq \limsup_{i\to\infty} F'(x_i,h).$$

Podle věty 322 (c) je funkce $F'(\cdot,.)$ shora polospojitá, takže $g^T h \leq F'(x,h)$.

(d) Nechť funkce F je konvexní v $\mathcal{B}(x,\varepsilon)$ a $g \in \partial F(x)$. Podle definice 115 a věty 322 (a) platí

$$g^T h \leq F'(x,h) \leq \frac{F(x+th)-F(x)}{t}$$

pro $0 < t \leq 1$ a $h \in \mathcal{B}(0,\varepsilon)$. Zvolíme-li $t = 1$, dostaneme (1122). Platí-li (1122), platí též

$$g^T h \leq \frac{F(x+th)-F(x)}{t}$$

pro $0 < t \leq 1$ a $h \in \mathcal{B}(0,\varepsilon)$. Provedeme-li limitní přechod tak jako v definici 114, dostaneme $g^T h \leq F'(x,h)$ $\forall h \in \mathcal{B}(0,\varepsilon)$, což podle definice 115 dává $g \in \partial F(x)$. □

Poznámka 419. Porovnáme-li větu 323 (b) s poznámkou 406, vidíme, že směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$F'(x,h) = \delta_{\partial F(x)}(h).$$

Je-li funkce F diferencovatelná, obsahuje subdiferenciál jediný prvek, gradient funkce F.

Věta 324. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$ a diferencovatelná v bodě $x \in R^n$. Pak platí $\partial F(x) = \{\nabla F(x)\}$.*

Důkaz Je-li F diferencovatelná v bodě $x \in R^n$, můžeme psát

$$F'(x,h) = (\nabla F(x))^T h.$$

Nechť $g \in \partial F(x)$. Pak podle definice 115 platí

$$(\nabla F(x))^T h \geq g^T h \quad \forall h \in R^n.$$

Pro žádný vektor $h \in R^n$ nemůže nastat případ, kdy $(\nabla F(x))^T h > g^T h$, neboť v tomto případě by muselo platit $(\nabla F(x))^T(-h) < g^T(-h)$, což je podle definice subdiferenciálu nemožné. Tedy $(\nabla F(x))^T h = g^T h$ $\forall h \in R^n$, neboli $g = \nabla F(x)$. □

Poznámka 420. Nechť $F(x) = \|x\|$ je eukleidovská norma. Tato funkce je konvexní a spojitě diferencovatelná v $R^n \setminus \{0\}$ a platí

$$\begin{aligned} \partial\|x\| &= \{g \in R^n : \ \|g\| \leq 1\}, & x = 0, \\ \partial\|x\| &= \left\{\frac{x}{\|x\|}\right\}, & x \neq 0. \end{aligned}$$

Nechť $x = 0$. Podle definice 115 je $g \in \partial\|0\|$ právě tehdy, když pro libovolný vektor $h \in R^n$ platí

$$g^T h \leq F'(0,h) = \lim_{t\downarrow 0} \frac{\|th\|}{t} = \|h\|,$$

Porovnáme-li tuto nerovnost se Schwarzovou nerovností $g^T h \leq \|g\|\|h\|$, vidíme, že $g \in \partial\|0\|$ právě tehdy když $\|g\| \leq 1$. Pro $x \neq 0$ lze použít lemma 64.

Poznámka 421. Nechť $F(x) = \|x\|_P$ je nějaká obecná (primární) norma. Tato funkce nemusí být spojitě diferencovatelná v $R^n \setminus \{0\}$ (norma $\|x\|_1$ není diferencovatelné v bodech kde je některá složka vektoru x nulová a norma $\|x\|_\infty$ v bodech kde se absolutní hodnota více složek vektoru x rovná $\|x\|_\infty$). Pak platí

$$\partial\|x\|_P = \{g \in R^n : \ \|g\|_D \leq 1, \ \ g^T x = \|x\|_P\}. \tag{1124}$$

Stejným způsobem jako v poznámce 420 lze ukázat, že $g \in \partial\|0\|_P$ právě tehdy, když pro libovolný vektor $h \in R^n$ platí $g^T h \leq \|h\|_P$, neboli $g^T h \leq 1$, pokud $\|h\|_P = 1$. Použijeme-li duální normu

$$\|g\|_D = \max_{\|h\|_P = 1} g^T h,$$

vidíme, že $g \in \partial\|0\|_P$ právě tehdy když $\|g\|_D \leq 1$. Pokud $x = 0$ je rovnost $g^T x = \|x\|_P$ splněna pro libovolný vektor $g \in R^n$. Nechť nyní $x \neq 0$ a $g \in \partial\|x\|_P$. Pak podle věty 323 (d) platí

$$\begin{aligned} 2\|x\|_P = \|x + x\|_P &\geq \|x\|_P + g^T x \\ 0 = \|x - x\|_P &\geq \|x\|_P - g^T x \end{aligned}$$

což dohromady dává $g^T x = \|x\|_P$. Z definice duální normy plyne, že $g^T x \leq \|g\|_D \|x\|_P$, takže $\|g\|_D \geq 1$. Ukážeme, že $\|g\|_D = 1$. Pokud by platilo $\|g\|_D > 1$, existoval by podle definice duální normy vektor $h \in R^n$ takový, že $\|h\|_P = 1$ a $g^T h = \|g\|_D \|h\|_P > 1$ a dostali bychom nerovnost

$$\|x\|_P + 1 = \|x\|_P + \|h\|_P \geq \|x + h\|_P \geq \|x\|_P + g^T h > \|x\|_P + 1,$$

což je spor. Platí tedy

$$\partial\|x\|_P \in \{g \in R^n : \ \|g\|_D \leq 1, \ \ g^T x = \|x\|_P\}.$$

Nechť naopak $x \neq 0$, $g \in \mathcal{G}$, kde $\mathcal{G}$ je množina uvedená na pravé straně rovnosti (1124), a $z \in R^n$. Jelikož $\|g\|_D \leq 1$ a $g^T x = \|x\|_P$, dostaneme použitím definice duální normy

$$\|z\|_P \geq \|z\|_P \|g\|_D \geq g^T z = g^T x + g^T (z - x) = \|x\|_P + g^T (z - x),$$

což podle věty 323 (d) dává $g \in \partial\|x\|_P$. Použijeme-li (1124) dostaneme

$$\partial\|x\|_1 = \{g \in R^n : \ \|g\|_\infty \leq 1, \ \ g^T x = \|x\|_1\}, \qquad \partial\|x\|_\infty = \{g \in R^n : \ \|g\|_1 \leq 1, \ \ g^T x = \|x\|_\infty\},$$

Subdiferenciál je užitečný nástroj, který lze použít k formulaci podmínek optimality pro nehladké konvexní funkce.

Věta 325. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak F má v bodě x lokální minimum právě tehdy, když $0 \in \partial F(x)$.*

Důkaz Podle věty 322 (a) má funkce $F : R^n \to R$ v bodě $x \in R^n$ lokální minimum právě tehdy, když $F'(x, h) \geq 0$, $\forall h \in R^n$. Podle definice 115 tedy platí $0 \in \partial F(x)$. Jestliže $0 \in \partial F(x)$, existuje podle věty 323 (d) číslo $\varepsilon > 0$ takové, že $F(x + h) - F(x) \geq 0 \ \forall h \in \mathcal{B}(x, \varepsilon)$, takže F má v bodě x lokální minimum. □

Věta 326. *Nechť $\mathcal{C} \in R^n$ je uzavřená konvexní množina, $x \in \mathcal{C}$ a funkce $F : R^n \to R$ je konvexní v okolí bodu x. Pak F má v bodě x lokální minimum na $\mathcal{C}$ právě tehdy, když $0 \in \partial F(x) + \mathcal{N}_\mathcal{C}(x)$.*

Důkaz (a) Jestliže $0 \in \partial F(x) + \mathcal{N}_\mathcal{C}(x)$, existuje vektor $g^* \in \partial F(x)$ takový že $-g^* \in \mathcal{N}_\mathcal{C}(x)$. Pro tento vektor platí $-(g^*)^T h \leq 0 \ \forall h \in \mathcal{T}_\mathcal{C}(x)$ a podle věty 323 (b) lze psát

$$F'(x, h) = \max\left\{g^T h : g \in \partial F(x)\right\} \geq (g^*)^T h \geq 0, \qquad \forall h \in \mathcal{T}_\mathcal{C}(x).$$

Podle věty 322 (a) existuje číslo $\varepsilon > 0$ takové, že $F(x+th) \geq F(x) + tF'(x,h) \geq F(x)$, pokud $t > 0$, $h \in \mathcal{T}_{\mathcal{C}}(x)$ a $x + th \in \mathcal{B}(x,\varepsilon)$, takže bod x je lokálním minimem funkce F na $\mathcal{T}_{\mathcal{C}}(x) \cap \mathcal{B}(x,\varepsilon)$ a tudíž i na $\mathcal{C} \cap \mathcal{B}(x,\varepsilon) \subset \mathcal{T}_{\mathcal{C}}(x) \cap \mathcal{B}(x,\varepsilon)$ (věta 312).

(b) Nutnost podmínky $0 \in \partial F(x) + \mathcal{N}_{\mathcal{C}}(x)$ plyne bezprostředně z obecnější věty 340, neboť konvexní funkce je regulární (ve smyslu definice 119). □

Spojitě diferencovatelné konvexní funkce lze vyšetřovat pomocí prostředků lineární algebry, neboť konvexita implikuje pozitivní semidefinitnost Hessovy matice a ryzí konvexita je důsledkem pozitivní definitnosti Hessovy matice.

Věta 327. *Nechť funkce $F : R^n \to R$ je spojitě diferencovatelná v otevřené konvexní množině $\mathcal{C} \subset R^n$. Pak*

(a) Funkce F je konvexní v $\mathcal{C}$ právě tehdy, když $F(y) - F(x) \geq (\nabla F(x))^T(y-x)$, pokud $x \in \mathcal{C}$ a $y \in \mathcal{C}$.
(b) Funkce F je ryze konvexní v $\mathcal{C}$ právě tehdy, když $F(y) - F(x) > (\nabla F(x))^T(y-x)$, pokud $x \in \mathcal{C}$, $y \in \mathcal{C}$ a $y \neq x$.

Důkaz (a) Je-li funkce F konvexní v $\mathcal{C}$ a platí-li $x \in \mathcal{C}$ a $y \in \mathcal{C}$, můžeme psát $F(x + t(y-x)) \leq F(x) + t(F(y) - F(x))$ pro $0 \leq t \leq 1$ (poznámka 414), takže

$$(\nabla F(x))^T(y-x) = F'(x, y-x) = \lim_{t \downarrow 0} \frac{F(x + t(y-x)) - F(x)}{t} \leq F(y) - F(x). \tag{1125}$$

Nechť naopak

$$F(x) - F(z) \geq (\nabla F(z))^T(x-z), \qquad F(y) - F(z) \geq (\nabla F(z))^T(y-z),$$

pokud $x \in \mathcal{C}$, $y \in \mathcal{C}$ a $z = \lambda_1 x + \lambda_2 y$, kde $\lambda_1 \geq 0$, $\lambda_2 \geq 0$ a $\lambda_1 + \lambda_2 = 1$ (takže $z \in \mathcal{C}$). Vynásobením uvedených nerovností čísly λ_1, λ_2 a jejich sečtením dostaneme

$$\lambda_1 F(x) + \lambda_2 F(y) - F(z) \geq (\nabla F(z))^T(\lambda_1 x + \lambda_2 y - z) = 0,$$

což dává $F(z) = F(\lambda_1 x + \lambda_2 y) \leq \lambda_1 F(x) + \lambda_2 F(y)$. Protože vektory $x \in \mathcal{C}$, $y \in \mathcal{C}$ lze vybrat libovolně, je funkce F konvexní v $\mathcal{C}$.

(b) Důkaz tvrzení (b) je v podstatě stejný jako důkaz tvrzení (a), používají se však ostré nerovnosti. Platí $F(x + t(y-x) < F(x) + t(F(y) - F(x))$ pro $0 < t < 1$ a jelikož podle věty 322 je diferenční podíl v (1125) neklesající, dostaneme $(\nabla F(x))^T(y-x) < F(y) - F(x)$. Ve zbylé části důkazu volíme body $x \in \mathcal{C}$, $y \in \mathcal{C}$ a $z \in \mathcal{C}$ tak, že $y \neq x$ a $\lambda_1 > 0$, $\lambda_2 > 0$. □

Věta 328. *Nechť funkce $F : R^n \to R$ je dvakrát spojitě diferencovatelná v otevřené konvexní množině $\mathcal{C} \subset R^n$. Pak*

(a) Funkce F je konvexní v $\mathcal{C}$ právě tehdy, je-li její Hessova matice pozitivně semidefinitní v $\mathcal{C}$.
(b) Funkce F je ryze konvexní v $\mathcal{C}$, je-li její Hessova matice pozitivně definitní v $\mathcal{C}$.

Důkaz (a) Nechť funkce F je konvexní v $\mathcal{C}$ ale její Hessova matice není pozitivně semidefinitní v $\mathcal{C}$. Pak existuje bod $x \in \mathcal{C}$ a vektor $v \in R^n$, $\|v\| = 1$, tak, že $v^T \nabla^2 F(x) v = \underline{\lambda}(\nabla^2 F(x)) < 0$. Ze spojitosti Hessovy matice a z otevřenosti množiny $\mathcal{C}$ plyne existence čísla $\overline{\alpha} > 0$ takového, že $y = x + \overline{\alpha} v \in \mathcal{C}$ a $v^T \nabla^2 F(x + \alpha v) v \leq \underline{\lambda}(\nabla^2 F(x))/2 < 0 \ \forall 0 \leq \alpha \leq \overline{\alpha}$. Pak podle věty o střední hodnotě existuje číslo $0 \leq \tilde{\alpha} \leq \overline{\alpha}$ takové, že

$$\begin{aligned} F(y) - F(x) &= (\nabla F(x))^T(y-x) + \frac{1}{2}(y-x)^T \nabla^2 F(x + \tilde{\alpha} v)(y-x) \\ &\leq (\nabla F(x))^T(y-x) + \frac{1}{4}\underline{\lambda}(\nabla^2 F(x)) < (\nabla F(x))^T(y-x), \end{aligned}$$

takže podle věty 327 není funkce F konvexní v $\mathcal{C}$. Nechť naopak Hessova matice funkce F je pozitivně semidefinitní v $\mathcal{C}$ a $x \in \mathcal{C}$. Podle věty o střední hodnotě pro libovolný vektor $y \in \mathcal{C}$ platí

$$F(y) - F(x) = (\nabla F(x))^T (y - x) + \frac{1}{2}(y - x)^T \nabla^2 F(z)(y - x), \tag{1126}$$

kde $z = x + \lambda(y - x)$ a $0 < \lambda < 1$ (takže $z \in \mathcal{C}$). Jelikož matice $\nabla^2 F(z)$ je pozitivně semidefinitní (takže $(y - x)^T \nabla^2 F(z)(y - x) \geq 0$), můžeme psát $F(y) - F(x) \geq (\nabla F(x))^T (y - x)$, což spolu s tvrzením (a) věty 327 implikuje konvexitu funkce F.

(b) Je-li Hessova matice funkce F pozitivně definitní v $\mathcal{C}$, platí $(y - x)^T \nabla^2 F(z)(y - x) > 0$, pokud $z \in \mathcal{C}$, takže po dosazení do (1126) dostaneme $F(y) - F(x) > (\nabla F(x))^T (y - x)$, což spolu s tvrzením (b) věty 327 implikuje ryzí konvexitu funkce F. □

Poznámka 422. Tvrzení (b) věty 328 nelze obrátit. Funkce $F : R \to R$, zadaná předpisem $F(x) = x^4$, je ryze konvexní podle věty 327, neboť pro $x \in R$, $y \in R$ a $y \neq x$ platí

$$F(y) - F(x) - F'(x)(y - x) = y^4 - x^4 - 4x^3(y - x) = (y - x)^2((y - x)^2 + 2x^2) > 0$$

(poslední rovnost lze ověřit roznásobením). Jelikož $F''(x) = 12x^2$, takže $F''(0) = 0$, není Hessova matice funkce F pozitivně definitní v bodě $x = 0$. Uvedený příklad zároveň ukazuje, že i u jednoduchých funkcí může být ověření konvexity podle definice 112 velmi komplikované, zatímco určení pozitivní semidefinitnosti nebo pozitivní definitnosti Hessovy matice nečiní potíže.

Některé další vlastnosti subdiferenciálů konvexních funkcí budou v obecnější podobě uvedeny v následujícím oddílu (věta 341 a věta 342). Ukážeme ještě, jak lze vlastnosti konvexních funkcí použít k vyšetřování konvexních množin.

Věta 329. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Pak platí*

$$\mathcal{T}_{\mathcal{C}}(x) = \{y \in R^n : d'_{\mathcal{C}}(x, y) = 0\}$$

($d'_{\mathcal{C}}(x, y)$ je směrová derivace funkce $d_{\mathcal{C}}(x)$ ve směru $y \in R^n$).

Důkaz Jelikož množina $\mathcal{C}$ je konvexní, je podle věty 288 funkce $d_{\mathcal{C}}(x)$ konvexní a podle věty 292 existuje v každém bodě $x \in \mathcal{C}$ její směrová derivace $d'_{\mathcal{C}}(x, y)$, $y \in R^n$.

(a) Označme $\mathcal{K} = \{y \in R^n : d'_{\mathcal{C}}(x, y) = 0\}$. Předpokládejme nejprve, že $y \in \mathcal{T}_{\mathcal{C}}(x)$. Pak existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in \mathcal{C}$. Jelikož $d'_{\mathcal{C}}(x, y) \geq 0$ (plyne to z toho, že $d_{\mathcal{C}}(x) = 0$ a $d_{\mathcal{C}}(z) \geq 0$ $\forall z \in R^n$), stačí dokázat, že $d'_{\mathcal{C}}(x, y) \leq 0$. Platí

$$\begin{aligned} d'_{\mathcal{C}}(x, y) &= \lim_{t \downarrow 0} \frac{d_{\mathcal{C}}(x + ty) - d_{\mathcal{C}}(x)}{t} = \lim_{i \to \infty} \frac{d_{\mathcal{C}}(x + t_i y) - d_{\mathcal{C}}(x)}{t_i} = \lim_{i \to \infty} \frac{\min_{z \in \mathcal{C}} \|x + t_i y - z\|}{t_i} \\ &\leq \lim_{i \to \infty} \frac{\min_{z \in \mathcal{C}} \|x + t_i y_i - z\| + t_i \|y - y_i\|}{t_i}. \end{aligned}$$

Ale $\min_{z \in \mathcal{C}} \|x + t_i y_i - z\| = 0$, neboť $x + t_i y_i \in \mathcal{C}$. Můžeme tedy psát

$$d'_{\mathcal{C}}(x, y) \leq \lim_{i \to \infty} \frac{t_i \|y - y_i\|}{t_i} = \lim_{i \to \infty} \|y - y_i\|$$

a jelikož $y_i \to y$, dostaneme $d'_{\mathcal{C}}(x, y) \leq 0$. Tedy $d'_{\mathcal{C}}(x, y) = 0$, čili $y \in \mathcal{K}$, což dává $\mathcal{T}_{\mathcal{C}}(x) \subset \mathcal{K}$.

(b) Nechť $y \in \mathcal{K}$ a $t_i \downarrow 0$. Z definice množiny $\mathcal{K}$ plyne, že

$$d'_{\mathcal{C}}(x, y) = \lim_{i \to \infty} \frac{d_{\mathcal{C}}(x + t_i y)}{t_i} = 0.$$

Nechť body $z_i \in \mathcal{C}$, $i \in N$, jsou zvoleny tak, že

$$\|x + t_i y - z_i\| \leq d_{\mathcal{C}}(x + t_i y) + \frac{t_i}{i}$$

(což je možné vzhledem k definici vzdálenosti $d_{\mathcal{C}}(x + t_i y)$). Položme $y_i = (z_i - x)/t_i$, $i \in N$. Pak platí

$$x + t_i y_i = x + (z_i - x) = z_i \in \mathcal{C}$$

a

$$\|y - y_i\| = \|y - \frac{z_i - x}{t_i}\| = \frac{1}{t_i}\|x + t_i y - z_i\| \leq \frac{d_{\mathcal{C}}(x + t_i y)}{t_i} + \frac{1}{i},$$

takže

$$\lim_{i\to\infty} \|y - y_i\| = d'_{\mathcal{C}}(x, y) + \lim_{i\to\infty} \frac{1}{i} = 0.$$

Tedy $y_i \to y$, $t_i \downarrow 0$ a $x + t_i y_i \in \mathcal{C}$, takže $y \in \mathcal{T}_{\mathcal{C}}(x)$, což dává $\mathcal{K} \subset \mathcal{T}_{\mathcal{C}}(x)$. □

Věta 330. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina. Pak funkce $d_{\mathcal{C}}(x)$ je konvexní a platí*

$$\begin{aligned} \partial d_{\mathcal{C}}(x) &= \left\{ \frac{x - P(x)}{\|x - P(x)\|} \right\}, \quad x \notin \mathcal{C}, \\ \partial d_{\mathcal{C}}(x) &= \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)}, \quad x \in \mathcal{C}. \end{aligned}$$

Důkaz Podle věty 288 je funkce $d_{\mathcal{C}}$ konvexní a ke každému bodu $x \in R^n$ existuje právě jeden bod $x_{\mathcal{C}} \in \mathcal{C}$ takový, že $d_{\mathcal{C}}(x) = \|x - x_{\mathcal{C}}\|$. Nechť $z \in R^n$. Pak $d_{\mathcal{C}}(z) \leq \|z - x_{\mathcal{C}}\|$ a pokud $g \in \partial d_{\mathcal{C}}(x)$, platí

$$g^T(z - x) \leq d_{\mathcal{C}}(z) - d_{\mathcal{C}}(x) \leq \|z - x_{\mathcal{C}}\| - \|x - x_{\mathcal{C}}\|,$$

takže $g \in \partial\|x - x_{\mathcal{C}}\|$, což dává $\partial d_{\mathcal{C}}(x) \subset \partial\|x - x_{\mathcal{C}}\|$ (bod $x_{\mathcal{C}}$ považujeme za pevný).

(a) Pokud $x \notin \mathcal{C}$, obsahuje množina $\partial\|x - x_{\mathcal{C}}\|$ jediný prvek $g = (x - P(x))/\|x - P(x)\|$ (poznámka 420).

(b) Nechť $x \in \mathcal{C}$ a $g \in \partial d_{\mathcal{C}}(x)$. Pak $d_{\mathcal{C}}(x) = 0$ a pro libovolný vektor $z \in \mathcal{C}$ platí

$$0 = d_{\mathcal{C}}(z) - d_{\mathcal{C}}(x) \geq g^T(z - x),$$

neboli $g \in N_{\mathcal{C}}(x)$. Již jsme dokázali, že $g \in \partial\|x - x_{\mathcal{C}}\|$, což podle poznámky 420 dává $g \in \overline{\mathcal{B}(0,1)}$. Lze tedy psát $\partial d_{\mathcal{C}}(x) \subset \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)}$. Nechť nyní $g \in \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)}$. Jelikož $\|g\| \leq 1$ a $d_{\mathcal{C}}(x) = 0$, můžeme pro libovolný vektor $z \in R^n$ psát

$$d_{\mathcal{C}}(z) - d_{\mathcal{C}}(x) = \min_{y\in\mathcal{C}} \|z - y\| \geq \min_{y\in\mathcal{C}} \|z - y\|\|g\| \geq \min_{y\in\mathcal{C}} g^T(z - y) = g^T(z - x) + \min_{y\in\mathcal{C}} g^T(x - y) \geq g^T(z - x)$$

(neboť $g \in \mathcal{N}_{\mathcal{C}}(x)$ a tudíž $g^T(y - x) \leq 0 \ \forall y \in \mathcal{C}$), takže podle věty 323 (d) platí $g \in \partial d_{\mathcal{C}}(x)$. □

Důsledek 37. *Nechť $\mathcal{C} \subset R^n$ je uzavřená konvexní množina a $x \in \mathcal{C}$. Pak platí*

$$\mathcal{N}_{\mathcal{C}}(x) = \operatorname{cone} \partial d_{\mathcal{C}}(x) = \bigcup_{\lambda\geq 0} \lambda \partial d_{\mathcal{C}}(x).$$

Důkaz Jelikož $\partial d_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)}$, platí $0 \in \mathcal{N}_{\mathcal{C}}(x)$ i $0 \in \operatorname{cone} \partial d_{\mathcal{C}}(x)$. Nechť $g \neq 0$ a $g \in \mathcal{N}_{\mathcal{C}}(x)$. Pak $g/\|g\| \in \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)} = \partial d_{\mathcal{C}}(x)$, takže $g \in \operatorname{cone} \partial d_{\mathcal{C}}(x)$. Nechť $g \neq 0$ a $g \in \operatorname{cone} \partial d_{\mathcal{C}}(x)$. Pak $g = \|g\|\tilde{g}$, kde $\tilde{g} \in \mathcal{N}_{\mathcal{C}}(x)$ (neboť $\|\tilde{g}\| = 1$ a $\partial d_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{C}}(x) \cap \overline{\mathcal{B}(0,1)}$), takže $g \in \mathcal{N}_{\mathcal{C}}(x)$. □

15.4 Lipschitzovské funkce

Definice 116. *Řekneme, že funkce $F : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ (s konstantou L), existuje-li číslo $\varepsilon > 0$ takové, že platí*

$$|F(x_2) - F(x_1)| \leq L\|x_2 - x_1\|, \tag{1127}$$

pokud $x_1 \in \mathcal{B}(x, \varepsilon)$ a $x_2 \in \mathcal{B}(x, \varepsilon)$. Řekneme, že funkce $F : R^n \to R$ je lokálně lipschitzovská v otevřené množině $\Omega \subset R^n$, je-li lipschitzovská v okolí každého bodu $x \in \Omega$.

Definice 117. *Zobecněnou (Clarkovu) směrovou derivaci funkce $F : R^n \to R$ v bodě $x \in R^n$ ve směru $h \in R^n$ definujeme předpisem*

$$F^0(x, h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + th) - F(y)}{t}. \tag{1128}$$

Poznámka 423. Je-li $F^0(x, h)$ zobecněnou směrovou derivací funkce F ve smyslu Definice 117, existují posloupnosti $x_i \to x$ a $t_i \downarrow 0$ tak, že

$$F^0(x, h) = \lim_{i \to \infty} \frac{F(x_i + t_i h) - F(x_i)}{t_i}$$

Věta 331. *Nechť $F : R^n \to R$ je lipschitzovská s konstantou L v okolí bodu $x \in R^n$. Pak:*

(a) Funkce $F^0(x, \cdot) : R^n \to R$ je positivně homogenní, subaditivní a lipschitzovská s konstantou L.
(b) Funkce $F^0(\cdot, \cdot) : R^n \times R^n \to R$ je shora polospojitá, neboli

$$\limsup_{i \to \infty} F^0(x_i, h_i) \leq F^0(x, h),$$

kdykoliv $x_i \to x$ a $h_i \to h$.
(c) Platí $F^0(x, -h) = (-F)^0(x, h)$ $\forall h \in R^n$.

Důkaz (a) Nechť $\lambda > 0$. Pak platí

$$F^0(x, \lambda h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + t\lambda h) - F(y)}{t} = \lambda \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + t\lambda h) - F(y)}{\lambda t} = \lambda F^0(y, h),$$

takže $F^0(x, \cdot)$ je positivně homogenní. Dále platí

$$\begin{aligned}
F^0(x, h_1 + h_2) &= \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + t(h_1 + h_2)) - F(y)}{t} \\
&= \limsup_{\substack{y \to x \\ t \downarrow 0}} \left(\frac{F(y + t(h_1 + h_2)) - F(y + th_1)}{t} + \frac{F(y + th_1) - F(y)}{t} \right) \\
&\leq \limsup_{\substack{y' \to x \\ t \downarrow 0}} \frac{F(y' + th_2) - F(y')}{t} + \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + th_1) - F(y)}{t} \\
&= F^0(x, h_2) + F^0(x, h_1),
\end{aligned}$$

kde $y' = y + th_1 \to x$, takže $F^0(x, h)$ je subaditivní. Jelikož F je lipschitzovská s konstantou L v okolí bodu x, platí v tomto okolí

$$F(y+th_2) \le F(y+th_1) + Lt\|h_2 - h_1\|$$

(viz (1127)), takže

$$\begin{aligned} F^0(x,h_2) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F(y+th_2)-F(y)}{t} \\ &\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F(y+th_1)-F(y)}{t} + L\|h_2-h_1\| \\ &= F^0(x,h_1) + L\|h_2-h_1\|, \end{aligned}$$

neboli

$$F^0(x,h_2) - F^0(x,h_1) \le L\|h_2-h_1\|.$$

Protože nezáleží na pořadí vektorů h_1 a h_2, platí

$$|F^0(x,h_2) - F^0(x,h_1)| \le L\|h_2-h_1\|.$$

Funkce $F^0(x,\cdot)$ je tedy lipschitzovská s konstantou L.

(b) Nechť $x_i \to x$ a $h_i \to h$. Z definice horní limity (limes superior) existují posloupnosti $y_i \to x$ a $t_i \downarrow 0$ takové, že

$$\begin{aligned} F^0(x_i,h_i) &\le \frac{F(y_i+t_ih_i)-F(y_i)}{t_i} + \frac{1}{i} \\ &= \frac{F(y_i+t_ih)-F(y_i)}{t_i} + \frac{F(y_i+t_ih_i)-F(y_i+t_ih)}{t_i} + \frac{1}{i}. \end{aligned}$$

Z lipschitzovské spojitosti funkce F plyne

$$\left\| \frac{F(y_i+t_ih_i)-F(y_i+t_ih)}{t_i} \right\| \le L\|h_i-h\|$$

pro dostatečně velké indexy i, takže

$$\limsup_{i\to\infty} F^0(x_i,h_i) \le F^0(x,h) + \lim_{i\to\infty}\left(L\|h_i-h\| + \frac{1}{i}\right) = F^0(x,h).$$

(c) Zřejmě

$$\begin{aligned} F^0(x,-h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F(y-th)-F(y)}{t} \\ &= \limsup_{\substack{z\to x\\ t\downarrow 0}} \frac{(-F)(z+th)-(-F)(z)}{t} = (-F)^0(x,h) \end{aligned}$$

(zde $z = y - th$). □

Poznámka 424. Podle definice 117 platí $F^0(x,0) = 0$, takže podle věty 331 (a) dostaneme

$$|F^0(x,h)| = |F^0(x,h) - F^0(x,0)| \le L\|h\|.$$

Definice 118. *Nechť funkce $F : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$. Pak množinu*

$$\partial F(x) = \{g \in R^n : F^0(x,h) \geq g^T h \quad \forall h \in R^n\}$$

nazveme subdiferenciálem funkce F v bodě x. Elementy $g \in \partial F(x)$ budeme nazývat subgradienty funkce F v bodě x.

Věta 332. *Nechť funkce $F : R^n \to R$ je lipschitzovská s konstantou L v okolí bodu $x \in R^n$. Pak:*

(a) Subdiferenciál $\partial F(x)$ je neprázdná konvexní kompaktní množina taková, že $\|g\| \leq L \ \forall g \in \partial F(x)$.
(b) Pro libovolný vektor $h \in R^n$ platí

$$F^0(x,h) = \max\left\{g^T h : g \in \partial F(x)\right\}.$$

(c) Jestliže $x_i \to x$, $g_i \in \partial F(x_i)$ a $g_i \to g$, pak $g \in \partial F(x)$ (polospojitost shora).
(d) Platí $\partial(-F)(x) = -\partial F(x)$.

Důkaz (a) Podle věty 331 (a) je funkce $F^0(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (důsledek 32) existuje vektor $g \in R^n$ takový, že

$$g^T h \leq F^0(x,h) \quad \forall h \in R^n,$$

takže subdiferenciál $\partial F(x)$ je neprázdný. Nechť $g_1 \in \partial F(x)$, $g_2 \in \partial F(x)$ a $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, $\lambda_1 + \lambda_2 = 1$. Pak pro libovolný vektor $h \in R^n$ platí

$$(\lambda_1 g_1 + \lambda_2 g_2)^T h = \lambda_1 g_1^T h + \lambda_2 g_2^T h \leq \lambda_1 F^0(x,h) + \lambda_2 F^0(x,h) = F^0(x,h),$$

takže subdiferenciál $\partial F(x)$ je konvexní. Nechť $g \in \partial F(x)$. Pak podle definice 118 a poznámky 424 platí

$$\|g\|^2 = g^T g \leq F^0(x,g) \leq L\|g\|,$$

čili $\|g\| \leq L$, takže subdiferenciál $\partial F(x)$ je omezený. Nechť $g_i \in \partial F(x)$ a $g_i \to g$. Pak platí

$$g^T h = \lim_{i\to\infty} g_i^T h \leq F^0(x,h),$$

takže $g \in \partial F(x)$ a subdiferenciál $\partial F(x)$ je uzavřený.

(b) Podle definice 118 platí

$$F^0(x,h) \geq \max\left\{g^T h : g \in \partial F(x)\right\}.$$

Předpokládejme, že pro nějaký vektor $\overline{h} \in R^n$ platí

$$F^0(x,\overline{h}) > \max\left\{g^T \overline{h} : g \in \partial F(x)\right\}. \tag{1129}$$

Uvažujme lineární formu $l(\lambda\overline{h}) = \lambda F^0(x,\overline{h})$ definovanou na jednorozměrném podprostoru $\{\lambda\overline{h} : \lambda \in R\} \subset R^n$. Jelikož $F^0(x,\cdot)$ je positivně homogenní a subaditivní, existuje podle Hahn-Banachovy věty vektor $\overline{g} \in R^n$ takový, že $F^0(x,h) \geq \overline{g}^T h \ \forall h \in R^n$ a $\overline{g}^T(\lambda\overline{h}) = l(\lambda\overline{h}) = \lambda F^0(x,\overline{h})$. Tedy $\overline{g} \in \partial F(x)$ a pro $\lambda = 1$ dostaneme $F^0(x,\overline{h}) = \overline{g}^T\overline{h}$, což je ve sporu s předpokladem (1129).

(c) Nechť $x_i \to x$ a $g_i \to g$, kde $g_i \in \partial F(x_i)$. Pak pro libovolný vektor $h \in R^n$ platí

$$g^T h = \lim_{i\to\infty} g_i^T h \leq \limsup_{i\to\infty} F^0(x_i,h).$$

Podle věty 331 (b) je funkce $F^0(\cdot,\cdot)$ shora polospojitá, takže $g^T h \leq F^0(x,h)$.

(d) Vztah $g \in \partial(-F)(x)$ platí podle definice 118 právě tehdy, jestliže $(-F)^0(x,h) \geq g^T h \ \forall h \in R^n$, což je podle věty 331 (c) ekvivalentní $F^0(x,-h) \geq g^T h \ \forall h \in R^n$, což podle definice 118 znamená $-g \in \partial F(x)$. Tedy $\partial(-F)(x) = -\partial F(x)$. □

Poznámka 425. Porovnáme-li větu 332 (b) s poznámkou 406 vidíme, že zobecněná směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$F^0(x,h) = \delta_{\partial F(x)}(h).$$

Věta 333. *Nechť funkce $F : R^n \to R$ je spojitě diferencovatelná v okolí bodu $x \in R^n$. Pak F je lipschitzovská v okolí bodu x a platí*

(a) $F^0(x,h) = F'(x,h) = (\nabla F(x))^T h \quad \forall h \in R^n$.
(b) $\partial F(x) = \{\nabla F(x)\}$.

Důkaz Je-li F spojitě diferencovatelná v okolí bodu $x \in R^n$, pak gradient $\nabla F(x)$ existuje a je omezený v okolí bodu x. Existují tedy čísla $\varepsilon > 0$ a $L > 0$ tak, že $\|\nabla F(y)\| \leq L \ \forall y \in \mathcal{B}(x,\varepsilon)$. Nechť $x_1 \in \mathcal{B}(x,\varepsilon)$ a $x_2 \in \mathcal{B}(x,\varepsilon)$. Pak podle věty o střední hodnotě platí

$$F(x_2) - F(x_1) = (\nabla F(y))^T (x_2 - x_1),$$

kde $y \in (x_1, x_2) \subset \mathcal{B}(x,\varepsilon)$. Můžeme tedy psát

$$|F(x_2) - F(x_1)| \leq \|\nabla F(y)\| \, \|x_2 - x_1\| \leq L\|x_2 - x_1\|,$$

takže funkce F je lipschitzovská v $\mathcal{B}(x,\varepsilon)$.

(a) Ze spojité diferencovatelnosti funkce F v bodě x plyne, že $F'(y,h) = (\nabla F(y))^T h$ pokud $y \in \mathcal{B}(x,\varepsilon)$. Předpokládejme, že $x_i \in \mathcal{B}(x,\varepsilon)$ a $x_i \to x$. Pak pro $h \in R^n$ platí

$$\begin{aligned} F'(x,h) &= (\nabla F(x))^T h = \lim_{x_i \to x} (\nabla F(x_i))^T h \\ &= \lim_{x_i \to x} F'(x_i,h) = \lim_{\substack{x_i \to x \\ t \downarrow 0}} \frac{F(x_i + th) - F(x_i)}{t} \\ &= \limsup_{\substack{x_i \to x \\ t \downarrow 0}} \frac{F(x_i + th) - F(x_i)}{t} = F^0(x,h) \end{aligned}$$

(existuje-li limita, rovná se horní limitě). Platí tedy $F^0(x,h) = (\nabla F(x))^T h = F'(x,h) \ \forall h \in R^n$.

(b) Z (a) plyne inkluze $\nabla F(x) \in \partial F(x)$. Předpokládejme nyní, že $g \in \partial F(x)$ a $g \neq \nabla F(x)$. Pak pro nějaký vektor $h \in R^n$ musí platit $F^0(x,h) = (\nabla F(x))^T h > g^T h$. Z definice $\partial F(x)$ však nutně plyne $F^0(x,-h) = -(\nabla F(x))^T h \geq -g^T h$, neboli (po vynásobení číslem -1) $(\nabla F(x))^T h \leq g^T h$, což je ve sporu s nerovností $(\nabla F(x))^T h > g^T h$. □

Poznámka 426. Je-li funkce $F : R^n \to R$ lipschitzovská v okolí bodu $x \in R^n$ a diferencovatelná v tomto bodě, platí $\nabla F(x) \in \partial F(x)$ (neboť $F^0(x,h) \geq F'(x,h) = (\nabla F(x))^T h \ \forall h \in R^n$). Vztah $\partial F(x) = \{\nabla F(x)\}$ lze dokázat pouze v případě spojité diferencovatelnosti.

Věta 334. *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak platí*

(a) $F^0(x,h) = F'(x,h) \quad \forall h \in R^n$.
(b) $\partial F(x) = \left\{ g \in R^n : F'(x,h) \geq g^T h \ \ \forall h \in R^n \right\}$.

Důkaz Vztah (b) plyne bezprostředně z (a) a z definice 118. Abychom dokázali (a), stačí dokázat, že $F^0(x,h) \leq F'(x,h)$, neboť obrácenou nerovnost dostaneme ihned z definice 117 (použijeme-li speciální volbu $y = x$). Nechť $x_i \to x$ a $t_i \downarrow 0$ tak, že

$$F^0(x,h) = \lim_{i\to\infty} \frac{F(x_i + t_i h) - F(x_i)}{t_i}$$

(poznámka 423). Položme $\bar{t}_i = \max\left(t_i, \sqrt{\|x_i - x\|}\right)$, takže $\|x_i - x\| \leq \bar{t}_i^2$, $t_i \leq \bar{t}_i$ a $\bar{t}_i \to 0$. Podle věty 321 je funkce F lipschitzovská (s nějakou konstantou L) v okolí bodu x (bez újmy na obecnosti budeme předpokládat, že body x_i, $x_i + \bar{t}_i h$ a $x + \bar{t}_i h$ leží v tomto okolí). Použijeme-li lemma 123 (levou nerovnost) dostaneme

$$\begin{aligned}
\frac{F(x_i + t_i h) - F(x_i)}{t_i} &\leq \frac{F(x_i + \bar{t}_i h) - F(x_i)}{\bar{t}_i} \\
&\leq \frac{F(x + \bar{t}_i h) - F(x)}{\bar{t}_i} + \frac{F(x_i + \bar{t}_i h) - F(x + \bar{t}_i h)}{\bar{t}_i} - \frac{F(x_i) - F(x)}{\bar{t}_i} \\
&\leq \frac{F(x + \bar{t}_i h) - F(x)}{\bar{t}_i} + \frac{2L\|x_i - x\|}{\bar{t}_i} \\
&\leq \frac{F(x + \bar{t}_i h) - F(x)}{\bar{t}_i} + 2L\bar{t}_i
\end{aligned}$$

pro dostatečně velké indexy i. Provedeme-li limitní přechod na obou stranách této nerovnosti, dostaneme

$$F^0(x,h) \leq F'(x,h) + \lim_{\bar{t}_i \to 0} 2L\bar{t}_i = F'(x,h)$$

□

Poznámka 427. Věta 334 říká, že v případě konvexních funkcí je zobecněná směrová derivace totožná s obyčejnou směrovou derivací a subdiferenciál podle definice 118 splývá se subdiferenciálem podle definice 115. Tato vlastnost je teoreticky i prakticky velmi výhodná, takže je účelné vyšetřovat funkce, pro než platí.

Definice 119. *Řekneme, že funkce $F : R^n \to R$ je regulární v bodě $x \in R^n$, existuje-li směrová derivace $F'(x,h)$ $\forall h \in R^n$ a platí-li $F^0(x,h) = F'(x,h)$ $\forall h \in R^n$.*

Věta 335. *Funkce spojitě diferencovatelné v okolí bodu x a funkce konvexní v okolí bodu x jsou regulární v bodě x. Dále jsou v bodě x regulární (a) nezáporné lineární kombinace regulárních funkcí a (b) bodová maxima regulárních funkcí.*

Důkaz Spojitě diferencovatelná funkce je regulární podle věty 333. Konvexní funkce je regulární podle věty 334.

(a) Stačí dokázat, že funkce $\lambda_1 F_1$ a $F_1 + F_2$ jsou regulární, jsou-li funkce F_1, F_2 regulární a platí-li $\lambda_1 \geq 0$. Nechť $h \in R^n$. Jsou-li funkce F_1, F_2 regulární a platí-li $\lambda_1 \geq 0$, pak použitím věty 322 (b) a věty 331 (a) dostaneme

$$(\lambda_1 F_1)^0(x,h) = F_1^0(x, \lambda_1 h) = F_1'(x, \lambda_1 h) = (\lambda_1 F_1)'(x,h).$$

Z definice 114 plyne, že $(F_1 + F_2)'$ existuje a platí $(F_1 + F_2)' = F_1' + F_2'$. Podle definice 117 platí $(F_1 + F_2)^0 \geq (F_1 + F_2)' = F_1' + F_2' = F_1^0 + F_2^0$. Z druhé strany

$$\begin{aligned}
(F_1+F_2)^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{(F_1+F_2)(y+th)-(F_1+F_2)(y)}{t}\\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_1(y+th)+F_2(y+th)-F_1(y)-F_2(y)}{t}\\
&\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_1(y+th)-F_1(y)}{t} + \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_2(y+th)-F_2(y)}{t}\\
&= F_1^0(x,h)+F_2^0(x,h),
\end{aligned}$$

takže

$$(F_1+F_2)' = F_1'+F_2' = F_1^0+F_2^0 \ge (F_1+F_2)^0,$$

což dohromady s předchozí nerovností dává $(F_1+F_2)^0 = (F_1+F_2)'$.

(b) Stačí dokázat, že funkce $F=\max(F_1,F_2)$ je regulární, jsou-li funkce F_1, F_2 regulární. Jestliže $F_1(x)>F_2(x)$, pak $F=F_1$, $F'=F_1'$ a $F^0=F_1^0=F_1'=F'$ (stejně se postupuje pokud $F_2(x)>F_1(x)$). Nechť tedy $F(x)=F_1(x)=F_2(x)$ a $h\in R^n$. Pak

$$\begin{aligned}
F'(x,h) &= \lim_{t\downarrow 0} \frac{F(x+th)-F(x)}{t}\\
&= \lim_{t\downarrow 0} \frac{\max(F_1(x+th),F_2(x+th))-F(x)}{t}\\
&= \max\left(\lim_{t\downarrow 0}\frac{F_1(x+th)-F_1(x)}{t}, \lim_{t\downarrow 0}\frac{F_2(x+th)-F_2(x)}{t}\right)\\
&= \max(F_1'(x,h),F_2'(x,h)),
\end{aligned}$$

takže $F'(x,h)$ existuje a platí $F'(x,h)=\max(F_1'(x,h),F_2'(x,h))$. Podle definice 117 platí $F^0(x,h)\ge F'(x,h)$. Z druhé strany

$$\begin{aligned}
F^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F(y+th)-F(y)}{t}\\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{\max(F_1(y+th),F_2(y+th))-\max(F_1(y),F_2(y))}{t}\\
&\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \max\left(\frac{F_1(y+th)-F_1(y)}{t}, \frac{F_2(y+th)-F_2(y)}{t}\right)\\
&\le \max(F_1^0(x,h),F_2^0(x,h)).
\end{aligned}$$

Platí tedy

$$F'(x,h)=\max(F_1'(x,h),F_2'(x,h))=\max(F_1^0(x,h),F_2^0(x,h))\ge F^0(x,h),$$

což dohromady s předchozí nerovností dává $F^0(x,h)=F'(x,h)$. □

Věta 336. *Nechť funkce $F_1:R^n\to R$, $F_2:R^n\to R$ jsou lipschitzovské v okolí bodu $x\in R^n$ a $\lambda_1\in R$. Pak*

(a) $\partial(\lambda_1 F_1)(x)=\lambda_1\partial F_1(x)$,

(b) $\partial(F_1 + F_2)(x) \subset \partial F_1(x) + \partial F_2(x)$.

Jsou-li funkce F_1, F_2 regulární v bodě x nebo je-li alespoň jedna z nich spojitě diferencovatelná v bodě x, nastává v (b) rovnost.

Důkaz (a) Jestliže $\lambda_1 \geq 0$, pak $(\lambda_1 F_1)^0(x,h) = \lambda_1 F_1^0(x,h)$, takže podle definice 118 platí $\partial(\lambda_1 F_1)(x) = \lambda_1 \partial F_1(x)$. V opačném případě s použitím věty 332 (d) a předchozího výsledku dostaneme

$$\partial(\lambda_1 F_1)(x) = \partial\left(-|\lambda_1| F_1\right)(x) = -\partial\left(|\lambda_1| F_1\right)(x) = -|\lambda_1| \partial F_1(x) = \lambda_1 \partial F_1(x).$$

(b) Zřejmě $(F_1+F_2)^0(x,h) \leq F_1^0(x,h)+F_2^0(x,h)\ \forall h \in R^n$ (důkaz věty 335 (a)). Použijeme-li poznámku 425 a větu 299, dostaneme

$$\delta_{\partial(F_1+F_2)(x)}(h) \leq \delta_{\partial F_1(x)}(h) + \delta_{\partial F_2(x)}(h) = \delta_{\partial F_1(x)+\partial F_2(x)}(h) \tag{1130}$$

$\forall h \in R^n$, takže podle věty 298 platí $\partial(F_1 + F_2)(x) \subset \partial F_1(x) + \partial F_2(x)$. Jsou-li funkce F_1, F_2 regulární, pak podle věty 335 (a) platí $(F_1 + F_2)^0 = (F_1 + F_2)' = F_1' + F_2' = F_1^0 + F_2^0$, takže v (1130) a tedy i v (b) nastane rovnost. Je-li funkce F_1 spojitě diferencovatelná v bodě x, pak podle definice 117 a věty o střední hodnotě ($z \in [y, y + th]$) platí

$$\begin{aligned}
(F_1 + F_2)^0(x,h) &= \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{(F_1 + F_2)(y + th) - (F_1 + F_2)(y)}{t} \\
&= \lim_{\substack{y \to x \\ t \downarrow 0}} (\nabla F_1(z))^T h + \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F_2(y + th) - F_2(y)}{t} \\
&= F_1^0(x,h) + F_2^0(x,h),
\end{aligned}$$

neboť $(\nabla F_1(z))^T h \to (\nabla F_1(x))^T h = F_1'(x,h) = F_1^0(x,h)$. □

Poznámka 428. *Indukcí se snadno dokáže, že*

$$\partial\left(\sum_{i=1}^{m} \lambda_i F_i\right)(x) \subset \sum_{i=1}^{m} \lambda_i \partial F_i(x),$$

přičemž rovnost nastane, jsou-li všechny funkce F_i regulární nebo jsou-li všechny funkce F_i až na jednu spojitě diferencovatelné.

Věta 337. *Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská v okolí bodu $x \in R^n$, který je jejím lokálním extrémem (minimem nebo maximem). Pak platí $0 \in \partial F(x)$.*

Důkaz Nechť $x \in R^n$ je lokálním minimem funkce $F : R^n \to R$. Pak nutně

$$0 \leq \limsup_{t \downarrow 0} \frac{F(x + th) - F(x)}{t} \leq F^0(x,h)$$

pro libovolný vektor $h \in R^n$, takže podle definice 118 platí $0 \in \partial F(x)$. Je-li bod x lokálním maximem funkce F, je nutně lokálním minimem funkce $-F$, takže $0 \in \partial(-F)(x)$ a podle věty 332 (d) platí $0 \in \partial F(x)$. □

Abychom mohli charakterizovat lokální extrém x lokálně lipschitzovské funkce F na uzavřené (obecně nekonvexní) množině $\mathcal{C}$, je třeba zobecnit pojem tečného kuželu množiny $\mathcal{C}$ v bodě x. Má-li funkce $d_{\mathcal{C}}(x)$ směrovou derivaci $d_{\mathcal{C}}'(x,y)$ v každém směru $y \in R^n$, platí $\mathcal{T}_{\mathcal{C}}(x) = \{y \in R^n : d_{\mathcal{C}}'(x,y) = 0\}$, neboť v částech (a) a (b) důkazu věty 329 nepředpokládáme konvexitu funkce $d_{\mathcal{C}}(x)$ (její konvexita slouží pouze k zaručení existence směrové derivace $d_{\mathcal{C}}'(x,y)$). Důsledek 37 však obecně neplatí, neboť není-li funkce $d_{\mathcal{C}}(x)$ regulární, nelze definovat subdiferenciál $\partial d_{\mathcal{C}}(x)$ tak, aby směrová derivace $d_{\mathcal{C}}'(x,y)$ byla jeho opěrnou funkcí. Můžeme však použít zobecněnou směrovou derivaci a definovat zobecněný (Clarkův) tečný kužel.

Definice 120. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Zobecněným tečným kuželem množiny $\mathcal{C}$ v bodě x nazveme množinu*

$$\tilde{\mathcal{T}}_{\mathcal{C}}(x) = \{y \in R^n : d^0_{\mathcal{C}}(x, y) = 0\}$$

Zobecněným normálovým kuželem množiny $\mathcal{C}$ v bodě x nazveme množinu $\tilde{\mathcal{N}}_{\mathcal{C}}(x) = \tilde{\mathcal{T}}^*_{\mathcal{C}}(x)$.

Poznámka 429. Je-li funkce $d_{\mathcal{C}}(x)$ regulární v bodě $x \in \mathcal{C}$, platí $\tilde{\mathcal{T}}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{C}}(x)$ a $\tilde{\mathcal{N}}_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{C}}(x)$.

Poznámka 430. Jelikož $d^0_{\mathcal{C}}(x, y) \geq d'_{\mathcal{C}}(x, y)\ \forall y \in R^n$, platí vždy $\tilde{\mathcal{T}}_{\mathcal{C}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$ (a tedy $\mathcal{N}_{\mathcal{C}}(x) \subset \tilde{\mathcal{N}}_{\mathcal{C}}(x)$). Nemusí však platit $\tilde{\mathcal{T}}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{C}}(x)$. Nechť $\mathcal{C}_1 = \{x \in R^2 : x_2 - x_1 \geq 0\}$, $\mathcal{C}_2 = \{x \in R^2 : x_2 + x_1 \geq 0\}$ a $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$. Pak se lze snadno přesvědčit, že $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{C}_1 \cup \mathcal{C}_2$, $\tilde{\mathcal{T}}_{\mathcal{C}}(x) = \mathcal{C}_1 \cap \mathcal{C}_2$ a $\mathcal{N}_{\mathcal{C}}(x) = \{0\}$, $\tilde{\mathcal{N}}_{\mathcal{C}}(x) = -(\mathcal{C}_1 \cap \mathcal{C}_2)$.

Věta 338. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Pak $\tilde{\mathcal{T}}_{\mathcal{C}}(x)$ je uzavřeným konvexním kuželem.*

Důkaz (a) Nechť $y_i \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$, $i \in N$, a $y_i \to y$. Pak $d^0_{\mathcal{C}}(x, y_i) = 0$, $i \in N$. Jelikož funkce $d_{\mathcal{C}}(x)$ je lipschitzovská v R^n (věta 288) je i funkce $d^0_{\mathcal{C}}(x, \cdot)$ lipschitzovská v R^n (věta 331 (a)) a tudíž spojitá. Platí tedy

$$d^0_{\mathcal{C}}(x, y) = \lim_{i \to \infty} d^0_{\mathcal{C}}(x, y_i) = 0,$$

což dává $y \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$.

(b) Nechť $y = \sum_{i=1}^m \lambda_i y_i$, kde $y_i \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$, $\lambda_i \geq 0$, $1 \leq i \leq m$ a $m \geq 1$. Pak podle věty 331 (a) platí

$$d^0_{\mathcal{C}}(x, y) = d^0_{\mathcal{C}}(x, \sum_{i=1}^m \lambda_i y_i) \leq \sum_{i=1}^m \lambda_i d^0_{\mathcal{C}}(x, y_i) = 0$$

a jelikož $d^0_{\mathcal{C}}(x, y) \geq 0$ (neboť $d_{\mathcal{C}}(x) = 0$ a $d_{\mathcal{C}}(z) \geq 0\ \forall z \in R^n$), dostaneme $d^0_{\mathcal{C}}(x, y) = 0$ a tedy $y \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$. □

Věta 339. *Nechť $\mathcal{C} \subset R^n$ je uzavřená množina a $x \in \mathcal{C}$. Pak platí*

$$\tilde{\mathcal{N}}_{\mathcal{C}}(x) = \overline{\operatorname{cone} \partial d_{\mathcal{C}}(x)} = \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_{\mathcal{C}}(x)}$$

(zde $\partial d_{\mathcal{C}}(x)$ je subdiferenciál podle definice 118).

Důkaz (a) Předpokládejme, že $z \in \partial d_{\mathcal{C}}(x)$. Pak podle definice 118 platí

$$d^0_{\mathcal{C}}(x, y) \geq z^T y \quad \forall y \in R^n.$$

Jestliže $y \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$, platí podle definice 120 $d^0_{\mathcal{C}}(x, y) = 0$, takže

$$z^T y \leq 0 \quad \forall y \in \tilde{\mathcal{T}}_{\mathcal{C}}(x),$$

což podle definice 107 a definice 120 dává $z \in \tilde{\mathcal{N}}_{\mathcal{C}}(x)$. Jelikož $\tilde{\mathcal{N}}_{\mathcal{C}}(x)$ je uzavřený konvexní kužel, platí

$$\overline{\bigcup_{\lambda \geq 0} \lambda \partial d_{\mathcal{C}}(x)} \subset \tilde{\mathcal{N}}_{\mathcal{C}}(x)$$

(b) Nechť $z \in \tilde{\mathcal{N}}_{\mathcal{C}}(x)$. Pak podle definice 107 a definice 120 pro $y \in \tilde{\mathcal{T}}_{\mathcal{C}}(x)$ platí

$$z^T y \leq 0 = d^0_{\mathcal{C}}(x, y) = \lambda(y) d^0_{\mathcal{C}}(x, y),$$

kde $\lambda(y) = 1 \geq 0$. Zbývá dokázat podobnou nerovnost i pro $y \notin \tilde{\mathcal{T}}_{\mathcal{C}}(x)$. Nechť $y \notin \tilde{\mathcal{T}}_{\mathcal{C}}(x)$. Jelikož $d^0_{\mathcal{C}}(x, y) > 0$ pro $y \notin \tilde{\mathcal{T}}_{\mathcal{C}}(x)$ (neboť $d^0_{\mathcal{C}}(x, y) \geq 0$ a podle definice 120 platí $d^0_{\mathcal{C}}(x, y) \neq 0$ pro $y \notin \tilde{\mathcal{T}}_{\mathcal{C}}(x)$), můžeme položit

$$\lambda(y) = \frac{\|z\|\|y\|}{d^0_{\mathcal{C}}(x,y)} \geq 0.$$

Použitím Schwarzovy nerovnosti dostaneme

$$z^T y \leq \|z\|\|y\| = \lambda(y) d^0_{\mathcal{C}}(x,y).$$

Dokázali jsme, že pro libovolný vektor $y \in R^n$ existuje $\lambda(y) \geq 0$ tak, že $z^T y \leq \lambda(y) d^0_{\mathcal{C}}(x,y)$. Pokud $z \neq 0$, dostaneme $(z/\lambda(y))^T y \leq d^0_{\mathcal{C}}(x,y)$, což podle definice 118 dává $z/\lambda(y) \in \partial d_{\mathcal{C}}(x)$, neboli $z \in \lambda(y)\partial d_{\mathcal{C}}(x)$ (pokud $z = 0$ je tato inkluze triviální). Odtud plyne, že $z \in \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_{\mathcal{C}}(x)}$, takže

$$\tilde{\mathcal{N}}_{\mathcal{C}}(x) \subset \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_{\mathcal{C}}(x)}$$

. □

Věta 340. *Nechť $\mathcal{C} \in R^n$ je uzavřená množina, $x \in \mathcal{C}$ a funkce $F : R^n \to R$ je lipschitzovská s konstantou L v okolí bodu x, který je jejím lokálním minimem na $\mathcal{C}$. Pak platí $0 \in \partial F(x) + \tilde{\mathcal{N}}_{\mathcal{C}}(x)$.*

Důkaz (a) Nechť F je lipschitzovská s konstantou L v $\mathcal{B}(x,\varepsilon)$. Ukážeme nejprve, že jsou-li splněny předpoklady dokazované věty, je bod x lokálním minimem funkce $F(x) + L\, d_{\mathcal{C}}(x)$ v $\mathcal{B}(x,\varepsilon)$. Předpokládejme naopak, že existuje posloupnost $x_i \to x$ taková, že $x_i \in \mathcal{B}(x,\varepsilon)$ a $F(x_i) + L\, d_{\mathcal{C}}(x_i) < F(x) + L\, d_{\mathcal{C}}(x)$, $i \in N$. Nechť $y_i = P_{\mathcal{C}}(x_i)$, $i \in N$, takže $y_i \in \mathcal{C}$ a $\|y_i - x_i\| = d_{\mathcal{C}}(x_i)$, $i \in N$. Jelikož $d_{\mathcal{C}}(x_i) \to 0$ (neboť $x_i \to x$ a $x \in \mathcal{C}$), platí $\|y_i - x\| \leq \|y_i - x_i\| + \|x_i - x\| = d_{\mathcal{C}}(x_i) + \|x_i - x\| \to 0$ a tedy $y_i \to x$. Jelikož funkce F je lipschitzovská s konstantou L v okolí bodu x, můžeme psát

$$\begin{aligned} F(y_i) &= F(x_i) + (F(y_i) - F(x_i)) \leq F(x_i) + L\,\|y_i - x_i\| \\ &= F(x_i) + L\, d_{\mathcal{C}}(x_i) < F(x) + L\, d_{\mathcal{C}}(x) = F(x), \end{aligned}$$

což je spor, neboť $y_i \in \mathcal{C}$, $y_i \to x$ a bod x je lokálním minimem funkce F na $\mathcal{C}$.

(b) Je-li bod x lokálním minimem funkce $F(x) + L\, d_{\mathcal{C}}(x)$ na R_n, platí podle věty 337 $0 \in \partial(F(x) + L\, d_{\mathcal{C}}(x))$, což s použitím věty 336 a věty 339 dává

$$0 \in \partial(F(x) + L\, \partial d_{\mathcal{C}}(x)) \subset \partial F(x) + \operatorname{cone} \partial d_{\mathcal{C}}(x) \subset \partial F(x) + \tilde{\mathcal{N}}_{\mathcal{C}}(x).$$

□

Poznámka 431. podmínka $0 \in \partial F(x) + \tilde{\mathcal{N}}_{\mathcal{C}}(x)$ je obecně mnohem slabší než podmínka $0 \in \partial F(x) + \mathcal{N}_{\mathcal{C}}(x)$. Je-li však funkce F regulární, obě podmínky splývají. Jelikož konvexní funkce je podle věty 334 regulární, plyne z věty 340 druhá část (nutná podmínka) tvrzení věty 326.

Pro další analýzu nehladkých funkcí je důležitá věta o střední hodnotě. Abychom zjednodušili symboliku, budeme pro libovolný vektor $v \in R^n$ používat označení

$$(\partial F(z))^T v = \left\{ g^T v : g \in \partial F(z) \right\}$$

(je to uzavřený interval).

Věta 341. *Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská na otevřené množině Ω obsahující úsečku $[x,y]$. Pak existuje bod $z \in (x,y)$ takový, že*

$$F(y) - F(x) \in (\partial F(z))^T (y - x). \tag{1131}$$

Důkaz Uvažujme funkci $\varphi(\lambda) = F(x + \lambda(y - x))$. Podle předpokladu je tato funkce lokálně lipschitzovská na množině obsahující interval $[0, 1]$. Ukážeme nejprve, že

$$\partial\varphi(\lambda) \subset (\partial F(x + \lambda(y - x)))^T (y - x). \tag{1132}$$

Podle věty 332 (a) jsou množiny na obou stranách této inkluze uzavřené intervaly. Podle poznámky 407 stačí dokázat, že

$$\delta_{\partial\varphi(\lambda)}(\beta) \leq \delta_{(\partial F(x+\lambda(y-x)))^T(y-x)}(\beta) \tag{1133}$$

pro $\beta = 1$ a $\beta = -1$. Podle definice 117 a věty 332 (b) platí

$$\begin{aligned}
\varphi^0(\lambda, \beta) &= \limsup_{\substack{\lambda' \to \lambda \\ t \downarrow 0}} \frac{\varphi(\lambda' + t\beta) - \varphi(\lambda')}{t} \\
&= \limsup_{\substack{\lambda' \to \lambda \\ t \downarrow 0}} \frac{F(x + (\lambda' + t\beta)(y - x)) - F(x + \lambda'(y - x))}{t} \\
&\leq \limsup_{\substack{y' \to x + \lambda(y-x) \\ t \downarrow 0}} \frac{F(y' + t\beta(y - x)) - F(y')}{t} \\
&= F^0(x + \lambda(y - x), \beta(y - x)) \\
&= \max\left\{\beta g^T (y - x) : g \in \partial F(x + \lambda(y - x))\right\}
\end{aligned}$$

pro $\beta = 1$ a $\beta = -1$, což podle poznámky 406 a poznámky 425 dává (1133) a tedy i (1132). Položme nyní

$$\psi(\lambda) = \varphi(\lambda) - \varphi(0) + \lambda(\varphi(0) - \varphi(1)) = F(x + \lambda(y - x)) - F(x) + \lambda(F(x) - F(y)).$$

Tato funkce je spojitá na intervalu $[0, 1]$ a platí $\psi(0) = \psi(1) = 0$. Musí tedy nabývat minima nebo maxima v nějakém bodě $\lambda^* \in (0, 1)$, což podle věty 337 dává $0 \in \partial\psi(\lambda^*)$. Použijeme-li větu 336 a vztah (1132), dostaneme

$$0 \in \partial\psi(\lambda^*) \subset \partial\varphi(\lambda^*) + (\varphi(0) - \varphi(1)) \subset (\partial F(x + \lambda^*(y - x)))^T (y - x) + (F(x) - F(y)),$$

protože $\partial(\lambda) = \{1\}$, což přičtením $F(y) - F(x)$ k oběma stranám inkluze dává (1131). □

Je-li funkce $F : R^n \to R$ lokálně lipschitzovská v otevřené množině $\Omega \subset R^n$, je podle Rademacherovy věty (tvrzení 10) diferencovatelná skoro všude v Ω neboli množina

$$\Omega_F = \{x \in \Omega : \nabla F(x) \text{ neexistuje}\}$$

má Lebesgueovu míru nula. V tomto případě můžeme subdiferenciál definovat též jiným způsobem.

Věta 342. *Nechť funkce $F : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$. Pak platí*

$$\partial F(x) = \operatorname{conv} \partial_B F(x),$$

kde

$$\partial_B F(x) = \left\{\lim_{i \to \infty} \nabla F(x_i) : x_i \to x,\ x_i \notin \Omega_F\right\}.$$

Důkaz (a) Dokážeme nejprve, že pro libovolné $h \in R^n$ platí

$$F^0(x, h) \leq \limsup_{\substack{y \to x \\ y \notin \Omega_F}} \nabla^T F(y) h. \tag{1134}$$

Zvolme $h \in R^n$, $\varepsilon > 0$ libovolně a označme α pravou stranu v (1134). Z definice horní limity (limes superior) plyne existence čísla $\delta > 0$ takového, že $\nabla^T F(y)h \leq \alpha + \varepsilon$ pokud $y \in \mathcal{B}(x, \delta)$ a $y \notin \Omega_F$. Bez újmy na obecnosti můžeme předpokládat, že F je lipschitzovská v $\mathcal{B}(x, \delta)$, takže podle Rademacherovy věty má $\mathcal{B}(x, \delta) \cap \Omega_F$ Lebesgueovu míru nula. Označme

$$\mathcal{L}_y = \{y + th : 0 < t < \delta/(2\|h\|)\},$$

takže $\mathcal{L}_y \subset \mathcal{B}(x, \delta)$, pokud $y \in \mathcal{B}(x, \delta/2)$. Z teorie Lebesgueovy míry plyne, že pro skoro všechny body $y \in \mathcal{B}(x, \delta/2)$ má množina $\mathcal{L}_y \cap \Omega_F$ Lebesgueovu míru nula. Pro skoro všechny body $y \in \mathcal{B}(x, \delta/2)$ a pro všechna $t \in (0, \delta/(2\|h\|))$ tedy existuje integrál

$$F(y + th) - F(y) = \int_0^t \nabla^T F(y + \vartheta h) h d\vartheta.$$

Jelikož $\nabla^T F(y + \vartheta h)h \leq \alpha + \varepsilon$ kdykoliv $\nabla F(y + \vartheta h)$ existuje, můžeme tento integrál majorizovat, takže

$$F(y + th) - F(y) \leq t(\alpha + \varepsilon). \tag{1135}$$

Tato nerovnost platí pro skoro všechny body $y \in \mathcal{B}(x, \delta/2)$ a pro všechna $t \in (0, \delta/(2\|h\|))$. Jelikož funkce F je spojitá, musí (1135) platit pro všechny body $y \in \mathcal{B}(x, \delta/2)$ a pro všechna $t \in (0, \delta/(2\|h\|))$, což podle definice 117 dává

$$F^0(x, h) \leq \alpha + \varepsilon.$$

Jelikož $\varepsilon > 0$ je libovolné, dostáváme (1134).

(b) Protože Ω_F má Lebesgueovu míru nula, existuje alespoň jedna posloupnost $y_i \to x$, $y_i \notin \Omega_F$. Podle poznámky 426 platí $\nabla F(y_i) \in \partial F(y_i)$, takže podle věty 332 (a) je posloupnost $\{\nabla F(y_i)\}$ omezená a existuje tedy konvergentní podposloupnost $\{\nabla F(y_i')\} \subset \{\nabla F(y_i)\}$. Množina $\partial_B F(x)$ je tedy neprázdná a podle věty 332 (c) platí

$$\lim_{i \to \infty} \nabla F(y_i') \in \partial F(x)$$

takže $\partial_B F(x) \subset \partial F(x)$. Jelikož $\partial F(x)$ je konvexní, platí také conv $\partial_B F(x) \subset \partial F(x)$. Jelikož $\partial F(x)$ je kompaktní, jsou i množiny $\partial_B F(x)$ a conv $\partial_B F(x)$ kompaktní. Použijeme-li poznámku 425 a nerovnost (1134), dostaneme

$$\begin{aligned} \delta_{\partial F(x)}(h) &= F^0(x, h) \leq \limsup_{\substack{y \to x \\ y \notin \Omega_F}} \nabla^T F(y) h = \sup_{g \in \partial_B F(x)} g^T h \\ &\leq \sup_{g \in \text{conv } \partial_B F(x)} g^T h = \delta_{\text{conv } \partial_B F(x)}(h) \end{aligned}$$

pro libovolný vektor $h \in R^n$, takže podle věty 298 platí $\partial F(x) \subset$ conv $\partial_B F(x)$. □

15.5 Lipschitzovská zobrazení

Přístup použitý ve větě 342 můžeme využít k definici zobecněného Jakobiánu lokálně lipschitzovského zobrazení $f : R^n \to R^m$. Stejně jako v případě lokálně lipschitzovské funkce zavedeme množinu

$$\Omega_f = \{x \in \Omega : \nabla f(x) \text{ neexistuje}\},$$

kde

$$\nabla f(x) = \begin{bmatrix} \dfrac{\partial f_1(x)}{\partial x_1}, & \dots, & \dfrac{\partial f_1(x)}{\partial x_n} \\ -- & -- & -- \\ \dfrac{\partial f_m(x)}{\partial x_1}, & \dots, & \dfrac{\partial f_m(x)}{\partial x_n} \end{bmatrix},$$

která má opět Lebesgueovu míru nula.

Definice 121. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak množinu*

$$\partial f(x) = \operatorname{conv} \partial_B f(x),$$

kde

$$\partial_B f(x) = \left\{ \lim_{i\to\infty} \nabla f(x_i) : x_i \to x, x_i \notin \Omega_f \right\},$$

nazveme zobecněným Jakobiánem zobrazení f.

Poznámka 432. Poznamenejme, že se dopouštíme jisté nedůslednosti, neboť pro $m = 1$ se $\nabla f(x)$ liší od $\nabla F(x)$ (platí $\nabla f(x) = (\nabla F(x))^T$). Tato konvence, která se běžně používá v literatuře, je výhodná proto, že pak $\nabla f(x) = J(x)$, kde $J(x)$ je Jacobiova matice zobrazení f.

Poznámka 433. Je-li zobrazení $f : R^n \to R^m$ diferencovatelné v bodě $x \in R^n$, pak přímo z definice 121 plyne, že

$$\nabla f(x) \in \partial f(x)$$

(stačí zvolit posloupnost $x_i = x \to x \notin \Omega_f$).

Věta 343. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské s konstantou L v okolí bodu $x \in R^n$. Pak*

(a) Platí $\partial f(x) \subset [\partial f_1(x), \dots, \partial f_m(x)]^T$, kde $\partial f_i(x)$, $1 \le i \le m$, jsou subdiferenciály funkcí $f_i : R^n \to R$ (i-tých složek zobrazení f) v bodě $x \in R^n$.

(b) Zobecněný Jakobián $\partial f(x)$ je neprázdná konvexní kompaktní množina taková, že $\|J\| \le L$ $\forall J \in \partial f(x)$.

(c) Jestliže $x_i \to x$, $J_i \in \partial f(x_i)$ a $J_i \to J$, pak $J \in \partial f(x)$ (polospojitost shora).

Důkaz (a) plyne bezprostředně z věty 342.

(b) Kompaktnost plyne bezprostředně z (a) a z věty 332 (a). Konvexita plyne přímo z definice 121. Neprázdnost plyne z existence alespoň jedné posloupnosti $x_i \to x$, $x_i \notin \Omega_f$, pro kterou $\{\nabla f(x_i)\}$ konverguje (argumentace je stejná jako v důkazu věty 342). Nerovnost $\|J\| \le L$ plyne z definice 121 a z toho, že $\|\nabla f(x_i)\| \le L$ pokud $\nabla f(x_i)$ existuje.

(c) Předpokládejme, že $x_i \to x$, $J_i \in \partial f(x_i)$ a $J_i \to J$. Bez újmy na obecnosti budeme předpokládat, že $x_i \in \mathcal{B}(x, 1/(2i))$ (v opačném případě lze vybrat vhodnou podposloupnost). Jestliže $J \notin \partial f(x)$, musí existovat číslo $\varepsilon > 0$ takové, že pro dostatečně velké indexy platí

$$J_i \notin \partial f(x) + \mathcal{B}(0, \varepsilon).$$

Protože množina $\partial f(x) + \mathcal{B}(0, \varepsilon)$ je konvexní, nemůže platit $\partial_B f(x_i) \subset \partial f(x) + \mathcal{B}(0, \varepsilon)$ (v opačném případě by muselo platit $J_i \in \operatorname{conv} \partial_B f(x_i) \subset \partial f(x) + \mathcal{B}(0, \varepsilon)$). Existuje tedy matice $\overline{J}_i \in \partial_B f(x_i)$ taková, že $\overline{J}_i \notin \partial f(x) + \mathcal{B}(0, \varepsilon)$. Podle definice 121 musí existovat bod $y_i \in \mathcal{B}(x_i, 1/(2i)) \subset \mathcal{B}(x, 1/i)$, $y_i \notin \Omega_f$ takový, že $\|\nabla f(y_i) - \overline{J}_i\| < \varepsilon/2$, takže

$$\nabla f(y_i) \notin \partial f(x) + \mathcal{B}(0, \varepsilon/2). \tag{1136}$$

Podle (a) jsou matice $\overline{J}_i$ a tedy i $\nabla f(y_i)$ stejnoměrně omezené v okolí bodu x. Můžeme tedy předpokládat, že existuje limita

$$\lim_{i\to\infty} \nabla f(y_i) = \overline{J}$$

(v opačném případě lze vybrat vhodnou podposloupnost). Zřejmě $y_i \to x$ (neboť $y_i \in \mathcal{B}(x, 1/i)$), $y_i \notin \Omega_f$ (neboť $\nabla f(y_i)$ existuje) a $\nabla f(y_i) \to \overline{J}$. Podle definice 121 tedy platí $\overline{J} \in \partial f(x)$, což je ve sporu s (1136). □

Lemma 124. *Nechť zobrazení $f : R^n \to R^m$ je lokálně lipschitzovské v okolí bodu $x \in R^n$ a funkce $F : R^m \to R$ je spojitě diferencovatelná v okolí bodu $f(x)$. Pak funkce $\varphi = F \circ f : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ a platí*

$$\partial\varphi(x) = (\partial f(x))^T \nabla F(f(x)).$$

Důkaz Lipschitzovskost funkce $F \circ f$ je zřejmá (stačí použít větu 333 a definici 116). Nechť $J \in \partial_B f(x)$. Pak existuje posloupnost $x_i \to x$, $x_i \notin \Omega_f$ taková, že $\nabla f(x_i) \to J$ a tudíž $\nabla\varphi(x_i) = (\nabla f(x_i))^T \nabla F(f(x_i)) \to J^T \nabla F(f(x))$. Platí tedy $J^T \nabla F(f(x)) \in \partial_B\varphi(x)$, což dává

$$(\partial_B f(x))^T \nabla F(f(x)) \subset \partial_B\varphi(x).$$

Nechť naopak $w \in \partial_B\varphi(x)$. Pak existuje posloupnost $x_i \to x$, $x_i \notin \Omega_\varphi$, kde $\Omega_f \subset \Omega_\varphi$, taková, že $\nabla\varphi(x_i) = (\nabla f(x_i))^T \nabla F(f(x_i)) \to w$. Jelikož Jacobiovy matice $\nabla f(x_i)$ jsou podle věty 343 (b) omezené v okolí bodu x, existuje podposloupnost $\{x_i'\} \subset \{x_i\}$ taková, že $\nabla f(x_i') \to J \in \partial_B f(x)$, což spolu s $(\nabla f(x_i'))^T \nabla F(f(x_i')) \to w$ dává

$$\partial_B\varphi(x) \subset (\partial_B f(x))^T \nabla F(f(x)).$$

Spojením obou inkluzí dostaneme $\partial_B\varphi(x) = (\partial_B f(x))^T \nabla F(f(x))$, což po přechodu ke konvexním obalům dává $\partial\varphi(x) = (\partial f(x))^T \nabla F(f(x))$. □

Abychom mohli zformulovat větu o střední hodnotě, zavedeme označení

$$\partial f([x,y]) = \operatorname{conv} \bigcup_{z\in[x,y]} \partial f(z). \tag{1137}$$

Lemma 125. *Množina $\partial f([x,y])$ je konvexní a kompaktní.*

Důkaz Konvexita plyne bezprostředně z (1137). Abychom dokázali kompaktnost, stačí podle věty 287 dokázat kompaktnost množiny $\bigcup_{z\in[x,y]} \partial f(z)$. Nechť $\{J_i\} \subset \bigcup_{z\in[x,y]} \partial f(z)$ je posloupnost taková, že $J_i \to J$. Zřejmě $J_i \in \partial f(z_i)$, kde $z_i \in [x,y]$. Jelikož množina $[x,y]$ je kompaktní, existuje podposloupnost $\{z_i'\} \subset \{z_i\}$ taková, že $z_i' \to z \in [x,y]$, a odpovídající podposloupnost $\{J_i'\} \subset \{J_i\}$ taková, že $J_i' \in \partial f(z_i')$. Jelikož vybraná posloupnost má stejnou limitu jako původní konvergentní posloupnost, lze psát $J_i' \to J$, a podle věty 343 (c) dostaneme $J \in \partial f(z) \subset \bigcup_{z\in[x,y]} \partial f(z)$. □

Věta 344. *Nechť zobrazení $f : R^n \to R^m$ je lokálně lipschitzovské na otevřené množině Ω obsahující úsečku $[x,y]$. Pak platí*

$$f(y) - f(x) \in \partial f([x,y])\,(y - x). \tag{1138}$$

Důkaz Podle lemmatu 124 pro libovolný bod $z \in (x,y)$ a pro libovolný vektor $v \in R^m$ platí $\partial(v^T f)(z) = v^T \partial f(z)$. Můžeme tedy použít větu 341, podle které pro libovolný vektor $v \in R^m$ existuje bod $z \in (x,y)$ takový, že

$$v^T(f(y) - f(x)) \in \partial(v^T f)(z)(y-x) = v^T \partial f(z)(y-x). \tag{1139}$$

Vztah (1138) dokážeme sporem. Předpokládejme, že $f(y) - f(x) \notin \partial f([x,y])(y-x)$. Jelikož množina na pravé straně je podle lemmatu 125 konvexní a kompaktní, musí podle věty 292 existovat vektor $v \in R^m$ a číslo $\alpha \in R$ tak, že

$$v^T(f(y) - f(x)) > \alpha \geq \max_{J \in \partial f([x,y])} v^T J(y-x),$$

což je ve sporu s (1139), neboť podle (1139) existuje prvek $J \in \partial f(z) \subset \partial f([x,y])$ takový, že $v^T(f(y) - f(x)) = v^T J(y-x)$. □

Věta 345. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$ a funkce $\varphi : R^m \to R$ je lipschitzovská v okolí bodu $f(x)$. Pak funkce $F = \varphi \circ f : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ a platí*

$$\partial F(x) \subset \operatorname{conv}(\partial f(x))^T \partial\varphi(f(x)) \triangleq \operatorname{conv}\left\{J^T v : J \in \partial f(x), v \in \partial\varphi(f(x))\right\}, \tag{1140}$$

přičemž rovnost nastává zejména v těchto případech

(a) Funkce φ je spojitě diferencovatelná v bodě $f(x)$. V tomto případě platí

$$\partial F(x) = (\partial f(x))^T \nabla\varphi(f(x)). \tag{1141}$$

(b) Funkce φ je regulární v bodě $f(x)$ a zobrazení f je spojitě diferencovatelné v bodě x. V tomto případě je funkce F regulární v bodě x a platí

$$\partial F(x) = (\nabla f(x))^T \partial\varphi(f(x)). \tag{1142}$$

(c) Funkce φ je regulární v bodě $f(x)$, funkce $f_i = e_i^T f$, $1 \leq i \leq m$, jsou regulární v bodě x a pro libovolný prvek $v \in \partial\varphi(f(x))$ platí $v_i \geq 0$, $1 \leq i \leq m$. V tomto případě je funkce F regulární v bodě x.

Důkaz Lipschitzovskost funkce $\varphi \circ f$ je zřejmá (stačí dvakrát použít definici 116). Označme $\mathcal{S}$ množinu na pravé straně (1140). Abychom dokázali inkluzi $\partial F(x) \subset \mathcal{S}$, použijeme větu 298 a poznámku 425. Jelikož podle věty 296 pro libovolný vektor $h \in R^n$ platí

$$\delta_{\mathcal{S}}(h) = \max\left\{v^T J h : J \in \partial f(x), v \in \partial\varphi(f(x))\right\},$$

stačí podle věty 298 a poznámky 425 ukázat, že pro libovolný vektor $h \in R^n$ existuje matice $J \in \partial f(x)$ a vektor $v \in \partial\varphi(f(x))$ tak, že

$$\delta_{\partial F}(h) = F^0(x,h) \leq v^T J h. \tag{1143}$$

Podle poznámky 423 můžeme vybrat posloupnosti $x_i \to x$ a $t_i \downarrow 0$ tak, že

$$F^0(x,h) = \lim_{i \to \infty} \frac{F(x_i + t_i h) - F(x_i)}{t_i}.$$

Je-li bod $x_i \in R^n$ dostatečně blízko k bodu x a je-li číslo $t_i > 0$ dostatečně malé, jsou i body $f(x_i)$ a $f(x_i + t_i h)$ dostatečně blízké k bodu $f(x)$. Jsou tedy splněny předpoklady věty 341 (aplikované na funkci φ) a existuje tedy bod $u_i \in [f(x_i), f(x_i + t_i h)]$ a subgradient $v_i \in \partial\varphi(u_i)$ tak, že

$$F(x_i + t_i h) - F(x_i) = \varphi(f(x_i + t_i h)) - \varphi(f(x_i)) = v_i^T(f(x_i + t_i h) - f(x_i)).$$

Podle věty 344 platí

$$\frac{f(x_i + t_i h) - f(x_i)}{t_i} \in \partial f([x_i, x_i + t_i h])h,$$

což podle vztahu (1137) a podle věty 284 znamená, že

$$\frac{F(x_i + t_i h) - F(x_i)}{t_i} = v_i^T \frac{f(x_i + t_i h) - f(x_i)}{t_i} = v_i^T \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

kde $J_i^k \in \partial f(y_i^k)$, $y_i^k \in [x_i, x_i + t_i h]$, $\lambda_i^k \geq 0$, $k \in [1, m+1]$, $\lambda_i^1 + \cdots + \lambda_i^{m+1} = 1$. Z tohoto důvodu musí alespoň pro jeden index $k \in [1, m+1]$ platit

$$\frac{F(x_i + t_i h) - F(x_i)}{t_i} \leq v_i^T J_i^k h. \tag{1144}$$

Jelikož $x_i \to x$ a $t_i \downarrow 0$, platí $u_i \to f(x)$ a $y_i^k \to x$. Z kompaktnosti subdiferenciálu a zobecněného Jakobiánu plyne existence podposloupností $\{x_i'\} \subset \{x_i\}$ a $\{t_i'\} \subset \{t_i\}$ takových, že odpovídající podposloupnosti $\{v_i'\} \subset \{v_i\}$ a $\{J_i'\} \subset \{J_i^k\}$ konvergují k v a J. Podle věty 332 (c) a věty 343 (c) platí $v \in \partial\varphi(f(x))$ a $J \in \partial f(x)$, takže z (1144) plyne (1143). Nyní vyšetříme speciální případy:

(a) Tento případ je tvrzením lemmatu 124.

(b) Je-li zobrazení f spojitě diferencovatelné, můžeme množinu $\mathcal{S}$ zapsat ve tvaru $\mathcal{S} = (\nabla f(x))^T \partial\varphi(f(x))$ (protože množina $\partial f(x) = \{\nabla f(x)\}$ je jednoprvková nemusíme používat její konvexní obal). Použijeme-li definici 104, poznámku 425 a regularitu funkce φ (definice 119), můžeme psát

$$\begin{aligned}
\delta_{\mathcal{S}}(h) &= \max_{v \in \partial\varphi(f(x))} v^T \nabla f(x) h = \max_{v \in \partial\varphi(f(x))} v^T f'(x,h) \\
&= \varphi^0(f(x), f'(x,h)) = \varphi'(f(x), f'(x,h)) \\
&= \lim_{t \downarrow 0} \frac{\varphi(f(x) + t f'(x,h)) - \varphi(f(x))}{t} = \lim_{t \downarrow 0} \left(\frac{\varphi(f(x + th)) - \varphi(f(x))}{t} + T(t) \right),
\end{aligned}$$

kde pro dostatečně malá t platí

$$\begin{aligned}
\|T(t)\| &= \frac{\|\varphi(f(x) + t f'(x,h)) - \varphi(f(x + th))\|}{t} \leq \frac{L \|f(x) + t f'(x,h) - f(x + th)\|}{t} \\
&= L \left\| f'(x,h) - \frac{f(x + th) - f(x)}{t} \right\|,
\end{aligned}$$

neboť funkce φ je lipschitzovská v nějakém okolí bodu $f(x)$ (konstantu jsme označili L). Ze spojité diferencovatelnosti zobrazení f plyne, že $(f(x + th) - f(x))/t \to f'(x,h)$, takže $T(t) \to 0$ pokud $t \downarrow 0$. Ukázali jsme tedy, že

$$F'(x,h) = \lim_{t \downarrow 0} \frac{\varphi(f(x + th)) - \varphi(f(x))}{t}$$

existuje a platí $\delta_{\mathcal{S}}(h) = F'(x,h) \leq F^0(x,h)$, což podle věty 298 dává $\mathcal{S} \subset \partial F(x)$, takže z (1140) plyne $\partial F(x) = \mathcal{S}$. Z nerovnosti $F^0(x,h) \leq \delta_{\mathcal{S}}(h) = F'(x,h) \leq F^0(x,h)$ pak plyne regularita funkce F v bodě x.

(c) Označme

$$\mathcal{S}' = \operatorname{conv} \left\{ \sum_{i=1}^{m} u_i v_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)) \right\}.$$

Podle (1140) platí $\partial F(x) \subset \mathcal{S}$ a podle věty 343 (a) platí $\mathcal{S} \subset \mathcal{S}'$, takže $\partial F(x) \subset \mathcal{S}'$. Jsou-li funkce φ a f_i, $1 \le i \le m$, regulární a platí-li $v_i \ge 0$, $1 \le i \le m$, můžeme psát

$$\begin{aligned}
\delta_{\mathcal{S}'}(h) &= \max\left\{\sum_{i=1}^{m} v_i u_i^T h : u_i \in \partial f_i(x), v \in \partial\varphi(f(x))\right\} \\
&\le \max\left\{\sum_{i=1}^{m} v_i \max_{u_i \in \partial f_i(x)} u_i^T h : v \in \partial\varphi(f(x))\right\} \\
&= \max\left\{\sum_{i=1}^{m} v_i f_i^0(x,h) : v \in \partial\varphi(f(x))\right\} \\
&= \max\left\{\sum_{i=1}^{m} v_i f_i'(x,h) : v \in \partial\varphi(f(x))\right\} \\
&= \varphi^0(f(x), f'(x,h)) = \varphi'(f(x), f'(x,h)).
\end{aligned}$$

Konec důkazu je již stejný jako konec důkazu tvrzení (b). Dostaneme $\delta_{\mathcal{S}'}(h) = F'(x,h) \le F^0(x,h)$, což podle věty 298 dává $\mathcal{S}' \subset \partial F(x)$, takže z $\partial F(x) \subset \mathcal{S} \subset \mathcal{S}'$ plyne $\partial F(x) = \mathcal{S} = \mathcal{S}'$. Z nerovnosti $F^0(x,h) \le \delta_{\mathcal{S}}(h) \le \delta_{\mathcal{S}'}(h) = F'(x,h) \le F^0(x,h)$ pak plyne regularita funkce F v bodě x. □

Důsledek 38. *Jsou-li splněny předpoklady věty 345, platí*

$$\partial F(x) \subset \operatorname{conv}\left\{\sum_{i=1}^{m} u_i v_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)),\right\} \tag{1145}$$

přičemž rovnost nastává zejména v těchto případech:

- *(a) Funkce φ je spojitě diferencovatelná v bodě $f(x)$ a $m = 1$.*
- *(b) Funkce φ je regulární v bodě $f(x)$ a funkce f_i, $1 \le i \le m$, jsou spojitě diferencovatelné v bodě x.*
- *(c) Funkce φ je regulární v bodě $f(x)$ a funkce f_i, $1 \le i \le m$, jsou regulární v bodě x a pro libovolný prvek $v \in \partial\varphi(f(x))$ platí $v_i \ge 0$, $1 \le i \le m$.*

Důkaz Stačí použít větu 345 a některé úvahy (například $\mathcal{S} \subset \mathcal{S}'$) z jejího důkazu. □

Důsledek 39. *Nechť funkce $f_1 : R^n \to R$, $f_2 : R^n \to R$ jsou lipschitzovské v okolí bodu $x \in R^n$. Pak funkce $F = f_1 f_2$ je lipschitzovská v okolí bodu x a označíme-li*

$$f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad P = \begin{bmatrix} 0, & 1 \\ 1, & 0 \end{bmatrix},$$

platí

$$\partial F(x) = (\partial f(x))^T P f(x) \subset \partial f_1(x) f_2(x) + f_1(x)\partial f_2(x)$$

přičemž rovnost nastává, jsou-li funkce f_1, f_2 regulární a platí-li $f_1(x) \ge 0$, $f_2(x) \ge 0$. V tomto případě je funkce $F = f_1 f_2$ regulární.

Důkaz Definujme funkci $\varphi : R^2 \to R$ předpisem $\varphi(u_1, u_2) = u_1 u_2$. Tato funkce je spojitě diferencovatelná a tedy (podle věty 333) lipschitzovská v okolí libovolného bodu $u \in R^2$, přičemž platí

$$\nabla\varphi(u) = \begin{bmatrix} u_2 \\ u_1 \end{bmatrix} = Pu.$$

Podle věty 345 je funkce $\varphi \circ f = f_1 f_2$ lipschitzovská v okolí bodu x a platí

$$\partial(f_1 f_2) = \left\{J^T \nabla\varphi(f(x)) : J \in \partial f(x)\right\} = (\partial f(x))^T P f.$$

Vztah $\partial(f_1 f_2) \subset \partial f_1(x) f_2(x) + f_1(x)\partial f_2(x)$ a podmínky pro rovnost dostaneme bezprostředně z důsledku 38 (c). □

Důsledek 40. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak funkce $F = (1/2)f^T f$ je lipschitzovská v okolí bodu x a platí*

$$\partial F(x) = \frac{1}{2}\partial(f^T f)(x) = (\partial(f(x))^T f(x) = \left\{J^T f(x) : J \in \partial(f(x))\right\}. \tag{1146}$$

Důkaz Definujme funkci $\varphi : R^m \to R$ předpisem

$$\varphi(u) = \frac{1}{2}u^T u = \frac{1}{2}\sum_{i=1}^{m} u_i^2.$$

Tato funkce je spojitě diferencovatelná a tedy (podle věty 333) lipschitzovská v okolí libovolného bodu $u \in R^m$, přičemž platí $\nabla\varphi(u) = u$. Podle věty 345 (a) je funkce $F = \varphi \circ f = (1/2)f^T f$ lipschitzovská v okolí bodu x a platí $\partial F(x) = (\partial f(x))^T \nabla\varphi(f(x)) = (\partial f(x))^T f(x)$. □

Věta 346. *Nechť funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou lipschitzovské v okolí bodu $x \in R^n$. Pak funkce*

$$F(x) = \max_{1\le i\le m} f_i(x)$$

je lipschitzovská v okolí bodu x a platí

$$\partial F(x) \subset \text{conv } \{\partial f_i(x) : i \in I(x)\}, \tag{1147}$$

kde $I(x) = \{i \in \{1,\dots,m\} : f_i(x) = F(x)\}$. Jsou-li funkce f_i, $1 \le i \le m$, regulární v bodě x, je funkce F regulární v bodě x a v (1147) platí rovnost.

Důkaz Definujme funkci $\varphi : R^m \to R$ předpisem $\varphi(u) = \max(u_1,\dots,u_m)$. Tato funkce je konvexní v R^m, neboť

$$\begin{aligned}\varphi(\lambda u + (1-\lambda)v) &= \max_{1\le i\le m}(\lambda u_i + (1-\lambda)v_i) \le \lambda \max_{1\le i\le m}(u_i) + (1-\lambda)\max_{1\le i\le m}(v_i)\\ &= \lambda\varphi(u) + (1-\lambda)\varphi(v)\end{aligned}$$

pro $u \in R^m$, $v \in R^m$ a $1 \le \lambda \le 1$, takže je lokálně lipschitzovská podle věty 321. Nechť $I(u) = \{i \in \{1,\dots,m\} : u_i = \varphi(u)\}$. Pak platí

$$\begin{aligned}\varphi'(u,d) &= \lim_{t\downarrow 0}\frac{\varphi(u+td)-\varphi(u)}{t} = \lim_{t\downarrow 0}\max_{1\le i\le m}\left(\frac{u_i + td_i - \varphi(u)}{t}\right)\\ &= \lim_{t\downarrow 0}\max_{i\in I(u)}\left(\frac{u_i + td_i - \varphi(u)}{t}\right) = \max_{i\in I(u)}(d_i),\end{aligned}$$

takže $\varphi^0(u,d) = \varphi'(u,d) = \max_{i\in I(u)}(d_i)$ a podle definice 118 platí

$$\partial\varphi(u) = \left\{v \in R^n : \max_{i\in I(u)}(d_i) \ge v^T d \ \ \forall d \in R^n\right\}.$$

Nechť e_i je i-tý sloupec jednotkové matice a $\delta > 0$. Jestliže $v_i \ne 0$ pro $i \notin I(u)$, dostaneme volbou $d_i = v_i e_i$ nerovnost $v^T d = v_i^2 > 0 = \max\{d_i, i \in I(u)\}$, takže $v \notin \partial\varphi(u)$. Jestliže $v_i < 0$ pro $i \in I(u)$,

dostaneme volbou $d_i = -\delta e_i$ nerovnost $v^T d = -\delta v_i > -\delta = \max\{d_i, i \in I(u)\}$, takže $v \notin \partial\varphi(u)$. Jestliže $v_i \geq 0\ \forall i \in I(u)$ a $\sum_{i\in I(u)} v_i > 1$, dostaneme volbou $d = \sum_{i\in I(u)} \delta e_i$ nerovnost $v^T d = \delta \sum_{i\in I(u)} v_i > \delta = \max\{d_i, i \in I(u)\}$, takže $v \notin \partial\varphi(u)$. Jestliže $v_i \geq 0\ \forall i \in I(u)$ a $\sum_{i\in I(u)} v_i < 1$, dostaneme volbou $d = -\sum_{i\in I(u)} \delta e_i$ nerovnost $v^T d = -\delta \sum_{i\in I(u)} v_i > -\delta = \max\{d_i, i \in I(u)\}$, takže $v \notin \partial\varphi(u)$. Musí tedy platit

$$\partial\varphi(u) = \left\{ v \in R^n : v_i \geq 0, \sum_{i\in I(u)} v_i = 1, \sum_{i\notin I(u)} v_i = 0 \right\}.$$

Podle důsledku 38 pak platí

$$\begin{aligned} \partial F(x) &\subset \operatorname{conv} \left\{ \sum_{i=1}^{m} v_i u_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)) \right\} \\ &= \operatorname{conv} \left\{ \sum_{i\in I(u)} v_i \partial f_i(x) : v_i \geq 0, \sum_{i\in I(u)} v_i = 1 \right\} \\ &= \operatorname{conv} \left\{ \partial f_i(x), i \in I(u) \right\}. \end{aligned}$$

Funkce φ je konvexní, takže je podle věty 335 regulární. Jsou-li funkce f_i, $1 \leq i \leq m$, regulární, je podle věty 335 i funkce F regulární a jelikož $v_i \geq 0$, $1 \leq i \leq m$, platí v (1147) rovnost. □

15.6 Polohladká zobrazení

Definice 122. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Jestliže pro každé $h \in R^n$ existuje limita*

$$\lim_{\substack{J\in\partial f(x+th) \\ t\downarrow 0}} Jh \tag{1148}$$

(nezávislá na volbě $J \in \partial f(x + th)$), řekneme, že zobrazení f je slabě polohladké v bodě x. Jestliže pro každé $h \in R^n$ existuje limita

$$\lim_{\substack{J\in\partial f(x+th') \\ h'\to h, t\downarrow 0}} Jh' \tag{1149}$$

(nezávislá na volbě $J \in \partial f(x + th')$), řekneme, že zobrazení f je polohladké v bodě x.

Poznámka 434. Jelikož $\partial f(x)$ je množinové zobrazení, mohlo by se zdát, že existence limity (1149) je výjimečná. V dalším textu však ukážeme (poznámka 438), že polohladkost je vlastnost převážné většiny zajímavých lokálně lipschitzovských zobrazení.

Poznámka 435. Z definice 122 plyne, že každé polohladké zobrazení je slabě polohladké. Slabá polohladkost se však nezachovává při skládání funkcí a také věta 352 vyžaduje platnost vztahu (1149).

Věta 347. *Nechť zobrazení $f : R^n \to R^m$ je slabě polohladké v bodě $x \in R^n$. Pak pro libovolný vektor $h \in R^n$ existuje směrová derivace $f'(x, h)$ a platí*

$$f'(x,h) = \lim_{t\downarrow 0} \frac{f(x+th) - f(x)}{t} = \lim_{\substack{J\in\partial f(x+th) \\ t\downarrow 0}} Jh.$$

Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$. Pak pro libovolný vektor $h \in R^n$ existuje směrová derivace $f'(x, h)$ a platí

$$f'(x,h) = \lim_{\substack{h'\to h\\ t\downarrow 0}} \frac{f(x+th')-f(x)}{t} = \lim_{\substack{J\in\partial f(x+th')\\ h'\to h,\ t\downarrow 0}} Jh'.$$

Důkaz (a) Zvolme libovolně vektor $h \in R^n$ a posloupnost $t_i \downarrow 0$. Jelikož zobrazení f je lipschitzovské v okolí bodu x, můžeme bez újmy na obecnosti předpokládat, že je lipschitzovské v každém z intervalů $[x, x+t_ih]$. Použijeme-li větu 344, dostaneme

$$\frac{f(x+t_ih)-f(x)}{t_i} \in \partial f\left([x,x+t_ih]\right)h = \left(\operatorname{conv}\bigcup_{t\in[0,t_i]}\partial f(x+th)\right)h = \operatorname{conv}\left(\bigcup_{t\in[0,t_i]}\partial f(x+th)h\right) \subset R^m$$

Podle věty 284 existuje nejvýše $m+1$ prvků $J_i^k \in \partial f(x+t_i^k h)$, $t_i^k \in [0,t_i]$, $1 \le k \le m+1$, tak, že

$$\frac{f(x+t_ih)-f(x)}{t_i} = \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

kde $0 \le \lambda_i^k \le 1$ a $\lambda_1^k + \cdots + \lambda_{m+1}^k = 1$. Jelikož interval [0,1] je kompaktní, můžeme předpokládat, že $\lambda_i^k \to \lambda^k$, $1 \le k \le m+1$ (v opačném případě vybereme vhodnou podposloupnost). Pak podle (1148) platí

$$\begin{aligned}\lim_{i\to\infty}\frac{f(x+t_ih)-f(x)}{t_i} &= \lim_{i\to\infty}\left(\sum_{k=1}^{m+1}\lambda_i^k J_i^k h\right) = \sum_{k=1}^{m+1}\left(\lim_{i\to\infty}\lambda_i^k\right)\left(\lim_{i\to\infty}J_i^k h\right)\\ &= \left(\sum_{k=1}^{m+1}\lambda^k\right)\lim_{\substack{J\in\partial f(x+th)\\ t\downarrow 0}} Jh = \lim_{\substack{J\in\partial f(x+th)\\ t\downarrow 0}} Jh,\end{aligned}$$

takže limita na levé straně nezávisí na výběru posloupnosti $t_i \downarrow 0$ a rovná se směrové derivaci $f'(x,h)$. Jelikož každé polohladké zobrazení je slabě polohladké a v $h' \to h$ lze volit $h' = h$, dostaneme ihned zbytek tvrzení. □

Poznámka 436. Zobrazení $f'(x,\cdot) : R^n \to R^m$ (směrová derivace) vystupující ve větě 347 je pozitivně homogení a lipschitzovské (poznámka 417). Nemusí však být subaditivní jako v případě konvexních funkcí.

Poznámka 437. Podle věty 347, pro polohladká zobrazení platí

$$f(x+th') = f(x) + tf'$$

kde $f' \to f'(x,h)$, pokud $h' \to h$ a $t \downarrow 0$

V dalším výkladu budeme často používat pojem funkce, tedy zobrazení $f : R^n \to R$, neboli $f : R^n \to R^m$, kde $m = 1$. V tomto případě je třeba mít na paměti konvenci zmíněnou v poznámce 432.

Věta 348. *Jsou-li funkce $f_i : R^n \to R$, $1 \le i \le m$, polohladké v bodě $x \in R^n$, je i zobrazení $f = [f_1, \ldots, f_m]^T$ polohladké v bodě x.*

Důkaz Nechť $h \in R^n$. Limita (1149) existuje právě tehdy, existují-li pro $1 \le i \le m$ limity

$$\lim_{\substack{J\in\partial f(x+th')\\ h'\to h, t\downarrow 0}} e_i^T Jh'.$$

(e_i je i-tý sloupec jednotkové matice řádu m). Tyto limity však existují, neboť pro $1 \le i \le m$ platí $J^T e_i \in \partial f_i(x+th')$ a funkce f_i, $1 \le i \le m$, jsou polohladké. □

Věta 349. *Je-li funkce $F : R^n \to R$ spojitě diferencovatelná v okolí bodu $x \in R^n$, je polohladká v bodě x.*

Důkaz Pro spojitě diferencovatelné funkce platí

$$\lim_{\substack{g \in \partial F(x+th') \\ h' \to h,\ t \downarrow 0}} g^T h' = \lim_{\substack{h' \to h \\ t \downarrow 0}} \left(\nabla F(x+th')\right)^T h' = \left(\nabla F(x)\right)^T h.$$

□

Věta 350. *Je-li funkce $F : R^n \to R$ konvexní v okolí bodu $x \in R^n$, je polohladká v bodě x.*

Důkaz Nechť funkce F je konvexní v $\mathcal{B}(x,\varepsilon)$, $x + th' \in \mathcal{B}(x,\varepsilon)$ a $g \in \partial F(x+th')$. Pak podle věty 323 (d) platí

$$F(x) - F(x+th') \geq g^T(x - (x+th')),$$

neboli

$$\frac{F(x+th') - F(x)}{t} \leq g^T h'.$$

Z druhé strany podle definice 115 platí

$$g^T h' \leq F'(x+th', h').$$

Jelikož funkce konvexní v okolí bodu $x \in R^n$ je v okolí tohoto bodu lipschitzovská s nějakou konstantou L (věta 321), můžeme psát

$$\lim_{h' \to h,\ t \downarrow 0} \frac{\|F(x+th') - F(x+th)\|}{t} \leq \lim_{h' \to h} L\|h' - h\| = 0$$

a jelikož podle věty 322 (a) existuje směrová derivace $F'(x,\cdot)$, platí

$$\lim_{h' \to h,\ t \downarrow 0} \frac{F(x+th') - F(x)}{t} = \lim_{t \downarrow 0} \frac{F(x+th) - F(x)}{t} + \lim_{h' \to h,\ t \downarrow 0} \frac{F(x+th') - F(x+th)}{t} = F'(x,h).$$

pro libovolný vektor $h \in R^n$, což spolu s předchozími nerovnostmi dává

$$\begin{aligned} F'(x,h) &= \lim_{\substack{h' \to h \\ t \downarrow 0}} \frac{F(x+th') - F(x)}{t} \leq \liminf_{\substack{g \in \partial F(x+th') \\ h' \to h,\ t \downarrow 0}} g^T h' \leq \limsup_{\substack{g \in \partial F(x+th') \\ h' \to h,\ t \downarrow 0}} g^T h' \\ &\leq \limsup_{h' \to h,\ t \downarrow 0} F'(x+th', h') \leq F'(x,h), \end{aligned}$$

(poslední nerovnost plyne z věty 322 (c)). Tím je dokázána existence limity (1149) (s g^T místo J). □

Věta 351. *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ a funkce $\varphi : R^m \to R$ je polohladká v bodě $f(x)$. Pak funkce $F = \varphi \circ f$ je polohladká v bodě x.*

Důkaz Nechť vektor $h \in R^n$ je libovolný. Nechť $x_k = x + t_k h_k$, kde $h_k \to h$ a $t_k \downarrow 0$. Podle věty 345 platí $\partial F(x_k) \subset \mathcal{S}_k$, kde symbol $\mathcal{S}_k \subset R^n$ označuje kompaktní množinu na pravé straně výrazu (1140) (s x_k místo x). Nechť

$$\begin{aligned} w_k^- &= (J_k^-)^T v_k^- = \arg\min_{w \in \mathcal{S}_k} w^T h, \quad v_k^- \in \partial\varphi(f(x_k)), \quad J_k^- \in \partial f(x_k), \\ w_k^+ &= (J_k^+)^T v_k^+ = \arg\max_{w \in \mathcal{S}_k} w^T h, \quad v_k^+ \in \partial\varphi(f(x_k)), \quad J_k^+ \in \partial f(x_k). \end{aligned}$$

Pak pro libovolný vektor $w_k \in \partial F(x_k) \subset \mathcal{S}_k$ platí

$$(w_k^-)^T h \leq w_k^T h \leq (w_k^+)^T h. \tag{1150}$$

Jelikož všechny veličiny v těchto vzorcích jsou podle věty 332 (a) omezené, můžeme předpokládat (po případném přechodu k podposloupnostem), že

$$\begin{aligned} J_k^- \to J^- \in \partial f(x), \qquad & v_k^- \to v^- \in \partial\varphi(f(x)), \\ J_k^+ \to J^+ \in \partial f(x), \qquad & v_k^+ \to v^+ \in \partial\varphi(f(x)) \end{aligned}$$

(používáme větu 332 (c)). Jelikož zobrazení f je polohladké, platí $J^-h = J^+h = f'(x,h)$, takže s použitím (1150) dostaneme

$$(v^-)^T f'(x,h) \leq \liminf_{k\to\infty} w_k^T h \leq \limsup_{k\to\infty} w_k^T h \leq (v^+)^T f'(x,h).$$

Jelikož funkce φ je polohladká a podle poznámky 437 platí $f(x_k) = f(x + t_k h_k) = f(x) + t_k f_k'$, kde $f_k' \to f'(x,h)$, pokud $h_k \to h$ a $t_k \downarrow 0$, můžeme použít definici 122, podle které

$$(v^-)^T f'(x,h) = \lim_{k\to\infty} (v_k^-)^T f'(x,h) = \lim_{k\to\infty} (v_k^+)^T f'(x,h) = (v^+)^T f'(x,h),$$

což dokazuje existenci limity posloupnosti $w_k^T h$ nezávislé na volbě vektoru $w_k \in \partial F(x_k)$. □

Důsledek 41. *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ a funkce $\varphi : R^m \to R$ je buď spojitě diferencovatelná nebo konvexní v okolí bodu $f(x)$. Pak funkce $F = \varphi \circ f$ je polohladká v bodě x.*

Důkaz Tvrzení plyne bezprostředně z věty 349, věty 350 a věty 351. □

Důsledek 42. *Nechť funkce $f_i : R^n \to R$, $1 \leq i \leq m$, jsou polohladké v bodě $x \in R^n$ a $\lambda_i \in R$, $1 \leq i \leq m$. Pak funkce $F_1 = \sum_{i=1}^m \lambda_i f_i$ (lineární kombinace) a $F_2 = \prod_{i=1}^m f_i$ (součin) jsou polohladké v bodě x.*

Důkaz Podle věty 348 je zobrazení $f = [f_1, \ldots, f_m]^T$ polohladké v bodě x. Funkce $\varphi_1(u) = \sum_{i=1}^m \lambda_i u_i$ a $\varphi_2(u) = \prod_{i=1}^m u_i$ jsou spojitě diferencovatelné, takže podle důsledku 41 jsou funkce $F_1 = \varphi_1 \circ f$ a $F_2 = \varphi_2 \circ f$ polohladké v bodě x. □

Důsledek 43. *Nechť funkce $f_i : R^n \to R$, $1 \leq i \leq m$, jsou polohladké v bodě $x \in R^n$ a $f = [f_1, \ldots, f_m]^T$. Pak funkce $F = \|f\|$, kde $\|\cdot\|$ je libovolná norma v R^m, je polohladká v bodě x. Speciálně funkce $F_1 = \max_{1\leq i\leq m}(|f_i|)$ (maximum absolutních hodnot) a $F_2 = \sum_{i=1}^m |f_i|$ (součet absolutních hodnot) jsou polohladké v bodě x. Dále funkce $F_3 = \max_{1\leq i\leq m}(f_i)$ (bodové maximum) a $F_4 = \min_{1\leq i\leq m}(f_i)$ (bodové minimum) jsou polohladké v bodě x.*

Důkaz Podle věty 348 je zobrazení $f = [f_1, \ldots, f_m]^T$ polohladké v bodě x. Funkce $\varphi(u) = \|u\|$ je konvexní, neboť z vlastností vektorové normy plyne, že pro $0 \leq \lambda \leq 1$ platí

$$\varphi(\lambda u + (1-\lambda)v) = \|\lambda u + (1-\lambda)v\| \leq \lambda\|u\| + (1-\lambda)\|v\|.$$

Funkce $F = \varphi \circ f$ je tedy podle důsledku 41 polohladká. Také funkce $\varphi_3(u) = \max_{1\leq i\leq m}(u_i)$ je konvexní (důkaz věty 346), takže funkce $F_3 = \varphi_3 \circ f$ je polohladká. Jelikož funkce f_i, $1 \leq i \leq m$, jsou polohladké, jsou podle důsledku 42 i funkce funkce $-f_i$, $1 \leq i \leq m$, polohladké, takže jejich bodové maximum je polohladké. Podle důsledku 42 je tedy i bodové minimum $F_4 = \min_{1\leq i\leq m}(f_i) = -\max_{1\leq i\leq m}(-f_i)$ polohladké. □

Důsledek 44. *(Obrácení věty 348). Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ přičemž $f = [f_1, \ldots, f_m]^T$. Pak funkce $f_i : R^n \to R$, $1 \leq i \leq m$, jsou polohladké v bodě x.*

Důkaz Zřejmě $f_i = \varphi_i \circ f$, $1 \leq i \leq m$, kde funkce $\varphi_i : R^m \to R$, definované předpisem $\varphi_i(u) = e_i^T u = u_i$, jsou spojitě diferencovatelné. Polohladkost funkcí $f_i : R^n \to R$, $1 \leq i \leq m$, tedy plyne z důsledku 41. □

Důsledek 45. *Lineární kombinace polohladkých zobrazení je polohladké zobrazení. Skalární součin polohladkých zobrazení je polohladká funkce.*

Důkaz Podle důsledku 44 jsou složky polohladkých zobrazení polohladkými funkcemi. Podle důsledku 42 je lineární kombinace polohladkých funkcí polohladkou funkcí, takže podle věty 348 je lineární kombinace polohladkých zobrazení polohladkým zobrazením. Polohladkost skalárního součinu plyne z důsledku 44, důsledku 42 a věty 348. □

Poznámka 438. Z předchozího textu vyplývá, že vycházíme-li ze spojitě diferencovatelných a konvexních zobrazení, dostáváme běžnými operacemi (součet, součin, absolutní hodnota, maximum, skládání funkcí) pouze polohladká zobrazení. Proto má teorie polohladkých zobrazení velké uplatnění v praktických aplikacích. Navíc je polohladkost základním předpokladem pro konstrukci numerických metod pro řešení nehladkých rovnic.

V následujících úvahách budeme používat symbol $o(\|h\|)$ pokud $h \to 0$. Tento symbol znamená, že pro libovolnou posloupnost $h_i \to 0$, $h_i \neq 0$ platí $o(\|h_i\|)/\|h_i\| \to 0$.

Věta 352. *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak f je polohladké v bodě x právě tehdy, existuje-li směrová derivace $f'(x,h)$ a platí-li*

$$Jh - f'(x,h) = o(\|h\|) \tag{1151}$$

pokud $h \to 0$ a $J \in \partial f(x+h)$.

Důkaz (a) Nechť zobrazení $f : R^n \to R^m$ je polohladké. Máme dokázat, že

$$\lim_{i\to\infty} \frac{J_i h_i - f'(x,h_i)}{\|h_i\|} = 0. \tag{1152}$$

pro libovolné posloupnosti $\{h_i\} \subset R^n$ a $\{J_i\} \subset R^{m\times n}$ takové, že $h_i \to 0$ a $J_i \in \partial f(x+h_i)$. Předpokládejme naopak, že existují posloupnosti $\{h_i\} \subset R^n$ a $\{J_i\} \subset R^{m\times n}$ takové, že $h_i \to 0$ a $J_i \in \partial f(x+h_i)$, a číslo $\varepsilon > 0$ takové, že že

$$\frac{\|J_i h_i - f'(x,h_i)\|}{\|h_i\|} = \|J_i h_i' - f'(x,h_i')\| \geq \varepsilon \quad \forall i \in N,$$

kde $h_i' = h_i/\|h_i\|$ a $t_i = \|h_i\|$ (takže $J_i \in \partial f(x+t_i h_i')$). Jelikož vektory h_i' jsou omezené (neboť $\|h_i'\| = 1$), můžeme tyto posloupnosti vybrat tak, že $h_i' \to h$. Pak podle věty 347 platí

$$\lim_{i\to\infty} J_i h_i' = f'(x,h)$$

což je však ve sporu s předchozí nerovností, neboť funkce $f'(x,.)$ je podle poznámky 417 spojitá.
(b) Předpokládejme nyní, že existuje směrová derivace $f'(x,.)$ a zobrazení $f : R^n \to R^m$ není polohladké. Pak musí existovat vektor $h \in R^n$ (bez újmy na obecnosti budeme předpokládat, že $\|h\| = 1$), posloupnosti $h_i' \to h$, $t_i \downarrow 0$, $J_i \in \partial f(x+t_i h_i')$ a číslo $\varepsilon > 0$ tak, že

$$\|J_i h_i' - f'(x,h)\| \geq 2\varepsilon \quad \forall i \in N \tag{1153}$$

(v opačném případě by existovala limita (1149) rovnající se $f'(x,h)$, takže zobrazení f by bylo podle definice 122 polohladké). Jelikož směrová derivace je podle poznámky 417 lipschitzovská, platí pro dostatečně velké indexy $\|f'(x,h_i') - f'(x,h)\| \leq \varepsilon$, což spolu s (1153) dává

$$\begin{aligned} \|J_i h_i' - f'(x,h_i')\| &= \|J_i h_i' - f'(x,h) - (f'(x,h_i') - f'(x,h))\| \\ &\geq \|J_i h_i' - f'(x,h)\| - \|(f'(x,h_i') - f'(x,h))\| \geq \varepsilon, \end{aligned}$$

Položme $h_i = t_i h_i'$. Jelikož $\|h_i'\| \to 1$ a $t_i \downarrow 0$, platí $\|h_i\| \to 0$. Z předchozí nerovnosti však plyne

$$\liminf_{i\to\infty} \frac{\|J_i h_i - f'(x,h_i)\|}{\|h_i\|} = \liminf_{i\to\infty} \frac{\|J_i h_i' - f'(x,h_i')\|}{\|h_i'\|} = \liminf_{i\to\infty} \|J_i h_i' - f'(x,h_i')\| \geq \varepsilon > 0,$$

takže neplatí (1152) a tudíž ani (1151). □

Poznámka 439. Vzhledem k platnosti věty 352 se polohladké zobrazení často definuje jako lokálně lipschitzovské zobrazení, které vyhovuje podmínce (1151).

Definice 123. *Řekneme, že zobrazení $f : R^n \to R^m$ je diferencovatelné v Bouligandově smyslu (B-diferencovatelné) v bodě $x \in R^n$, jestliže existuje positivně homogenní zobrazení $f'(x,.) : R^n \to R^m$ (směrová derivace) takové, že*

$$f(x+h) - f(x) - f'(x,h) = o(\|h\|), \tag{1154}$$

pokud $h \to 0$ (to znamená, že zobrazení $f'(x,.)$ má stejné aproximační vlastnosti jako Frechetova derivace).

Věta 353. *Polohladké zobrazení je B-diferencovatelné .*

Důkaz Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$. Máme dokázat, že

$$\lim_{i\to\infty} \frac{f(x+h_i) - f(x) - f'(x,h_i)}{\|h_i\|} = 0.$$

pro libovolnou posloupnost $\{h_i\} \subset R^n$ takovou, že $h_i \to 0$. Předpokládejme naopak, že existuje posloupnost $\{h_i\} \subset R^n$ taková, že $h_i \to 0$, a číslo $\varepsilon > 0$ takové, že

$$\frac{|f(x+h_i) - f(x) - f'(x,h_i)|}{\|h_i\|} = \left| \frac{f(x+t_i h_i') - f(x)}{t_i} - f'(x,h_i') \right| \geq \varepsilon \quad \forall i. \tag{1155}$$

kde $h_i' = h_i/\|h_i\|$ a $t_i = \|h_i\|$. Jelikož vektory h_i' jsou omezené (neboť $\|h_i'\| = 1$), můžeme tuto posloupnost vybrat tak, že $h_i' \to h$. Pak podle věty 347 platí

$$\lim_{i\to\infty} \frac{f(x+t_i h_i') - f(x)}{t_i} = f'(x,h),$$

což je však ve sporu s (1155), neboť funkce $f'(x,\cdot)$ je podle poznámky 417 spojitá. □

Důsledek 46. *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$. Pak platí*

$$f(x+h) - f(x) - Jh = o(\|h\|), \tag{1156}$$

pokud $h \to 0$ a $J \in \partial f(x+h)$.

Důkaz Tvrzení plyne bezprostředně z věty 352 a věty 353. □

16 Metody pro řešení soustav nehladkých rovnic

16.1 Newtonova metoda

Nyní se budeme zabývat řešením soustavy rovnic

$$f(x) = 0, \tag{1157}$$

kde $f : R^n \to R^n$ je polohladké zobrazení. Nejprve se budeme věnovat nepřesné Newtonově metodě, která je iterační a generuje posloupnost $\{x_k\}$ předpisem

$$x_{k+1} = x_k + d_k, \tag{1158}$$

kde vektor d_k se vybírá tak, aby platilo

$$\omega_k = \frac{\|A_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega \tag{1159}$$

a matice A_k se vybírá tak, aby platilo

$$\Delta_k = \|A_k - J_k\| \leq \Delta \tag{1160}$$

pro nějaký prvek $J_k \in \partial_B f(x_k)$. Přitom $\omega \geq 0$, $\Delta \geq 0$ a normy v (1159) a (1160) jsou euklidovské.

Definice 124. *Řekneme, že lokálně lipschitzovské zobrazení $f : R^n \to R^n$ je silně BD-regulární v bodě $x \in R^n$, jestliže všechny matice $J \in \partial_B f(x)$ jsou regulární (množina $\partial_B f(x)$ je uvedena v definici 121).*

Poznámka 440. V iteračním procesu (1158)-(1160) předpokládáme, že A_k aproximuje prvek z $\partial_B f(x_k)$, neboť regularitu všech prvků z $\partial_B f(x_k) \subset \partial f(x_k)$ lze zajistit snadněji než regularitu všech prvků z $\partial f(x_k)$.

Věta 354. *Nechť lokálně lipschitzovské zobrazení $f : R^n \to R^n$ je silně BD-regulární v bodě $x \in R^n$. Pak existuje číslo $\delta > 0$ a konstanta $c \geq 0$ tak, že všechny matice $J \in \partial_B f(y)$ jsou regulární a platí $\|J^{-1}\| \leq c$ pokud $y \in \mathcal{B}(x, \delta)$.*

Důkaz Nejprve dokážeme existenci čísla $\delta > 0$ a konstanty $c \geq 0$ tak, že všechny Jacobiho matice $\nabla f(z)$ jsou regulární a platí

$$\|(\nabla f(z))^{-1}\| \leq c, \tag{1161}$$

pokud $z \in \mathcal{B}(x, \delta) \backslash \Omega_f$ (množina Ω_f je uvedena v definici 121). Předpokládejme, že (1161) neplatí. Pak musí existovat posloupnost $x_i \to x$, $x_i \in \mathcal{B}(x, \delta) \backslash \Omega_f$ taková, že buď všechny Jacobiho matice $\nabla f(x_i)$ jsou singulární nebo $\|(\nabla f(x_i))^{-1}\| \to \infty$. Jelikož zobrazení f je lipschitzovské v okolí bodu x, jsou podle věty 343 (b) Jacobiovy matice $\nabla f(x_i)$ omezené v okolí bodu x. Existuje tedy podposloupnost $\{x_i'\} \subset \{x_i\}$ taková, že $\nabla f(x_i') \to J$. Ze spojité závislosti vlastních čísel na koeficientech matice plyne, že J musí být singulární. Podle definice 121 platí $J \in \partial_B f(x)$, což je v rozporu s definicí 124. Nechť nyní $y \in \mathcal{B}(x, \delta) \cap \Omega_f$ a $J \in \partial_B f(y)$. Pak existuje číslo $0 < \delta' < \delta$ tak, že $\mathcal{B}(y, \delta') \subset \mathcal{B}(x, \delta)$ a (1161) platí pokud $z \in \mathcal{B}(y, \delta') \backslash \Omega_f$. Jelikož podle definice 121 platí

$$J = \lim_{i \to \infty} \nabla f(y_i)$$

pro nějakou posloupnost $y_i \to y$, $y_i \in \mathcal{B}(y, \delta') \backslash \Omega_f$, dostaneme z (1161) a ze spojité závislosti vlastních čísel na koeficientech matice nerovnost $\|J^{-1}\| \leq c$. □

Věta 355. *Nechť zobrazení $f : R^n \to R^n$ je polohladké a silně BD-regulární v bodě $x^* \in R^n$ takovém, že $f(x^*) = 0$. Pak existují čísla $\varepsilon > 0$, $\omega > 0$ a $\Delta > 0$ taková, že pokud $x_1 \in \mathcal{B}(x^*, \varepsilon)$, je iterační proces (1158)-(1160) dobře definován (matice A_k jsou regulární) a posloupnost $\{x_k\}$ konverguje k bodu x^* Q-lineárně. Jestliže navíc $\omega_k \to 0$ a $\Delta_k \to 0$, pak posloupnost $\{x_k\}$ konverguje k bodu x^* Q-superlineárně a také posloupnost $\{f(x_k)\}$ konverguje k nule Q-superlineárně.*

Důkaz Nechť c a δ jsou čísla, jejichž existence plyne z věty 354. Položme $\Delta = 1/(5c)$ a zvolme $\varepsilon \leq \delta$ tak, aby zobrazení f bylo lipschitzovské (s nějakou konstantou L) v $\mathcal{B}(x^*, \varepsilon)$ a aby platilo

$$\|f(x) - f(x^*) - J(x - x^*)\| \leq \frac{\Delta}{2}\|x - x^*\| \quad \forall J \in \partial_B f(x), \tag{1162}$$

pokud $x \in \mathcal{B}(x^*, \varepsilon)$ (to je možné vzhledem k (1156)). Dále položme $\omega = \Delta/(2L)$. Předpokládejme, že $x_k \in \mathcal{B}(x^*, \varepsilon)$ (platí to pro $k = 1$). Pak podle věty 354 platí $\|J_k^{-1}\| \leq c$. Zřejmě

$$A_k^{-1} + J_k^{-1}(A_k - J_k)A_k^{-1} = J_k^{-1}.$$

Jelikož rozdíl norem není větší než norma rozdílu, můžeme psát

$$\|A_k^{-1}\| - \|J_k^{-1}\|\|A_k - J_k\|\|A_k^{-1}\| \leq \|J_k^{-1}\|,$$

neboli

$$\|A_k^{-1}\| \leq \frac{\|J_k^{-1}\|}{1 - \|J_k^{-1}\|\|A_k - J_k\|} \leq \frac{c}{1 - c\Delta} = \frac{5}{4}c,$$

což podle (1158)-(1160) a (1162) (s využitím vztahu $f(x^*) = 0$) dává

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k + d_k - x^*\| = \|x_k + A_k^{-1}(A_k d_k + f(x_k) - f(x_k)) - x^*\| \\
&= \|A_k^{-1}(A_k d_k + f(x_k) - (f(x_k) - J_k(x_k - x^*)) + (A_k - J_k)(x_k - x^*))\| \\
&\leq \|A_k^{-1}\| \left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \|A_k - J_k\|\|x_k - x^*\| + \omega_k\|f(x_k) - f(x^*)\|\right) \\
&\leq \frac{5}{4}c \left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&\leq \frac{5}{4}c\left(\frac{1}{2}\Delta + \Delta + \frac{1}{2}\Delta\right)\|x_k - x^*\| = \frac{1}{2}\|x_k - x^*\|.
\end{aligned} \tag{1163}$$

Odtud plyne, že $x_{k+1} \in \mathcal{B}(x^*, \varepsilon)$, takže můžeme pokračovat stejným způsobem dále. Dokázali jsme tak indukcí, že ve všech iteračních krocích platí $x_{k+1} \in \mathcal{B}(x^*, \varepsilon)$ a $\|x_{k+1} - x^*\| \leq (1/2)\|x_k - x^*\|$ čili, že posloupnost $\{x_k\}$ konverguje k bodu x^* Q-lineárně. Nechť nyní $\omega_k \to 0$ a $\Delta_k \to 0$. Pak podle (1156) a (1163) platí

$$\begin{aligned}
\|x_{k+1} - x^*\| &\leq \frac{5}{4}c\left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&= \frac{5}{4}c\left(o\left(\|x_k - x^*\|\right) + o\left(\|x_k - x^*\|\right) + o\left(\|x_k - x^*\|\right)\right) \\
&= o\left(\|x_k - x^*\|\right)
\end{aligned} \tag{1164}$$

a posloupnost $\{x_k\}$ konverguje k bodu x^* Q-superlineárně. Jelikož $f(x^*) = 0$, můžeme podle (1164) psát

$$\lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|x_k - x^*\|} \leq L \lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0. \tag{1165}$$

S použitím (1158)-(1160) a (1163) dostaneme

$$\begin{aligned}\|x_k - x^*\| &\leq \|x_{k+1} - x_k\| + \|x_{k+1} - x^*\| \\ &\leq \|A_k^{-1}\|\|A_k d_k + f(x_k)\| + \|A_k^{-1}\|\|f(x_k)\| + \|x_{k+1} - x^*\| \\ &\leq \frac{5}{4}c(1+\omega)\|f(x_k)\| + \frac{1}{2}\|x_k - x^*\|,\end{aligned}$$

neboli

$$\|x_k - x^*\| \leq \frac{5}{2}c(1+\omega)\|f(x_k)\|,$$

takže podle (1165) platí

$$\lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|f(x_k)\|} \leq \frac{5}{2}c(1+\omega) \lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|x_k - x^*\|} = 0$$

a $\{f(x_k)\}$ konverguje k nule Q-superlineárně. □

Věta 355 říká, že nepřesná Newtonova metoda (1158)-(1160) je lokálně konvergentní, čili že konverguje, pokud počáteční bod $x_1 \in R^n$ je dostatečně blízko k řešení $x^* \in R^n$. K zaručení globální konvergence (konvergence z libovolného počátečního bodu) je třeba vztah (1158) nahradit výběrem délky kroku. V následujícím algoritmu se pro výběr délky kroku používá funkce

$$F(x) = \frac{1}{2}f^T(x)f(x)$$

a matice A_k se vybírají tak, že $A_k = J_k$ (takže $\Delta_k = 0$).

Algoritmus 4.1

Data $\varrho, \sigma \in (0,1)$, $\omega \in (0, 1-\sigma)$, $\varepsilon > 0$.

Krok 1 (Inicializace). Zvolíme počáteční bod $x_1 \in R^n$ a položíme $k = 1$.

Krok 2 (Směrový vektor). Jestliže $F(x) \leq \varepsilon$, ukončíme výpočet. V opačném případě zvolíme $J_k \in \partial_B f(x_k)$ a určíme směrový vektor d_k tak, aby platilo

$$\omega_k = \frac{\|J_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega. \tag{1166}$$

Krok 3 (Délka kroku). Nechť $t_k = \varrho^{i_k}$, kde i_k je nejmenší nezáporné celé číslo i vyhovující podmínce

$$F(x_k + \varrho^i d_k) - F(x_k) \leq -2\sigma\varrho^i F(x_k). \tag{1167}$$

Krok 4 (Aktualizace). Položíme $x_{k+1} := x_k + t_k d_k$ a $k := k+1$. Přejdeme na Krok 2.

Věta 356. *Nechť množina $X = \{x \in R^n : F(x) \leq F(x_1)\}$ je kompaktní, nechť zobrazení $f : R^n \to R$ je polohladké a silně BD-regulární na $X \subset R^n$ a funkce $F(x)$ je spojitě diferencovatelná na $X \subset R^n$. Pak:*

(a) Každý hromadný bod posloupnosti $\{x_k\}$, generovaný Algoritmem 4.1, je řešením rovnice (1157).
(b) Jestliže $\sigma < 1/2$ a $\omega_k \to 0$, pak $x_k \to x^*$ superlineárně.

Důkaz (a) Jelikož f je silně BD-regulární na $X \subset R^n$ a množina X je kompaktní, existuje konstanta $c > 0$ tak, že v každém iteračním kroku platí $\|J_k^{-1}\| \leq c$. Krok 2 algoritmu je tedy dobře definován a podle (1166) platí

$$\|d_k\| = \|J_k^{-1}(J_k d_k + f(x_k)) - J_k^{-1} f(x_k)\| \leq (1+\omega)\|J_k^{-1}\|\|f_k\| \leq c(1+\omega)\sqrt{2F(x_1)}. \tag{1168}$$

Ukážeme, že i Krok 3 algoritmu je dobře definován. Předpokládejme naopak, že pro libovolný exponent i platí

$$F(x_k + \varrho^i d_k) - F(x_k) > -2\sigma\varrho^i F(x_k),$$

neboli v limitě

$$F'(x_k, d_k) \geq -2\sigma F(x_k).$$

Jelikož F je spojitě diferencovatelná, podle důsledku 40 a podle (1166) platí

$$\begin{aligned} F'(x_k, d_k) &= (\nabla F(x_k))^T d_k = f^T(x_k) J_k d_k \\ &= f^T(x_k) f(x_k) + f^T(x_k) J_k d_k - f^T(x_k) f(x_k) \\ &\leq \|f(x_k)\|\|f(x_k) + J_k d_k\| - \|f(x_k)\|^2 \\ &\leq (\omega - 1)\|f(x_k)\|^2 = -2(1-\omega)F(x_k). \end{aligned} \tag{1169}$$

Jelikož platí $F(x_k) \neq 0$ (v opačném případě by došlo k ukončení výpočtu v Kroku 2 algoritmu) dostaneme porovnáním obou nerovností $\sigma \geq 1 - \omega$, což je ve sporu s předpokladem $\sigma < 1 - \omega$. Uvažujme nyní posloupnost $\{x_k\}$ generovanou Algoritmem 4.1. Jelikož $x_k \in X$ a $X \subset R^n$ je kompaktní, musí existovat alespoň jeden hromadný bod $x^* \in X$ posloupnosti $\{x_k\}$. Existuje tedy podmnožina $\mathcal{K}$ množiny všech indexů taková, že $x_k \overset{\mathcal{K}}{\to} x^*$. Vyšetříme nyní dva případy.

(α) Předpokládejme nejprve, že $t_k \geq \tau > 0\ \forall k \in \mathcal{K}$. Pak podle (1167) platí

$$\begin{aligned} F(x_1) &\geq F(x_1) - \lim_{k\to\infty} F(x_k) = \sum_{k=1}^{\infty}(F(x_k) - F(x_{k+1})) \\ &\geq \sum_{k=1}^{\infty} 2\sigma t_k F(x_k) \geq 2\tau\sigma \sum_{k\in\mathcal{K}} F(x_k), \end{aligned}$$

takže nutně $F(x_k) \overset{\mathcal{K}}{\to} 0$, což spolu s $x_k \overset{\mathcal{K}}{\to} x^*$ dává $F(x^*) = 0$ (neboť funkce F je spojitá).

(β) Předpokládejme nyní, že $t_k \overset{\mathcal{K}_1}{\to} 0$ pro nějakou podmnožinu $\mathcal{K}_1 \subset \mathcal{K}$. Odtud plyne, že $i_k \overset{\mathcal{K}_1}{\to} \infty$, takže pro dostatečně velké indexy $k \in \mathcal{K}_1$ platí $i_k > 0$ a jelikož (1167) neplatí pro $i = i_k - 1$, můžeme s použitím věty o střední hodnotě psát

$$(\nabla F(x'_k))^T d_k = \frac{F\left(x_k + \frac{t_k}{\varrho} d_k\right) - F(x_k)}{\frac{t_k}{\varrho}} > -2\sigma F(x_k),$$

kde $x'_k \in (x_k, x_k + (t_k/\varrho)d_k)$. Jelikož posloupnost $\{\|d_k\|\}_{\mathcal{K}_1}$ je podle (1168) omezená, má tato posloupnost alespoň jeden hromadný bod d^*. Existuje tedy podmnožina $\mathcal{K}_2 \subset \mathcal{K}_1$ taková, že $d_k \overset{\mathcal{K}_2}{\to} d^*$, což spolu s $x_k \overset{\mathcal{K}_2}{\to} x^*$ a $t_k \overset{\mathcal{K}_2}{\to} 0$ (takže $x'_k \overset{\mathcal{K}_2}{\to} x^*$) v limitě dává

$$(\nabla F(x^*))^T d^* \geq -2\sigma F(x^*).$$

Z druhé strany podle (1169) platí $(\nabla F(x_k))^T d_k \leq -2(1-\omega)F(x_k)$, což v limitě dává

$$(\nabla F(x^*))^T d^* \leq -2(1-\omega)F(x^*).$$

Jelikož podle předpokladu platí $\sigma < 1-\omega$, dostaneme porovnáním obou nerovností $F(x^*) = 0$. Dokázali jsme tedy, že pokud x^* je hromadným bodem posloupnosti generované algoritmem, platí $F(x^*) = 0$ a tedy i $f(x^*) = 0$.

(b) Nechť $\mathcal{K}$ je indexová množina použitá v části (a) důkazu. Našim cílem je ukázat, že pro dostatečně velké indexy $k \in \mathcal{K}$ platí $x_{k+1} = x_k + d_k$, a pak použít indukční postup z důkazu věty 355. Jelikož $x_k \overset{\mathcal{K}}{\to} x^*$, $\omega_k \overset{\mathcal{K}}{\to} 0$ (a $\Delta_k = 0$), jsou pro dostatečně velké indexy $k \in \mathcal{K}$ splněny předpoklady použité v důkazu věty 355 ($x_k \in \mathcal{B}(x^*, \varepsilon)$ a $\omega_k \leq 1/(10cL)$), takže pro bod $x_k + d_k$ platí (1163) (s $x_k + d_k$ místo x_{k+1}) a

$$\lim_{k \overset{\mathcal{K}}{\to} \infty} \frac{\|x_k + d_k - x^*\|}{\|x_k - x^*\|} = 0,$$
$$\lim_{k \overset{\mathcal{K}}{\to} \infty} \frac{\|f(x_k + d_k)\|}{\|f(x_k)\|} = 0.$$

Jelikož $\sigma < 1/2$, existuje index $\overline{k} \in \mathcal{K}$ takový, že $\|f(x_k + d_k)\| \leq (1-2\sigma)\|f(x_k)\|$, pokud $k \in \mathcal{K}$ a $k \geq \overline{k}$. Pro tyto indexy platí

$$\begin{aligned} \frac{F(x_k + d_k) - F(x_k)}{F(x_k)} &= \frac{(\|f(x_k + d_k)\| - \|f(x_k)\|)\,(\|f(x_k + d_k)\| + \|f(x_k)\|)}{\|f(x_k)\|^2} \\ &\leq \frac{\|f(x_k + d_k)\| - \|f(x_k)\|}{\|f(x_k)\|} \leq -2\sigma, \end{aligned}$$

takže podmínka (1167) je splněna s $i_k = 0$. Platí tedy $x_{k+1} = x_k + d_k$ a vzhledem k (1163) můžeme množinu $\mathcal{K}$ formálně doplnit o index $k+1$. Pokračujeme-li takto pro další hodnoty indexu, vidíme (tak jako v důkazu věty 355), že $x_k \to x^*$ superlineárně. □

Poznámka 441. Požadavek spojité diferencovatelnosti funkce $F = (1/2)f^T f$ se zdá být na první pohled nerealistický, neboť zobrazení f není spojitě diferencovatelné. Ve skutečnosti je však tento požadavek splněn v mnoha významných aplikacích.

Poznámka 442. V Algoritmu 4.1 se používá matice $J_k \in \partial_B f(x_k)$. Jelikož zobrazení f je podle Rademacherovy věty diferencovatelné skoro všude, platí obvykle $x_k \notin \Omega_f$, takže $J_k = \nabla f(x_k)$. Pokud $x_k \in \Omega_f$, bývá výpočet $J_k \in \partial_B f(x_k)$ obtížnější. Z definice 121 plyne, že

$$\partial_B f(x_k) \subset [\partial_B f_1(x_k), \ldots, \partial_B f_n(x_k)]^T \overset{\Delta}{=} \partial_b f(x_k),$$

přičemž určení $\partial_b f(x_k)$ bývá obvykle snadnější než určení $\partial_B f(x_k)$. Proto se naskýtá otázka, zda by nebylo možné volit $J_k \in \partial_b f(x_k)$. Odpověď na tuto otázku je kladná. Nechť $J \in \partial_b f(x)$. Protože funkce $f_1, \ldots, f_n$ jsou podle důsledku 44 polohladké, podle důsledku 46 platí

$$\begin{aligned} f_1(x+h) - f_1(x) - e_1^T Jh &= o(\|h\|), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\ f_n(x+h) - f_1(x) - e_n^T Jh &= o(\|h\|) \end{aligned}$$

a n je konečné, zůstává klíčový vztah (1156) v platnosti i pro $J \in \partial_b f(x)$ a v důkazech věty 355 a věty 356 se v podstatě nic nezmění.

16.2 Aplikace nehladkých rovnic

Definice 125. *Nechť zobrazení $p : R^n \to R^n$ je spojitě diferencovatelné. Pak úlohou nelineární komplementarity (NCP) rozumíme nalezení bodu $x^* \in R^n_+$ takového, že $p(x^*) \in R^n_+$ a $(x^*)^T p(x^*) = 0$, tedy*

$$x_i^* \geq 0, \quad p_i(x^*) \geq 0, \quad x_i^* p_i(x^*) = 0 \tag{1170}$$

pro libovolný index $1 \leq i \leq n$.

Úlohu nelineární komplementarity lze snadno převést na řešení ekvivalentní soustavy polohladkých rovnic $f(x) = 0$, kde

$$f(x) = \begin{bmatrix} \psi(x_1, p_1(x)) \\ \ldots\ldots\ldots\ldots \\ \psi(x_n, p_n(x)) \end{bmatrix} \tag{1171}$$

a $\psi : R^n \to R$ je polohladká funkce, pro kterou platí $\psi(u_1, u_2) = 0$ právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$. Funkce, která má tuto vlastnost se nazývá NCP funkcí. Vyšetříme tři základní NCP funkce, Pangovu funkci

$$\psi(u) = \min(u_1, u_2), \tag{1172}$$

Fischerovu-Burmeisterovu funkci

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2) \tag{1173}$$

a Kanzowovu funkci

$$\psi(u) = u_1 u_2 - \frac{1}{2}\min^2(0, u_1 + u_2). \tag{1174}$$

Lemma 126. *Funkce $\psi : R^2 \to R$ definovaná vztahem (1172) je spojitě diferencovatelná v $R^2 \backslash \mathcal{L}$, kde $\mathcal{L} = \{u \in R^2 : u_1 = u_2\}$, a polohladká na $\mathcal{L}$, přičemž $\partial_B \psi(u) = \{e_1, e_2\}$ a $\partial\psi(u) = \operatorname{conv}\{e_1, e_2\}$ pro $u \in \mathcal{L}$, kde $e_1 = [1, 0]^T$ a $e_2 = [0, 1]^T$. Pro každý vektor $g(u) \in \partial\psi(u)$, $u \in \mathcal{L}$, a $g(u) = \nabla\psi(u)$, $u \notin \mathcal{L}$, platí $\|g(u)\| \geq \sqrt{2}/2$. Rovnost $\psi(u_1, u_2) = 0$ nastává právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$. Druhá mocnina funkce ψ není spojitě diferencovatelná na $\mathcal{L}\backslash\{0\}$.*

Důkaz Spojitá diferencovatelnost funkce ψ v $R^2\backslash\mathcal{L}$ je zřejmá: Pokud $u_1 > u_2$, platí $\psi(u) = u_1$, $\nabla\psi(u) = e_1$, a pokud $u_2 > u_1$, platí $\psi(u) = u_2$, $\nabla\psi(u) = e_2$. Vztahy pro $\partial_B\psi(u)$, $\partial\psi(u)$, $u \in \mathcal{L}$, plynou z věty 342 a věty 346. Polohladkost funkce ψ na $\mathcal{L}$ plyne z důsledku 43. Vztah $\|g(u)\| \geq \sqrt{2}/2$, $u \in R^2$, plyne z toho, že $g(u)$ leží vždy na úsečce spojující body e_1 a e_2. To, že rovnost $\psi(u_1, u_2) = 0$ nastává právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$, plyne bezprostředně ze vztahu (1172). Podle důsledku 40, kde $m = 1$, platí $\partial\psi^2(u) = 2\psi(u)\partial\psi(u)$. Nechť $u \in \mathcal{L}\backslash\{0\}$. Pak $\psi(u) \neq 0$ a jelikož množina $\partial\psi(u)$ není jednobodová, není ani množina $\partial\psi^2(u)$ jednobodová. □

Věta 357. *Nechť zobrazení $p : R^n \to R^n$ je spojitě diferencovatelné v bodě $x \in R^n$. Nechť $f : R^n \to R^n$ je zobrazení definované předpisem (1171), kde $\psi : R^2 \to R$ je funkce definovaná předpisem (1172). Pak:*

(a) Zobrazení f je polohladké v bodě x.
(b) Platí $\partial_B f(x) \subset [\partial_B f_1(x), \ldots, \partial_B f_n(x)]^T$, kde

$$\partial_B f_i(x) = \nabla f_i(x) = e_i, \qquad x_i > p_i(x), \tag{1175}$$

$$\partial_B f_i(x) = \nabla f_i(x) = \nabla p_i(x), \qquad x_i < p_i(x), \tag{1176}$$

a

$$\partial_B f_i(x) = \{e_i, \nabla p_i(x)\}, \qquad x_i = p_i(x). \tag{1177}$$

Důkaz (a) Polohladkost zobrazení f plyne z věty 348 a věty 351, neboť $f_i(x) = \psi(x_i, p_i(x))$, funkce ψ je polohladká podle lemmatu 126 a zobrazení p je spojitě diferencovatelné.

(b) Podle lemmatu 126 je funkce $\psi(x_i, p_i)$ spojitě diferencovatelná, pokud $x_i \neq p_i$. Vztahy (1175) a (1176) plynou ze vztahů $\nabla\psi(u) = e_1$, $u_1 > u_2$, a $\nabla\psi(u) = e_2$, $u_2 > u_1$ (důkaz lemmatu 126) použitím pravidla pro derivování složené funkce. Vztah (1177) plyne z věty 342, neboť z $x \notin \Omega_{f_i}$ plyne $x_i \neq p_i(x)$, neboli $\nabla f_i(x) \in \{e_i, \nabla p_i(x)\}$. □

Lemma 127. *Funkce $\psi : R^2 \to R$ definovaná vztahem (1173) je spojitě diferencovatelná v $R^2\backslash\{0\}$ a polohladká v bodě 0, přičemž $\partial_B\psi(0) = S(-e,1)$ a $\partial\psi(0) = \overline{\mathcal{B}(-e,1)}$, kde $e = [1,1]^T$ ($S(-e,1)$ je kružnice a $\overline{\mathcal{B}(-e,1)} = \operatorname{conv} S(-e,1)$ kruh se středem $-e$ a poloměrem 1). Pro každý vektor $g(u) \in \partial\psi(u)$, $u = 0$, a $g(u) = \nabla\psi(u)$, $u \neq 0$, platí $\|g(u)\| \geq \sqrt{2} - 1$. Rovnost $\psi(u_1, u_2) = 0$ nastává právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$. Druhá mocnina funkce ψ je spojitě diferencovatelná v R^2.*

Důkaz Spojitá diferencovatelnost funkce ψ v $R^2\backslash\{0\}$ je zřejmá: Pro $u \in R^2\backslash\{0\}$ platí

$$\nabla\psi(u) = \begin{bmatrix} \frac{u_1}{\sqrt{u_1^2+u_2^2}} - 1 \\ \frac{u_2}{\sqrt{u_1^2+u_2^2}} - 1 \end{bmatrix}. \tag{1178}$$

Polohladkost funkce ψ v bodě 0 plyne z věty 350, neboť funkce ψ je konvexní (je součtem euklidovské normy $\sqrt{u_1^2 + u_2^2}$ a lineární funkce $-(u_1 + u_2)$). Uvažujme posloupnost $\{u_i\}$, kde $u_i = [t_i \cos\varphi_i, t_i \sin\varphi_i]^T$ a $t_i \downarrow 0$. Pak platí $\nabla\psi(u_i) = [\cos\varphi_i - 1, \sin\varphi_i - 1]^T$ a posloupnost $\{\nabla\psi(u_i)\}$ má limitu $[\cos\varphi - 1, \sin\varphi - 1]^T$ právě tehdy, když $\varphi_i \to \varphi$. Odtud plyne, že

$$\partial_B\psi(0) = \bigcup_{\varphi\in[0,2\pi]} [\cos\varphi - 1, \sin\varphi - 1]^T = S(-e,1)$$

a

$$\partial\psi(0) = \operatorname{conv} \partial_B\psi(0) = \operatorname{conv} S(-e,1) = \overline{\mathcal{B}(-e,1)}.$$

Vztah $\|g(u)\| \geq \sqrt{2} - 1$, $u \in R^2$, plyne z toho, že $g(u)$ leží vždy na kružnici $S(-e,1)$. Pokud $u_1 < 0$, platí

$$\psi(u) = \sqrt{|u_1|^2 + u_2^2} + |u_1| - u_2 \geq |u_2| + |u_1| - u_2 > 0$$

(stejný výsledek dostaneme pro $u_2 < 0$). Pokud $u_1 > 0$, $u_2 > 0$, platí

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2) < \sqrt{u_1^2 + 2u_1u_2 + u_2^2} - (u_1 + u_2) = 0.$$

Pokud $u_1 = 0$ a $u_2 > 0$, platí

$$\psi(u) = |u_2| - u_2 = 0$$

(stejný výsledek dostaneme pro $u_1 > 0$ a $u_2 = 0$). Rovnost $\psi(0) = 0$ je zřejmá. Druhou mocninu funkce ψ můžeme vyjádřit ve tvaru

$$\psi^2(u) = u_1^2 + u_2^2 + (u_1 + u_2)^2 - 2(u_1 + u_2)\sqrt{u_1^2 + u_2^2}.$$

Tato funkce je spojitě diferencovatelná v $R^2\backslash\{0\}$ a je spojitě diferencovatelná v bodě 0 právě tehdy, je-li funkce $\overline{\psi}(u) = (u_1 + u_2)\sqrt{u_1^2 + u_2^2}$ spojitě diferencovatelná v bodě 0. Ale

$$\lim_{\|u\|\to 0} \frac{\overline{\psi}(u) - \overline{\psi}(0)}{\|u\|} = \lim_{\|u\|\to 0} (u_1 + u_2)\frac{\sqrt{u_1^2 + u_2^2}}{\sqrt{u_1^2 + u_2^2}} = 0,$$

takže $\overline{\psi}$ je diferencovatelná v bodě 0 a platí $\nabla\overline{\psi}(0) = 0$. Spojitost parciální derivace $\partial\overline{\psi}/\partial u_1$ v bodě 0 plyne z nerovnosti

$$\left|\frac{\partial\overline{\psi}(u)}{\partial u_1}\right| = \left|\frac{u_1}{\sqrt{u_1^2+u_2^2}}(u_1+u_2)+\sqrt{u_1^2+u_2^2}\right|$$
$$\leq \frac{|u_1|}{\sqrt{u_1^2+u_2^2}}|u_1+u_2|+\sqrt{u_1^2+u_2^2} \leq |u_1+u_2|+\sqrt{u_1^2+u_2^2}$$

a z toho, že pravá strana této nerovnosti konverguje k nule pokud $u \to 0$ (stejný výsledek dostaneme pro parciální derivaci $\partial\overline{\psi}/\partial u_2$). □

Věta 358. *Nechť zobrazení $p : R^n \to R^n$ je spojitě diferencovatelné v bodě $x \in R^n$. Nechť $f : R^n \to R^n$ je zobrazení definované předpisem (1171), kde $\psi : R^2 \to R$ je funkce definovaná předpisem (1173). Pak:*

(a) Zobrazení f je polohladké v bodě x.
(b) Platí $\partial_B f(x) \subset [\partial_B f_1(x), \ldots, \partial_B f_n(x)]^T$, kde

$$\partial_B f_i(x) = \nabla f_i(x) = \left(\frac{x_i}{\sqrt{x_i^2+p_i^2(x)}}-1\right)e_i + \left(\frac{p_i(x)}{\sqrt{x_i^2+p_i^2(x)}}-1\right)\nabla p_i(x), \tag{1179}$$

pokud $x_i^2 + p_i^2(x) \neq 0$ a

$$\partial_B f_i(x) = \bigcup_{\varphi\in[0,2\pi]} [(\cos\varphi - 1)e_i + (\sin\varphi - 1)\nabla p_i(x)], \tag{1180}$$

pokud $x_i^2 + p_i^2(x) = 0$.
(c) Funkce $F = (1/2)f^T f$ je spojitě diferencovatelná v bodě x.

Důkaz (a) Polohladkost zobrazení f plyne z věty 348 a věty 351, neboť $f_i(x) = \psi(x_i, p_i(x))$, funkce ψ je polohladká podle lemmatu 127 a zobrazení p je spojitě diferencovatelné.

(b) Podle lemmatu 127 je funkce $\psi(x_i, p_i)$ spojitě diferencovatelná, pokud $x_i^2 + p_i^2 \neq 0$. Vztah (1179) plyne z (1178) použitím pravidla pro derivování složené funkce. V případě, že $x_i^2 + p_i^2 = 0$, můžeme použít stejný limitní proces jako v lemmatu 127, takže

$$\partial_B f_i(x) = [e_i, \nabla p_i(x)]\partial_B\psi(0) = [e_i, \nabla p_i(x)]S(-e, 1),$$

což dává (1180).

(c) Platí

$$F(x) = \frac{1}{2}f^T(x)f(x) = \frac{1}{2}\sum_{i=1}^{n}\psi^2(x_i, p_i(x)).$$

Zobrazení p je spojitě diferencovatelné. Podle lemmatu 127 je druhá mocnina funkce ψ spojitě diferencovatelná, takže i funkce F je spojitě diferencovatelná. □

Lemma 128. *Funkce $\psi : R^2 \to R$ definovaná vztahem (1174) je spojitě diferencovatelná v R^2. Rovnost $\psi(u_1, u_2) = 0$ nastává právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$. Platí $\nabla\psi(0) = 0$.*

Důkaz Pro funkci (1174) platí

$$\psi(u) = u_1u_2, \qquad \nabla\psi(u) = \begin{bmatrix} u_2 \\ u_1 \end{bmatrix}, \quad u_1+u_2 > 0, \tag{1181}$$

$$\psi(u) = -\frac{1}{2}(u_1^2+u_2^2), \qquad \nabla\psi(u) = \begin{bmatrix} -u_1 \\ -u_2 \end{bmatrix}, \quad u_1+u_2 < 0. \tag{1182}$$

Odtud je vidět, že funkce $\psi(u)$ je spojitá a spojitě diferencovatelná i v bodech, kde $u_1 + u_2 = 0$, přičemž $\nabla\psi(0) = 0$. Jestliže $u_1 + u_2 \geq 0$, pak $\psi(u) = u_1 u_2 = 0$ právě tehdy, když $u_1 = 0$, a tedy $u_2 \geq 0$, nebo $u_2 = 0$, a tedy $u_1 \geq 0$. Jestliže $u_1 + u_2 < 0$, pak buď $u_1 < 0$ nebo $u_2 < 0$, takže $\psi(u) = -(u_1^2 + u_2^2)/2 < 0$. □

Věta 359. *Nechť zobrazení $p : R^n \to R^n$ je spojitě diferencovatelné v bodě $x \in R^n$. Nechť $f : R^n \to R^n$ je zobrazení definované předpisem (1171), kde $\psi : R^2 \to R$ je funkce definovaná předpisem (1174). Pak:*

(a) Zobrazení f je spojitě diferencovatelné v bodě x.
(b) Platí $\nabla f(x) = [\nabla f_1(x), \ldots, \nabla f_n(x)]^T$, kde

$$\begin{aligned} \nabla f_i(x) &= p_i(x)e_i + x_i \nabla p_i(x), \qquad x_i + p_i(x) \geq 0, && (1183) \\ \nabla f_i(x) &= -x_i e_i - p_i(x)\nabla p_i(x), \qquad x_i + p_i(x) < 0. && (1184) \end{aligned}$$

Důkaz (a) Spojitá diferencovatelnost plyne z toho, že $f_i(x) = \psi(x_i, p_i(x))$, funkce ψ je spojitě diferencovatelná podle lemmatu 128 a zobrazení p je spojitě diferencovatelné.

(b) Vztahy (1183) a (1184) plynou ze vztahů (1181) a (1182) použitím pravidla pro derivování složené funkce (funkce $\psi(u)$ je spojitě diferencovatelná i v bodě $u \in R_n$ takovém, že $u_1 + u_2 = 0$). □

Poznámka 443. Fischerovu-Burmeistrovu funkci lze zobecnit tak že místo eukleidovské normy vystupující v (1173) použijeme jinou vhodnou normu. Dostaneme tak funkci

$$\psi(u) = \|u\| - (u_1 + u_2). \tag{1185}$$

Vhodné jsou například normy $\|u\|_p = (u_1^p + u_2^p)^{1/p}$, $p > 1$ a $\|u\|_\infty = \max(|u_1|, |u_2|)$. Nelze však použít normu $\|u\|_1 = |u_1| + |u_2|$. Předpokládejme bez újmy na obecnosti, že $u_1 \geq u_2$. Pokud $u_2 > 0$, platí

$$\begin{aligned} \|u\|_p - (u_1 + u_2) &= (|u_1|^p + |u_2|^p)^{1/p} - (u_1 + u_2) < ((|u_1| + |u_2|)^p)^{1/p} - (|u_1| + |u_2|) = 0, \\ \|u\|_\infty - (u_1 + u_2) &= \max(|u_1|, |u_2|) - (u_1 + u_2) = |u_1| - (|u_1| + |u_2|) = -|u_2| < 0, \\ \|u\|_1 - (u_1 + u_2) &= |u_1| + |u_2| - (u_1 + u_2) = |u_1| + |u_2| - (|u_1| + |u_2|) = 0. \end{aligned}$$

První nerovnost plyne ze zobecněné binomické věty podle které platí $|u_1|^p + |u_2|^p < (|u_1| + |u_2|)^p$. Poslední rovnost ukazuje, že normu $\|u\|_1$ nelze použít ke konstrukci NCP funkce. Pokud $u_2 = 0$, přejdou uvedené nerovnosti v rovnosti a platí $\psi(u) = 0$. Pokud $u_2 < 0$, můžeme použít zobecněné Schwarzovy nerovnosti

$$\begin{aligned} |u^T v| &\leq \|u\|_p \|v\|_q, \quad 1/p + 1/q = 1, \\ |u^T v| &\leq \|u\|_\infty \|v\|_1, \end{aligned}$$

které přejdou v rovnosti pouze tehdy, jsou-li vektory u a v lineárně závislé. Nechť $e_1 = [1, 0]^T$. Jelikož předpokládáme, že $u_2 < 0$, jsou vektory u a e_1 lineárné nezávislé, a jelikož $\|e_1\|_q = 1$ a $\|e_1\|_1 = 1$ (lze to snadno ověřit dosazením do definičních vztahů pro uvedené normy), můžeme psát

$$\begin{aligned} |u_1| &= |u^T e_1| < \|u\|_p \, \|e_1\|_q = \|u\|_p, \\ |u_1| &= |u^T e_1| < \|u\|_\infty \|e_1\|_1 = \|u\|_\infty, \end{aligned}$$

takže

$$\begin{aligned} \|u\|_p - (u_1 + u_2) &> |u_1| - |u_1| + |u_2| = |u_2| > 0, \\ \|u\|_\infty - (u_1 + u_2) &= |u_1| - |u_1| + |u_2| = |u_2| > 0. \end{aligned}$$

V obou případech tedy platí $\psi(u) = 0$ právě tehdy, když $u_1 \geq 0$, $u_2 \geq 0$ a $u_1 u_2 = 0$.

Věta 358 naznačuje jednu z možností jak řešit úlohy nelineární komplementarity. Úloha nelineární komplementarity se převede na ekvivalentní soustavu nehladkých rovnic (1171), které se řeší pomocí Algoritmu 4.1. Podle poznámky 442 lze volit $J_k \in \partial_b f(x_k)$, kde množinu $\partial_b f(x_k) = [\partial_B f_1(x_k), \ldots, \partial_B f_n(x_k)]^T$ lze určit podle (1179)-(1180). Funkce $F = (1/2)f^T f$ používaná při výběru délky kroku je v tomto případě spojitě diferencovatelná.

Ukážeme ještě jednu aplikaci nehladkých rovnic. Uvažujme úlohu nelineárního programování: Najít minimum spojitě diferencovatelné funkce $F : R^n \to R$ na množině určené omezeními $c_i(x) \leq 0$, $1 \leq i \leq m$, kde $c : R^n \to R^m$, je spojitě diferencovatelné zobrazení. Jsou-li splněny podmínky regularity, musí řešení této úlohy vyhovovat podmínkám

$$\nabla F(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0, \tag{1186}$$

$$\left.\begin{array}{ll} -c_i(x) \geq 0, & \lambda_i \geq 0, \\ \lambda_i c_i(x) = 0, & 1 \leq i \leq m \end{array}\right\} \tag{1187}$$

(tvrzení 4). Podmínky (1187) jsou v podstatě podmínkami nelineární komplementarity (1170). Můžeme tedy sestavit soustavu $n + m$ nehladkých rovnic

$$f(x, \lambda) \triangleq \left[\begin{array}{l} \nabla F(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) \\ \psi(\lambda_1, -c_1(x)) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \psi(\lambda_m, -c_m(x)) \end{array}\right] = 0, \tag{1188}$$

kde ψ je Fischerova-Burmeisterova funkce (1173). Zobrazení $f : R^{n+m} \to R^{n+m}$ je polohladké a funkce $F = (1/2)F^T F$ je spojitě diferencovatelná, takže soustavu rovnic (1188) lze řešit pomocí Algoritmu 4.1.

17 Metody pro nehladkou optimalizaci

17.1 Svazkové metody

Budeme předpokládat, že funkce $F : R^n \to R$ je lokálně lipschitzovská a že umíme v každém bodě $x \in R^n$ spočítat nějaký subgradient $g \in \partial F(x)$. Jelikož lokálně lipschitzovská funkce je podle Rademacherovy věty diferencovatelná skoro všude, platí obvykle $g = \nabla F(x)$. Zvláštností úloh nehladké optimalizace je, že se gradient $\nabla F(x)$ může měnit skokem a že nemusí být malý v okolí extrému funkce F. Z tohoto důvodu nestačí chování funkce F vystihnout hodnoty $F_k = F(x_k)$, $g_k \in \partial F(x_k)$, v jediném bodě x_k, ale je zapotřebí celý svazek hodnot

$$F_j = F(y_j), \quad g_j \in \partial F(y_j), \tag{1189}$$

získaných v pokusných bodech y_j, $j \in \mathcal{J}_k \subset \{1, \ldots, k\}$, který slouží ke konstrukci po částech lineární funkce

$$F_L^k(x) = \max_{j \in \mathcal{J}_k}(F_j + g_j^T(x - y_j)) = \max_{j \in \mathcal{J}_k}(F_j^k + g_j^T(x - x_k)) = \max_{j \in \mathcal{J}_k}(F(x_k) + g_j^T(x - x_k) - \alpha_j^k),$$

kde

$$F_j^k = F_j + g_j^T(x_k - y_j), \tag{1190}$$

$$\alpha_j^k = F(x_k) - F_j^k \tag{1191}$$

pro $j \in \mathcal{J}_k$. Tato po částech lineární funkce je v konvexním případě majorizována funkcí F.

Věta 360. *Nechť funkce $F : R^n \to R$ je konvexní. Pak pro libovolný index k platí $\alpha_j^k \geq 0$ $\forall j \in \mathcal{J}_k$ a $F(x) \geq F_L^k(x)$ $\forall x \in R^n$.*

Důkaz Jelikož $g_j \in \partial F(y_j)$, platí podle věty 323 (d) $F(x) \geq F_j + g_j^T(x - y_j)$ $\forall j \in \mathcal{J}_k$, takže podle (1190) dostaneme $F(x_k) \geq F_j^k$, což podle (1191) dává $\alpha_j^k \geq 0$. Navíc

$$F(x) \geq \max_{j \in \mathcal{J}_k}(F_j + g_j^T(x - y_j)) = F_L^k(x).$$

□

V případě, že funkce F není konvexní, věta 360 neplatí. Abychom v tomto případě zaručili vhodnost po částech lineárního modelu $F_L^k(x)$, je třeba čísla α_j^k, $j \in \mathcal{J}_k$, definovat jiným způsobem. Jednou z možností je pro $j \in \mathcal{J}_k$ položit

$$\alpha_j^k = \max\left(|F(x_k) - F_j^k|, \gamma\|x_k - y_j\|^\nu\right),$$

kde $\gamma \geq 0$ a $\nu \geq 1$. Jelikož by však bylo nutné ukládat body y_j, $j \in \mathcal{J}_k$, využívá se toho, že pro $j \in \mathcal{J}_k$ platí

$$\|x_k - y_j\| \leq \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \triangleq s_j^k \tag{1192}$$

a čísla α_j^k se určují podle vzorce

$$\alpha_j^k = \max\left(|F(x_k) - F_j^k|, \gamma(s_j^k)^\nu\right), \quad j \in \mathcal{J}_k. \tag{1193}$$

Funkce F_L^k není sama o sobě vhodná k určení nové aproximace minima, neboť její minimum nemusí existovat (F_L^k je po částech lineární) a pokud existuje, může být příliš daleko od minima funkce F. Proto se k funkci F_L^k přidává tlumící kvadratický člen. Dostáváme tak po částech kvadratickou funkci

$$\begin{aligned} F_Q^k(x) &= \frac{1}{2}(x-x_k)^T G_k(x-x_k) + F_L^k(x) \\ &= \frac{1}{2}(x-x_k)^T G_k(x-x_k) + \max_{j\in\mathcal{J}_k}(F(x_k) + g_j^T(x-x_k) - \alpha_j^k), \end{aligned}$$

kde G_k je nějaká symetrická positivně definitní matice. Tato po částech kvadratická funkce může být interpretována různým způsobem buď k určení směrového vektoru v metodách spádových směrů nebo k určení oblasti přijatelnosti v metodách s lokálně omezeným krokem. V tomto textu se omezíme na metody spádových směrů.

Protože je z praktických důvodů možné pracovat pouze s omezenými svazky, kdy $|\mathcal{J}_k| \leq m$ ($|\mathcal{J}_k|$ je mohutnost množiny $\mathcal{J}_k$), určuje se množina $\mathcal{J}_k$ obvykle tak, že $\mathcal{J}_k = \{1,\ldots,k\}$, pokud $k \leq m$, a $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\}\backslash\{k+1-m\}$, pokud $k \geq m$. Poznamenejme, že to není jediný a dokonce ani nejvhodnější způsob jak určovat svazky, je to však způsob jednoduchý, který vyhovuje všem teoretickým požadavkům, takže se ho v tomto textu přidržíme.

Jestliže $\mathcal{J}_k \neq \{1,\ldots,k\}$, je třeba používat agregované hodnoty, které v sobě kumulují informace z předchozích iteračních kroků. Agregace bude podrobně popsána později (definiční vztahy (1200), (1206), (1207) a transformační vztahy (1211)). Zde pouze uvedeme, že v bodě x_k máme k disposici hodnoty $F_a^k \in R$, $g_a^k \in R^n$, $s_a^k \in R$ reprezentující jistou lineární funkci, která se přidává k lineárním funkcím obsaženým ve svazku a že v průběhu k-tého iteračního kroku se řešením úlohy kvadratického programování určují nové hodnoty $\tilde{F}_a^k \in R$, $\tilde{g}_a^k \in R^n$, $\tilde{s}_a^k \in R$, které se pak transformují do bodu x_{k+1}.

Použijeme-li agregované hodnoty, má po částech kvadratická funkce tvar

$$F_Q^k(x) = \frac{1}{2}(x-x_k)^T G_k(x-x_k) + \max(F_L^k(x), F(x_k) + (x-x_k)^T g_a^k - \alpha_a^k),$$

kde

$$\alpha_a^k = \max\left(|F(x_k) - F_a^k|, \gamma(s_a^k)^\nu\right). \tag{1194}$$

Minimum této funkce lze vyjádřit ve tvaru $x_{k+1} = x_k + d_k$, kde směrový vektor d_k je řešením úlohy kvadratického programování: Minimalizovat funkci

$$\frac{1}{2}d^T G_k d + v \tag{1195}$$

na množině určené omezeními

$$-\alpha_j^k + d^T g_j \leq v, \quad j \in \mathcal{J}_k, \tag{1196}$$

$$-\alpha_a^k + d^T g_a^k \leq v, \tag{1197}$$

(minimalizuje se přes všechny dvojice $(d,v) \in R^{n+1}$ vyhovující nerovnostem (1196), (1197)).

Věta 361. *Řešení úlohy (1195)-(1197) lze vyjádřit ve tvaru*

$$d_k = -G_k^{-1}\tilde{g}_a^k, \tag{1198}$$

$$v_k = -d_k^T G_k d_k - \tilde{\alpha}_a^k, \tag{1199}$$

kde

$$\tilde{g}_a^k = \sum_{j\in\mathcal{J}_k} \lambda_j^k g_j + \lambda_a^k g_a^k, \tag{1200}$$

$$\tilde{\alpha}_a^k = \sum_{j\in\mathcal{J}_k} \lambda_j^k \alpha_j^k + \lambda_a^k \alpha_a^k \tag{1201}$$

a kde Lagrangeovy multiplikátory λ_j^k, $j \in \mathcal{J}_k$, λ_a^k, jsou řešením duální úlohy kvadratického programování: Minimalizovat funkci

$$\frac{1}{2}\left(\sum_{j\in\mathcal{J}_k}\lambda_j g_j + \lambda_a g_a^k\right)^T G_k^{-1}\left(\sum_{j\in\mathcal{J}_k}\lambda_j g_j + \lambda_a g_a^k\right) + \sum_{j\in\mathcal{J}_k}\lambda_j\alpha_j^k + \lambda_a\alpha_a^k \tag{1202}$$

na množině určené omezeními

$$\left.\begin{array}{ll} \lambda_j \geq 0, & j \in \mathcal{J}_k, \quad \lambda_a \geq 0, \\ \sum_{j\in\mathcal{J}_k}\lambda_j + \lambda_a = 1. & \end{array}\right\} \tag{1203}$$

Minimální hodnota funkce (1202), odpovídající řešení úlohy (1202)-(1203), je

$$w_k = \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \tilde{\alpha}_a^k = -v_k - \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k. \tag{1204}$$

Důkaz Jelikož matice G_k je positivně definitní, je funkce (1195) konvexní. Omezení (1196)-(1197) jsou lineární a tudíž také konvexní, takže pár $(d_k, v_k) \in R^{n+1}$ je podle tvrzenví 4 řešením úlohy (1195)-(1197) právě tehdy, existují-li Lagrangeovy multiplikátory $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, takové, že

$$\begin{bmatrix} G_k d_k \\ 1 \end{bmatrix} + \sum_{j\in\mathcal{J}_k}\lambda_j^k\begin{bmatrix} g_j \\ -1 \end{bmatrix} + \lambda_a^k\begin{bmatrix} g_a^k \\ -1 \end{bmatrix} = 0, \tag{1205}$$

přičemž

$$\begin{array}{rcl} \lambda_j^k > 0 & \Rightarrow & -\alpha_j^k + d_k^T g_j = v_k, \\ \lambda_a^k > 0 & \Rightarrow & -\alpha_a^k + d_k^T g_a^k = v_k \end{array}$$

(podmínky komplementarity). Z poslední rovnice soustavy (1205) dostaneme

$$\sum_{j\in\mathcal{J}_k}\lambda_j^k + \lambda_a^k = 1.$$

Platí tedy (1198) (1200) a (1203). Použijeme-li označení (1200)-(1201) a podmínky komplementarity, můžeme psát

$$-\tilde{\alpha}_a^k + d_k^T\tilde{g}_a^k = v_k,$$

což spolu s (1198) dává (1199). Zbývá dokázat, že Lagrangeovy multiplikátory $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$ jsou řešením duální úlohy kvadratického programování (1202)-(1203). Tato úloha je opět konvexní, takže čísla $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, jsou podle tvrzení 4 jejím řešením právě tehdy, existují-li Langrangeovy multiplikátory v_k (odpovídající rovnosti v (1203)) a $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$ (odpovídající nerovnostem v (1203)) tak, že

$$\begin{array}{rcll} -(g_j)^T d_k + \alpha_j^k + v_k - \mu_j^k & = & 0, & j \in \mathcal{J}_k, \\ -(g_a^k)^T d_k + \alpha_a^k + v_k - \mu_a^k & = & 0, & \end{array}$$

přičemž $\lambda_j^k\mu_j^k = 0$, $j \in \mathcal{J}_k$, $\lambda_a^k\mu_a^k = 0$ (pro zjednodušení jsme použili označení (1198) a (1200)). Poslední rovnosti však nejsou nic jiného než nerovnosti (1196), (1197), neboť $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$, a podmínky $\lambda_j^k\mu_j^k = 0$, $j \in \mathcal{J}_k$, $\lambda_a^k\mu_a^k = 0$ jsou ekvivalentní podmínkám komplementarity pro úlohu (1195)-(1197). □

Poznámka 444. Poznamenejme, že omezení (1197) není třeba používat pokud $\mathcal{J}_k = \{1, \ldots, k\}$, neboť je v tomto případě lineární kombinací omezení (1196). Pak ale $\lambda_a^k = 0$ v (1200)-(1201).

Poznámka 445. Kromě agregovaných gradientů (1200) se pomocí Lagrangeových multiplikátorů $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$ definují agregované hodnoty

$$\tilde{F}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k F_j^k + \lambda_a^k F_a^k, \tag{1206}$$

$$\tilde{s}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k. \tag{1207}$$

Máme-li k dispozici směrový vektor d_k, je třeba určit novou aproximaci minima funkce F. Abychom zaručili globální konvergenci svazkové metody, nelze jednoduše položit $x_{k+1} = x_k + d_k$, ale je třeba použít složitější proceduru jejímž výstupem jsou dva body

$$\begin{aligned} x_{k+1} &= x_k + t_L^k d_k, \\ y_{k+1} &= x_k + t_R^k d_k, \end{aligned}$$

kde $0 \leq t_L^k \leq t_R^k \leq 1$ jsou délky kroku. Délky kroku se vybírají takovým způsobem (Algoritmus 5.2), aby nastala právě jedna z možností popsaných v definici 126 a definici 127. V obou definicích používáme označení

$$\beta_{k+1} = \max\left(|F(x_k) - F_{k+1} - (x_k - y_{k+1})^T g_{k+1}|, \gamma |x_k - y_{k+1}|^\nu\right) \tag{1208}$$

a konstanty $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $0 < \tau < 1$ a $D > 0$.

Definice 126. *(Spádový krok) Spádovým krokem nazveme krok, ve kterém platí $t_R^k = t_L^k > 0$,*

$$F(x_{k+1}) \leq F(x_k) - \sigma_L t_L^k w_k \tag{1209}$$

a buď $t_L^k \geq \tau$ nebo $\beta_{k+1} > \sigma_A w_k$.

Definice 127. *(Nulový krok) Nulovým krokem nazveme krok, ve kterém platí $t_R^k > t_L^k = 0$,*

$$d_k^T g_{k+1} \geq \beta_{k+1} - \sigma_R w_k \tag{1210}$$

a $\|y_{k+1} - z_{k+1}\| \leq D$, kde z_{k+1} je libovolný bod, pro který platí $F(z_{k+1}) \leq F(x_k)$.

Máme-li určen nový bod x_{k+1} je třeba do něj transformovat všechny svazkové i agregované hodnoty. To se provádí pomocí vzorců

$$\left.\begin{array}{ll} F_j^{k+1} = F_j^k + (x_{k+1} - x_k)^T g_j, & j \in J_k \\ F_a^{k+1} = \tilde{F}_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k & \\ F_{k+1}^{k+1} = F_{k+1} + (x_{k+1} - y_{k+1}) g_{k+1} & \\ g_a^{k+1} = \tilde{g}_a^k & \\ s_j^{k+1} = s_j^k + \|x_{k+1} - x_k\|, & j \in J_k \\ s_a^{k+1} = \tilde{s}_a^k + \|x_{k+1} - x_k\| & \\ s_{k+1}^{k+1} = \|x_{k+1} - y_{k+1}\| & \end{array}\right\} \tag{1211}$$

Zbývá uvést podmínky, které by měly splňovat matice G_k. Abychom zaručili globální konvergenci svazkové metody, použijeme tento předpoklad.

Předpoklad M. *atice G_k jsou stejnoměrně pozitivně definitní a stejnoměrně omezené (jejich vlastní čísla leží v kompaktním intervalu neobsahujícím nulu). Je-li k-tý krok nulový, platí $h^T G_{k+1}^{-1} h \leq h^T G_k^{-1} h$ $\forall h \in R^n$.*

Nyní můžeme popsat základní algoritmus svazkových metod.

Algoritmus 11

Data $\varepsilon > 0$, $\gamma \geq 0$, $\nu \geq 1$, $m \geq 1$.

Krok 1 (Inicializace). Určíme počáteční bod $x_1 \in R^n$ a počáteční symetrickou pozitivně definitní matici G_1. Položíme $y_1 = x_1$ a vypočteme hodnoty $F_1 = F(y_1)$, $g_1 \in \partial F(y_1)$. Položíme $s_1^1 = s_a^1 = 0$, $F_1^1 = F_a^1 = F_1$, $g_1^1 = g_a^1 = g_1$, $J_1 = \{1\}$ a $k = 1$.

Krok 2 (Směrový vektor). Najdeme řešení úlohy kvadratického programování (1195)-(1197) (omezení (1197) používáme pouze tehdy, jestliže $J_k \neq \{1, \ldots, k\}$.) Dostaneme tak Lagrangeovy multiplikátory λ_j^k, $j \in J_k$ a λ_a^k ($\lambda_a^k \neq 0$ pouze tehdy, jestliže $J_k \neq \{1, \ldots, k\}$), agregované hodnoty $\tilde{g}_a^k$, $\tilde{\alpha}_a^k$, $\tilde{F}_a^k$, $\tilde{s}_a^k$, směrový vektor d_k a čísla v_k, w_k (věta 361). Jestliže $w_k \leq \varepsilon$, ukončíme výpočet.

Krok 3 (Délka kroku). Pomocí Algoritmu 5.2 určíme délky kroku t_L^k, t_R^k tak, abychom dostali buď spádový krok (definice 126) nebo nulový krok (definice 127). Položíme $x_{k+1} = x_k + t_L d_k$, $y_{k+1} = x_k + t_R d_k$ a vypočteme hodnoty $F_{k+1} = F(y_{k+1})$, $g_{k+1} \in \partial F(y_{k+1})$.

Krok 4 (Aktualizace). Vypočteme transformované hodnoty podle (1211) a určíme matici G_{k+1} tak, aby vyhovovala Předpokladu M. Jestliže $|\mathcal{J}_k| < m$, položíme $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\}$. Jestliže $|\mathcal{J}_k| = m$, položíme $\mathcal{J}_{k+1} = \mathcal{J}_k \cup \{k+1\} \backslash \{k+1-m\}$. Položíme $k := k+1$ a přejdeme na Krok 2.

Poznámka 446. Množinu $\mathcal{J}_{k+1}$ můžeme určovat i jiným způsobem než je uvedeno v Kroku 4 algoritmu. V podstatě jde o to, aby obsahovala dostatečný počet indexů a aby platilo $k+1 \in \mathcal{J}_{k+1}$.

Výběr délky kroku (Krok 3 algoritmu) je poměrně komplikovaná procedura, kterou uvedeme ve formě samostatného algoritmu. Abychom zjednodušili označení vynecháme index k a index $k+1$ nahradíme symbolem $+$.

Algoritmus 12

Data $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $\gamma > 0$, $\nu \geq 1$, $0 < \kappa < 1/2$, $0 < \tau < 1/2$, $D > 0$.

Vstup $x \in R^n$, $d \in R^n$, $F = F(x)$, $w > 0$.

Krok 1 (Inicializace). Položíme $t^1 = 1$, $t_A^1 = 0$, $t_U^1 = 1$ a $i = 1$.

Krok 2 (Nové hodnoty). Vypočteme hodnoty $F^i = F(x + t^i d)$, $g^i \in \partial F(x + t^i d)$ a

$$\beta^i = \max\left(|F - F^i + t^i d^T g^i|, \gamma(t^i \|d\|)^\nu\right).$$

Jestliže $F^i \leq F - \sigma_T t^i w$, položíme $t_A^i = t^i$. V opačném případě položíme $t_U^i = t^i$.

Krok 3 (Spádový krok). Jestliže $F^i \leq F - \sigma_L t^i w$ a buď $t^i \geq \tau$ nebo $\beta^i > \sigma_A w$, položíme $t_R = t_L = t^i$, $t_A = t_A^i$, $\beta^+ = \beta^i$ a ukončíme výpočet.

Krok 4 (Nulový krok). Jestliže $d^T g^i \geq \beta^i - \sigma_R w$ a $(t^i - t_A^i)\|d\| \leq D$, položíme $t_R = t^i$, $t_L = 0$, $t_A = t_A^i$, $\beta^+ = \beta^i$ a ukončíme výpočet.

Krok 5 (Aktualizace). Zvolíme $t^{i+1} \in [t_A^i + \kappa(t_U^i - t_A^i), t_U^i - \kappa(t_U^i - t_A^i)]$, položíme $i := i+1$ a přejdeme na Krok 2.

Věta 362. *Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská a nechť pro libovolnou posloupnost $t^i \downarrow 0$ platí*

$$\limsup_{\substack{g^i \in \partial F(x+t^i d) \\ i \to \infty}} d^T g^i \geq \liminf_{i \to \infty} \frac{F(x + t^i d) - F(x)}{t^i}. \tag{1212}$$

Pak Algoritmus 5.2 najde po konečném počtu kroků délky kroku t_L, t_R, t_A takové, že pro body $x^+ = x + t_L d$, $y^+ = x + t_R d$, $z^+ = x + t_A d$ nastane právě jeden z těchto případů:

(a) Spádový krok: Platí $t_R = t_L > 0$,

$$F(x^+) \leq F(x) - \sigma_L t_L w$$

a buď $t_L \geq \tau$ *nebo* $\beta^+ > \sigma_A w$.

(b) Nulový krok: Platí $t_R > t_L = 0$,

$$d^T g(y^+) \geq \beta^+ - \sigma_R w,$$

$\|y^+ - z^+\| \leq D$ *a* $F(z^+) \leq F(x)$.

V obou případech se používá označení

$$\beta^+ = \max\left(|F(x) - F(y^+) - (x - y^+)^T g^+|, \gamma\|x - y^+\|)^\nu\right)$$

Důkaz K ukončení algoritmu dojde buď v Kroku 3, pak zřejmě platí (a), nebo v Kroku 4, pak platí (b). Zbývá tedy dokázat, že k ukončení algoritmu dojde po konečném počtu kroků. Abychom to dokázali, budeme naopak předpokládat, že k ukončení algoritmu nedojde po konečném počtu kroků. Nechť $\{t^i\}$, $\{t^i_A\}$, $\{t^i_U\}$, $\{g^i\}$, $\{\beta^i\}$ jsou posloupnosti hodnot generovaných algoritmem (takže buď $t^i = t^i_A$ nebo $t^i = t^i_U$). Jelikož $t^i_A \leq t^{i+1}_A \leq t^{i+1}_U \leq t^i_U$ a $t^{i+1}_U - t^{i+1}_A \leq (1-\kappa)(t^i_U - t^i_A)$ pro všechny indexy i, existuje nutně hodnota $t^* \geq 0$ taková, že $t^i_A \uparrow t^*$, $t^i_U \downarrow t^*$ a $t_i \to t^*$. Navíc pro dostatečně velké indexy platí $(t^i - t^i_A)\|d\| \leq D$. Označme $\mathcal{S} = \{t \geq 0 : F(x + td) \leq F - \sigma_T t w\}$. Protože $\{t^i_A\} \subset \mathcal{S}$, $t^i_A \uparrow t^*$ a funkce F je spojitá, musí platit

$$F(x + t^* d) \leq F - \sigma_T t^* w, \tag{1213}$$

takže $t^* \in \mathcal{S}$. Nechť $I = \{i : t^i \notin \mathcal{S}\}$. Ukážeme nejprve, že množina I je nekonečná. Pokud by existoval index $\bar{i} \in I$ takový, že $t^i \in \mathcal{S}\ \forall i > \bar{i}$, muselo by platit $t^{\bar{i}}_U = t^i_U \downarrow t^*\ \forall i > \bar{i}$, neboli $t^* = t^{\bar{i}}_U \notin \mathcal{S}$, což je ve sporu s $t^* \in \mathcal{S}$. Množina I je tedy nekonečná a platí $F(x + t^i d) > F - \sigma_T t^i w\ \forall i \in I$, což spolu s (1213) dává

$$\frac{F(x + t^i d) - F(x + t^* d)}{t^i - t^*} > -\sigma_T w \quad \forall i \in I.$$

Použijeme-li předpoklad (1212), dostaneme

$$-\sigma_T w \leq \liminf_{i \xrightarrow{I} \infty} \frac{F(x + t^* d + (t^i - t^*)d) - F(x + t^* d)}{t^i - t^*} \leq \limsup_{i \xrightarrow{I} \infty} d^T g^i. \tag{1214}$$

Vyšetříme nyní dva případy.

(a) Nechť $t^* > 0$. Podle (1213) pro dostatečně velké indexy platí $F(x + t^i d) \leq F - \sigma_L t^i w$, neboť $\sigma_L < \sigma_T$, $t^i \to t^*$ a funkce F je spojitá. Protože nedojde k ukončení algoritmu, musí pro dostatečně velké indexy platit $\beta^i \leq \sigma_A w$ (Krok 3 algoritmu) a $\beta^i - d^T g^i > \sigma_R w$ (Krok 4 algoritmu), což dohromady dává

$$d^T g^i < \beta^i - \sigma_R w \leq -(\sigma_R - \sigma_A) w < -\sigma_T w$$

(neboť $w > 0$) a což je pro $i \in I$ (I je nekonečná) ve sporu s (1214).

(b) Nechť $t^* = 0$. Pak $t^i \to 0$ implikuje $\beta^i \to 0$ (neboť funkce F je spojitá a subgradienty g^i jsou podle věty 332 (a) omezené v okolí bodu x). Protože nedojde k ukončení výpočtu, musí pro velké indexy platit $\beta^i - d^T g^i > \sigma_R w$ (Krok 4 algoritmu), takže

$$\limsup_{i \xrightarrow{I} \infty} d^T g^i \leq -\sigma_R w < -\sigma_T w,$$

což je opět ve sporu s (1214). □

Poznámka 447. Podle věty 347 splňuje podmínku (1212) každá slabě polohladká funkce, neboť výraz na pravé straně (1212) je v tomto případě směrovou derivací (která existuje) a výraz na levé straně je roven limitě (1148).

Nyní dokážeme globální konvergenci Algoritmu 5.1. Vzhledem k tomu, že budeme vyšetřovat vlastnosti nekonečné posloupnosti bodů generovaných tímto algoritmem, budeme předpokládat, že $\varepsilon = 0$ (Krok 2). Dále budeme používat následující předpoklad.

Předpoklad F. *unkce $F : R^n \to R$ je lokálně lipschitzovská na množině $\mathcal{D}_F(F_1) + \overline{\mathcal{B}(0, D)}$, kde množina $\mathcal{D}_F(F_1) = \{x \in \mathcal{D} : F(x) \leq F(x_1)\}$ je kompaktní, a je splněna podmínka (1212) (například, když F je slabě polohladká).*

Poznámka 448. Protože ve spádových krocích hodnota funkce F neroste, platí $x_k \in X$ a protože $\mathcal{D}_F(F_1)$ je kompaktní, je posloupnost $\{x_k\}$ omezená. Jelikož podle věty 362 platí $\|y_k - z_k\| \leq D$, kde $z_k \in X$, můžeme psát $y_k \in \mathcal{D}_F(F_1) + \overline{\mathcal{B}(0, D)}$. Množina $\mathcal{D}_F(F_1) + \overline{\mathcal{B}(0, D)}$ je kompaktní, takže posloupnost $\{y_k\}$ je omezená. Z lokální lipschitzovskosti funkce F na $\mathcal{D}_F(F_1) + \overline{\mathcal{B}(0, D)}$ plyne omezenost posloupnosti $\{g_k\}$. Podle (1215) je i posloupnost $\{\tilde{g}_a^k\}$ omezená. Z (1198) a Předpokladu M pak plyne omezenost posloupnosti $\{d_k\}$.

Lemma 129. *Existují čísla $\tilde{\lambda}_i^k \geq 0$, $1 \leq i \leq k$, $\tilde{\lambda}_1^k + \cdots + \tilde{\lambda}_k^k = 1$ taková, že hodnoty $\tilde{F}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$ získané v Kroku 2 Algoritmu 5.1 vyhovují vztahům*

$$\left(\tilde{F}_a^k, \tilde{g}_a^k, \tilde{s}_a^k\right) = \sum_{i=1}^{k} \tilde{\lambda}_i^k (F_i^k, g_i, s_i^k) \tag{1215}$$

(závorky v (1215) značí, že tato rovnost platí pro všechny prvky dané trojice).

Důkaz Důkaz provedeme indukcí. Předpokládejme, že hodnoty $\tilde{F}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$ vyhovují vztahům (1215) (platí to zřejmě pokud $\mathcal{J}_k = \{1, \ldots, k\}$, kdy $\lambda_a^k = 0$, takže vztahy (1200), (1206), (1207) implikují (1215) s $\tilde{\lambda}_i^k = \lambda_i^k$). Nechť $\lambda_i^{k+1} \geq 0$, $i \in \mathcal{J}_{k+1}$, jsou Lagrangeovy multiplikátory určené řešením úlohy (1195)-(1197) (nebo úlohy (1202)-(1203)), kde index k je nahražen indexem $k+1$, a nechť $\lambda_i^{k+1} = 0$, $i \notin \mathcal{J}_{k+1}$. Položme $\tilde{\lambda}_i^{k+1} = \lambda_i^{k+1} + \lambda_a^{k+1}\tilde{\lambda}_i^k$, $i \leq k$ a $\tilde{\lambda}_{k+1}^{k+1} = \lambda_{k+1}^{k+1}$. Pak podle (1203) platí $\tilde{\lambda}_i^{k+1} \geq 0$, $1 \leq i \leq k+1$, a

$$\sum_{i=1}^{k+1} \tilde{\lambda}_i^{k+1} = \sum_{i=1}^{k+1} \lambda_i^{k+1} + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k = \sum_{i \in \mathcal{J}_{k+1}} \lambda_i^{k+1} + \lambda_a^{k+1} = 1.$$

Dále s použitím (1211), (1200), (1206), (1207) dostaneme

$$
\begin{aligned}
(\tilde{F}_a^{k+1}, \tilde{g}_a^{k+1}, \tilde{s}_a^{k+1}) &= \sum_{i \in \mathcal{J}_{k+1}} \lambda_i^{k+1} (F_i^{k+1}, g_i, s_i^{k+1}) + \lambda_a^{k+1} \left(F_a^{k+1}, g_a^{k+1}, s_a^{k+1}\right) \\
&= \sum_{i \in \mathcal{J}_{k+1}} \lambda_i^{k+1} (F_i^{k+1}, g_i, s_i^{k+1}) \\
&\quad + \lambda_a^{k+1} \left(\tilde{F}_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k, \tilde{g}_a^k, \tilde{s}_a^k + \|x_{k+1} - x_k\|\right) \\
&= \sum_{i=1}^{k+1} \lambda_i^{k+1} (F_i^{k+1}, g_i, s_i^{k+1}) \\
&\quad + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k \left(F_i^k + (x_{k+1} - x_k)^T g_i, g_i, s_i^k + \|x_{k+1} - x_k\|\right) \\
&= \left(\sum_{i=1}^{k+1} \lambda_i^{k+1} + \lambda_a^{k+1} \sum_{i=1}^{k} \tilde{\lambda}_i^k\right) (F_i^{k+1}, g_i, s_i^{k+1}) \\
&= \sum_{i=1}^{k+1} \tilde{\lambda}_i^{k+1} \left(F_i^{k+1}, g_i, s_i^{k+1}\right).
\end{aligned}
$$

□

Lemma 130. *Jestliže posloupnost $\{x_k\}$ generovaná Algoritmem 5.1 má hromadný bod $x^* \in R^n$ a existuje podposloupnost $\{x_k\}_{\mathcal{K}} \subset \{x_k\}$ taková, že $x_k \overset{\mathcal{K}}{\to} x^*$ a $w_k \overset{\mathcal{K}}{\to} 0$, pak bod x^* je stacionárním bodem funkce F (platí $0 \in \partial F(x^*)$).*

Důkaz Podle lemmatu 129 platí (1215). Podle věty 284 existuje nanejvýš $n+2$ dvojic $(g^{k,i}, s^{k,i}), g^{k,i} \in \partial F(y^{k,i})$, $(y^{k,i}, g^{k,i}, s^{k,i}) \in \{(y_i, g_i, s_i) : i = 1, \ldots, k\}$ tak, že platí

$$
(\tilde{g}_a^k, \tilde{s}_a^k) = \sum_{i=1}^{n+2} \lambda^{k,i} \left(g^{k,i}, s^{k,i}\right), \tag{1216}
$$

kde $\lambda^{k,i} \geq 0$, $1 \leq i \leq n+2$, $\lambda^{k,1} + \cdots + \lambda^{k,n+2} = 1$. Podle poznámky 448 jsou vektory $y^{k,i}$, $g^{k,i}$, $1 \leq i \leq n+2$, omezené, takže existuje podmnožina $\overline{\mathcal{K}} \subset \mathcal{K}$ taková, že $y^{k,i} \overset{\overline{\mathcal{K}}}{\to} y_i^*$, $g^{k,i} \overset{\overline{\mathcal{K}}}{\to} g_i^*$, $\lambda^{k,i} \overset{\overline{\mathcal{K}}}{\to} \lambda_i^*$, $1 \leq i \leq n+2$. Podle věty 332 (c) platí $g_i^* \in \partial F(y_i^*)$, $1 \leq i \leq n+2$. Z (1216) pak plyne $(\tilde{g}_a^k, \tilde{s}_a^k) \to (\tilde{g}_a^*, \tilde{s}_a^*)$, kde

$$
(\tilde{g}_a^*, \tilde{s}_a^*) = \sum_{i=1}^{n+2} \lambda_i^* (g_i^*, s_i^*) \tag{1217}
$$

a $\lambda_i^* \geq 0$, $1 \leq i \leq n+2$, $\lambda_1^* + \cdots + \lambda_{n+2}^* = 1$. Navíc (1192) implikuje $s^{k,i} \geq \|x_k - y^{k,i}\|$, což spolu s $x_k \overset{\overline{\mathcal{K}}}{\to} x^*$, $y^{k,i} \overset{\overline{\mathcal{K}}}{\to} y_i^*$ a $s^{k,i} \overset{\overline{\mathcal{K}}}{\to} s_i^*$ dává

$$
s_i^* \geq \|x^* - y_i^*\| \tag{1218}
$$

pro $1 \leq i \leq n+2$. Jelikož $w_k \overset{\overline{\mathcal{K}}}{\to} 0$, matice G_k jsou stejnoměrně pozitivně definitní a $\tilde{\alpha}_a^k \geq 0$, musí podle (1204) platit $\tilde{g}_a^k \overset{\overline{\mathcal{K}}}{\to} 0$, $\tilde{\alpha}_a^k \overset{\overline{\mathcal{K}}}{\to} 0$. Podle (1193), (1194) a (1201) dostaneme

$$\begin{aligned}
\tilde{\alpha}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k \max\left(|F(x_k)-F_j^k|, \gamma(s_j^k)^\nu\right) + \lambda_a^k \max\left(|F(x_k)-F_a^k|, \gamma(s_a^k)^\nu\right) \\
&\geq \max\left(\sum_{j\in\mathcal{J}_k} \lambda_j^k |F(x_k)-F_j^k| + \lambda_a^k |F(x_k)-F_a^k|, \gamma\left(\sum_{j\in\mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k\right)^\nu\right) \\
&\geq \max\left(|F(x_k)-\tilde{F}_a^k|\right), \gamma\left(\tilde{s}_a^k\right)^\nu), \qquad (1219)
\end{aligned}$$

neboť funkce $\max(\cdot,\cdot)$ a $|\cdot|^\nu$, $\nu \geq 1$, jsou konvexní. Platí tedy $\tilde{g}_a^k \overset{\overline{\mathcal{K}}}{\to} 0$, $\tilde{s}_a^k \overset{\overline{\mathcal{K}}}{\to} 0$, což s použitím (1217) a (1218) dává

$$(0,0) = \sum_{i=1}^{n+2} \lambda_i^*(g_i^*, s_i^*)$$

a $y_i^* = x^*$, $1 \leq i \leq n+2$. Tedy $g_i^* \in \partial F(y_i^*) = \partial F(x^*)$ a $0 = \lambda_1^* g_1^* + \cdots + \lambda_{n+2}^* g_{n+2}^* \in \partial F(x^*)$. □

Poznámka 449. *Pokud výpočet skončí předčasně, čili pokud v některém iteračním kroku platí $w_k = 0$, má bod x_k stejné vlastnosti jako bod x^* v lemmatu 130. Platí $\tilde{g}_a^k = 0$ a $\tilde{s}_a^k = 0$, což jako v důkazu lemmatu 130 dává $0 \in \partial F(x_k)$.*

Lemma 131. *Nechť počet spádových kroků v Algoritmu 5.1 je konečný a nechť l-tý iterační krok je posledním spádovým krokem. Pak bod x_{l+1} je stacionárním bodem funkce F (platí $0 \in \partial F(x_{l+1})$).*

Důkaz Nejprve poznamenejme, že pro $k > l$ platí $x_{k+1} = x_k$, takže z (1211) a (1194) plyne

$$\alpha_a^{k+1} = \max\left(|F(x_k)-F_a^{k+1}|, \gamma(s_a^{k+1})^\nu\right) = \max\left(|F(x_k)-\tilde{F}_a^k|, \gamma(\tilde{s}_a^k)^\nu\right),$$

což spolu s (1219) dává $\alpha_a^{k+1} \leq \tilde{\alpha}_a^k$. Nechť $0 \leq \lambda \leq 1$. Označme

$$\begin{aligned}
g_{k+1}(\lambda) &= \lambda g_{k+1} + (1-\lambda) g_a^{k+1} = \lambda g_{k+1} + (1-\lambda)\tilde{g}_a^k \triangleq \tilde{g}_k(\lambda), \\
\alpha_{k+1}(\lambda) &= \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\alpha_a^{k+1} \leq \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\tilde{\alpha}_a^k \triangleq \tilde{\alpha}_k(\lambda).
\end{aligned}$$

Vzhledem k tomu, že w_{k+1} je podle věty 361 minimem funkce (1202) (s indexem $k+1$ místo k), musí pro $k > l$ platit

$$w_{k+1} \leq \frac{1}{2} g_{k+1}^T(\lambda) G_{k+1}^{-1} g_{k+1}(\lambda) + \alpha_{k+1}(\lambda) \leq \frac{1}{2}\tilde{g}_k^T(\lambda) G_k^{-1}\tilde{g}_k(\lambda) + \tilde{\alpha}_k(\lambda) \triangleq w_k(\lambda),$$

neboť pro $k > l$ je $h^T G_{k+1}^{-1} h \leq h^T G_k^{-1} h \ \forall h \in R^n$ (Předpoklad M). Dále poznamenejme, že pro $k > l$ z (1198) a (1210) plyne

$$\alpha_{k+1}^{k+1} + g_{k+1}^T G_k^{-1}\tilde{g}_a^k \leq \sigma_R w_k.$$

neboť v nulových krocích podle (1208) platí $\alpha_{k+1}^{k+1} = \beta_{k+1}$. Postupnými úpravami dostaneme

$$\begin{aligned}
w_k(\lambda) &= \frac{1}{2}\tilde{g}_k^T(\lambda) G_k^{-1}\tilde{g}_k(\lambda) + \tilde{\alpha}_k(\lambda) \\
&= \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \tilde{\alpha}_a^k + \lambda\left(g_{k+1}^T G_k^{-1}\tilde{g}_a^k - (\tilde{g}_a^k)^T G_k^{-1}\tilde{g}_a^k + \alpha_{k+1}^{k+1} - \tilde{\alpha}_a^k\right) \\
&\quad +\lambda^2\left(g_{k+1}-\tilde{g}_a^k\right)^T G_k^{-1}(g_{k+1}-\tilde{g}_a^k) \\
&\leq w_k + \lambda\sigma_R w_k - \lambda w_k + \lambda^2\left(g_{k+1}-\tilde{g}_a^k\right)^T G_k^{-1}(g_{k+1}-\tilde{g}_a^k) \\
&\leq w_k + \lambda(\sigma_R w_k - w_k) + \lambda^2 M,
\end{aligned}$$

kde existence konstanty M plyne z omezenosti hodnot g_{k+1}, $\tilde{g}_a^k$ (poznámka 448) a ze stejnoměrné pozitivní definitnosti matic G_k (Předpoklad M). Výraz na pravé straně nerovnosti nabývá minima pro $\lambda = (1 - \sigma_R)w_k/(2M)$ a jeho minimální hodnota se rovná $w_k - (1-\sigma_R)^2 w_k^2/(4M)$. Platí tedy

$$w_{k+1} \leq w_k - \frac{(1-\sigma_R)^2 w_k^2}{4M}. \tag{1220}$$

Nyní již snadno dokončíme důkaz lemmatu. Ukážeme, že pro $k > l$ platí $w_k \to 0$. Kdyby tomu tak nebylo, musela by existovat konstanta $\delta > 0$ taková, že $w_k \geq \delta \ \forall k > l$ (neboť posloupnost kladných čísel $\{w_k\}$ je podle (1220) nerostoucí pro $k > l$). Pak bychom z (1220) dostali $w_{k+1} \leq w_k - (1-\sigma_R)^2\delta^2/(4M) \ \forall k > l$, takže pro dostatečně velké indexy by platilo $w_k < \delta$, což je spor. Jelikož $x_k = x_{l+1} \ \forall k > l$, platí $x_k \to x_{l+1}$, což spolu s $w_k \to 0$ dává $0 \in \partial F(x_{l+1})$ podle lemmatu 130. □

Věta 363. *Nechť funkce $F : R^n \to R$ splňuje Předpoklad F. Pak každý hromadný bod posloupnosti $\{x_k\}$ generované Algoritmem 5.1 je stacionárním bodem funkce F.*

Důkaz Je-li počet spádových kroků v Algoritmu 5.1 konečný, existuje podle lemmatu 131 právě jeden hromadný bod posloupnosti $\{x_k\}$, který je stacionárním bodem funkce F. Předpokládejme, že $x_k \overset{\mathcal{K}}{\to} x^*$ (množina $\mathcal{K}$ a bod x^* existují, protože posloupnost $\{x_k\}$ je omezená). Utvořme nekonečnou množinu

$$\overline{\mathcal{K}} = \left\{k = k(i) : k(i) \geq i, i \in \mathcal{K}, x_i = \cdots = x_{k(i)} \neq x_{k(i)+1}\right\},$$

takže krok s indexem $k \in \overline{\mathcal{K}}$ je spádový a $x_k \overset{\overline{\mathcal{K}}}{\to} x^*$. Jelikož posloupnost $\{F(x_k)\}$ je nerostoucí a zdola omezená (protože F je lokálně lipschitzovská na kompaktní množině), musí mít limitu a tudíž $F(x_k) - F(x_{k+1}) \overset{\overline{\mathcal{K}}}{\to} 0$. Jelikož pro $k \in \overline{\mathcal{K}}$ platí (1209), můžeme psát

$$0 \leq \sigma_L t_L^k w_k \leq F(x_k) - F(x_{k+1}),$$

takže $t_L^k w_k \overset{\overline{\mathcal{K}}}{\to} 0$. Podle věty 362 platí $\overline{\mathcal{K}} = \mathcal{K}_1 \cup \mathcal{K}_2$, kde $\mathcal{K}_1 = \{k \in \overline{\mathcal{K}} : t_L^k \geq \tau\}$ a $\mathcal{K}_2 = \{k \in \overline{\mathcal{K}} : \beta_{k+1} > \sigma_A w_k\}$. Je-li množina $\mathcal{K}_1$ nekonečná, pak z $t_L^k w_k \overset{\mathcal{K}_1}{\to} 0$ plyne $w_k \overset{\mathcal{K}_1}{\to} 0$ a podle lemmatu 130 je bod x^* stacionárním bodem funkce F. Je-li množina $\mathcal{K}_1$ konečná, musí být množina $\mathcal{K}_2$ nekonečná. Předpokládejme, že existuje číslo δ takové, že množina $\mathcal{K}_3 = \{k \in \mathcal{K}_2, w_k > \delta\}$ je nekonečná. Pak z $t_L^k w_k \overset{\mathcal{K}_3}{\to} 0$ plyne $t_L^k \overset{\mathcal{K}_3}{\to} 0$. Z Předpokladu M a z omezenosti směrových vektorů (poznámka 448) plyne existence čísla $\overline{M} > 0$ takového, že

$$\|x_{k+1} - x_k\| = t_L^k \|d_k\| \leq t_L^k \overline{M},$$

takže $t_L^k \overset{\mathcal{K}_3}{\to} 0$ implikuje $\|x_{k+1} - x_k\| \overset{\mathcal{K}_3}{\to} 0$. Protože ve spádových krocích platí $y_{k+1} = x_{k+1}$, dostaneme $\|y_{k+1} - x_k\| \overset{\mathcal{K}_3}{\to} 0$. To po dosazení do (1208) a využití spojitosti funkce F dává $\beta_{k+1} \overset{\mathcal{K}_3}{\to} 0$. Jelikož $\mathcal{K}_3 \subset \mathcal{K}_2$, platí $0 \leq \sigma_A w_k < \beta_{k+1}$, takže $w_k \overset{\mathcal{K}_3}{\to} 0$, což je ve sporu s definicí množiny $\mathcal{K}_3$. Platí tedy $w_k \overset{\mathcal{K}_2}{\to} 0$ a podle lemmatu 130 je bod x^* stacionárním bodem funkce F. □

Algoritmus 5.1 reprezentuje jednu třídu globálně konvergentních svazkových metod pro minimalizaci nehladkých funkcí. Jednotlivé metody se liší výběrem matice G_k. Nejjednodušší svazková metoda používá matici

$$G_k = u_k I$$

kde $u_k > 0$ jsou váhové koeficienty. Tyto váhové koeficienty se adaptivně nastavují podle jistých (víceméně heuristických) pravidel tak, aby $u_{\min} \leq u_k \leq u_{\max}$ a aby v nulových krocích platilo $u_{k+1} \geq u_k$ (tím je splněn Předpoklad M). Matice G_k může být také určena pomocí kvazinewtonovských aktualizací V tom případě musí být v nulových krocích použita aktualizace hodnosti jedna, která vyhovuje Předpokladu M. Výhodou kvazinewtonovských svazkových metod je to, že matice G_k obsahuje poměrně kvalitní informaci

o minimalizované nehladké funkci, takže je možné používat malé svazky (například s $m = 1$ nebo $m = 2$) což vede ke značné úspoře času při řešení úlohy kvadratického programování (1195)-(1197).

18 Úvod do problematiky nelineárního programování

18.1 Základní pojmy

Zatím jsme se zabývali hledáním lokálních minim spojitých funkcí bez jakýchkoliv omezujících podmínek, tedy v R^n. Nyní budeme hledat lokální minima spojitých funkcí na podmnožině $\mathcal{C} \subset R^n$, definované soustavou rovností a nerovností. Tedy

$$\mathcal{C} = \{x \in R^n : c_i(x) \leq 0,\, i \in I,\; c_i(x) = 0,\, i \in E\}, \tag{1221}$$

kde I a E jsou nějaké indexové množiny (obvykle $I = \{1, \ldots, m_I\}$, $E = \{m_I + 1, \ldots, m_I + m_E = m\}$) a $c_i : R^n \to R$, $i \in I \cup E$ jsou spojité funkce. Množina $\mathcal{C}$ se nazývá množinou přípustných bodů nebo stručněji přípustnou množinou. Omezující podmínky budeme stručně zapisovat ve tvaru $c_I \leq 0$, $c_E = 0$, kde $c_I : R^n \to R^{m_I}$, $c_E : R^n \to R^{m_E}$ jsou spojitá zobrazení ($c_I(x) \leq 0$ je míněno po složkách).

Definice 128. *Minimalizace spojité funkce $F : R^n \to R$ na množině $\mathcal{C}$ definované soustavou rovností a nerovností (1221) se nazývá úlohou matematického programování. Je-li některá z funkcí F a c_i, $i \in I \cup E$, nelineární, jde o úlohu nelineárního programování.*

V definici 128 předpokládáme, že $\mathcal{D}_F \subset R^n$. Do oblasti matematického programování jsou často zařazovány i jiné úlohy. Je-li $\mathcal{D}_F \subset Z^n$, mluvíme o celočíselném programování. Je-li $\mathcal{D}_F$, částí obecnější diskrétní množiny, mluvíme o diskrétním programování. Tyto úlohy však představují zcela odlišnou problematiku a k jejich řešení se používají speciální metody, kterými se zde zabývat nebudeme.

Přípustná množina může být obecně prázdná (zejména tehdy, vyskytují-li se v (1221) vylučující se nerovnosti, například $x \leq 0$ a $1 - x \leq 0$, nebo tehdy, není-li daná nerovnost nikdy splněna, například $x^2 + 1 \leq 0$). Tyto případy odstraníme zavedením vhodných předpokladů. V dalším textu budeme používat tyto základní předpoklady.

Předpoklad C1. *Přípustná množina $\mathcal{C} \subset R^n$ určená vztahy (1221) je neprázdná. Funkce $F : \mathcal{C} \to R$ a $c_i : \mathcal{C} \to R$, $i \in I \cup E$, jsou definované a spojité na $\mathcal{C}$.*

Předpoklad C2. *Přípustná množina $\mathcal{C} \subset R^n$ je kompaktní.*

Poznámka 450. Předpoklad C2 je rozumný a snadno zajistitelný. Uzavřenost $\mathcal{C}$ plyne přímo z (1221) neboť limitní přechod zachovává neostré nerovnosti. Jelikož nás obvykle nezajímají úlohy, kde fukce F má infimum v nekonečnu, lze omezenost množiny $\mathcal{C}$ snadno zajistit použitím dodatečného omezení

$$\sum_{i=1}^{n} x_i^2 \leq R^2, \tag{1222}$$

kde $R > 0$ je vhodné dostatečně velké číslo.

Předpoklad C3. *Funkce $F : \mathcal{C} \to R$ a $c_i : \mathcal{C} \to R$, $i \in I \cup E$, jsou spojitě diferencovatelné na $\mathcal{C}$.*

Předpoklad C4. *Funkce $F : \mathcal{C} \to R$ a $c_i : \mathcal{C} \to R$, $i \in I \cup E$, jsou dvakrát spojitě diferencovatelné na $\mathcal{C}$.*

Z předpokladu C4 plyne C3. Jsou-li splněny předpoklady C1–C2, jsou funkce F a c_i, $i \in I \cup E$ omezené na $\mathcal{C}$. Existují tedy konstanty $\overline{F}$ a $\overline{c}$ takové, že $|F(x)| \leq \overline{F}$ a $|c_i(x)| \leq \overline{c}$, $i \in I \cup E$. Jsou-li splněny předpoklady C1–C3 mají funkce F a c_i, $i \in I \cup E$ omezené první derivace na $\mathcal{C}$. Jsou-li splněny předpoklady C1–C4 mají funkce F a c_i, $i \in I \cup E$ omezené i druhé derivace na $\mathcal{C}$.

I když je snahou nalézt globální minimum funkce F na množině $\mathcal{C}$, omezíme se zde na metody určené k hledání lokálních minim. Úlohy globální optimalizace vyžadují odlišný přístup (nelze zformulovat obecné podmínky pro globální minimum a příslušné metody nemají standardní iterační charakter).

Definice 129. *Řekneme, že bod $x^* \in R^n$ je lokálním řešením úlohy matematického programování (definice 128), existuje-li číslo $\varepsilon > 0$ takové, že*

$$F(x^*) \leq F(x) \quad \forall x \in \mathcal{C} \cap \mathcal{B}(x^*, \varepsilon).$$

kde $\mathcal{C} \subset R^n$ je přípustná množina definovaná předpisem (1221) a $F : R^n \to R$ je funkce spojitá na $\mathcal{C} \cap \mathcal{B}(x^*, \varepsilon)$.

Přítomnost omezení (1221) značně komplikuje řešení úloh matematického programování. Velmi přitom záleží na typu omezujících podmínek, takže je účelné klasifikovat úlohy matematického programování podle typu omezení:

- Úlohy s konvexními omezeními. V tomto případě jsou funkce c_i, $i \in I$, konvexní a funkce c_i, $i \in E$, lineární, takže přípustná množina $\mathcal{C}$ je konvexní (věta 364). Je-li navíc funkce F konvexní, jde o úlohu konvexního programování.
- Úlohy s lineárními omezeními. V tomto případě jsou funkce c_i, $i \in I \cup E$, lineární. Je-li navíc funkce F kvadratická, jde o úlohu kvadratického programování, a je-li funkce F lineární, jde o úlohu lineárního programování.
- Úlohy s nelineárními omezeními ve tvaru rovností, kdy $I = \emptyset$ a $E \neq \emptyset$.
- Úlohy s nelineárními omezeními ve tvaru nerovností, kdy $I \neq \emptyset$ a $E = \emptyset$.
- Smíšené úlohy s nelineárními omezeními, kdy $I \neq \emptyset$ a $E \neq \emptyset$.

Je-li množina $\mathcal{C}$, konvexní, je odvození podmínek optimality značně jednodušší než v obecném případě (věta 364). Navíc pro úlohy konvexního programování existují speciální metody mající výhodné teoretické vlastnosti (oddíl **??**). Řešení úloh s lineárními omezeními (která jsou nutně konvexní) neřináší zásadní teoretické problémy, neboť lze použít metody aktivních omezení, které převádějí úlohu matematického programování na posloupnost optimalizačních úloh na lineárních varietách (oddíl 19.3). Tyto úlohy jsou po vhodné transformaci proměnných ekvivalentní úlohám nepodmíněné minimalizace. Jelikož se přitom pohybujeme v přípustné množině není zapotřebí používat pokutové funkce ani jiné prostředky zaručující konvergenci metody k bodu ležícímu v přípustné množině. Navíc pro úlohy kvadratického a zejména lineárního programování existují speciální účinné metody. Výhodou úloh s omezeními ve tvaru rovností je snadné použití metod rekursivního kvadratického programování (oddíl **??**) a na úlohy s omezeními ve tvaru nerovností lze s výhodou použít metody vnitřních bodů (oddíl **??**). V případě smíšených úloh je použití metod rekursivního kvadratického programování i metod vnitřních bodů poněkud komplikovanější.

Jak již bylo poznamenáno, velký význam pro formulaci podmínek optimality a pro konstrukci efektivních algoritmů má pojem aktivních omezení.

Definice 130. *Nechť $\mathcal{C} \subset R^n$ je přípustná množina úlohy matematického programování (definice 128) a $x \in \mathcal{C}$. Pak množinu $\bar{E}(x) = \bar{I}(x) \cup E$, kde $\bar{I}(x) = \{i \in I : c_i(x) = 0\}$, nazveme množinou indexů omezení aktivních v bodě x. Omezení s indexy $i \in \bar{E}(x)$ nazveme aktivními omezeními.*

Význam aktivních omezení spočívá v tom, že při formulaci podmínek optimality v bodě $x \in \mathcal{C}$, můžeme neaktivní omezení vynechat (tečný kužel $\mathcal{T}_{\mathcal{C}}(x)$ je určen pouze omezeními s indexy $i \in \bar{E}(x)$).

Abychom zjednodušili označení budeme v souladu s běžnou konvencí gradienty omezujících funkcí označovat symboly $a_i(x) = \nabla c_i(x)$. Dále označíme $A_I(x)$ a $A_E(x)$ matice, jejichž sloupci jsou vektory $a_i(x)$, $i \in I$, a $a_i(x)$, $i \in E$ a položíme $A(x) = [A_I(x), A_E(x)]$. Symbolem $\bar{A}(x)$ označíme matici jejímiž sloupci jsou vektory $a_i(x)$, $i \in \bar{E}(x)$, tedy gradienty omezení aktivních v bodě x. Hessovy matice funkcí $c_i(x)$, $i \in I \cup E$ budeme označovat symboly $G_i(x)$, $i \in I \cup E$.

18.2 Podmínky optimality pro úlohy s konvexními omezeními

V tomto oddílu budeme předpokládat, že funkce $c_i : R^n \to R$, $i \in I$, jsou konvexní a funkce $c_i : R^n \to R$, $i \in E$, jsou lineární. Pro zjednodušení zápisu položíme $\bar{I} = \bar{I}(x)$ a $\bar{E} = \bar{E}(x)$ (o který bod $x \in R^n$ se jedná

bude zřejmé z kontextu). Pro zjednodušení výkladu budeme používat označení $\mathcal{C}_i = \{x \in R^n : c_i(x) \leq 0\}$, $i \in I \cup E$, a

$$\mathcal{C}_I = \bigcap_{i \in I} \mathcal{C}_i, \qquad \mathcal{C}_E = \bigcap_{i \in E} \mathcal{C}_i, \qquad \mathcal{C} = \mathcal{C}_I \cap \mathcal{C}_E = \bigcap_{i \in I \cup E} \mathcal{C}_i. \tag{1223}$$

Omezení ve tvaru nerovností rozdělíme na lineární a nelineární odpovídající indexovým množinám I_L a I_N (takže $I = I_L \cup I_N$ a $\bar{I} = \bar{I}_L \cup \bar{I}_N$). Místo předpokladu C1 budeme používat silnější předpoklad.

Předpoklad C1a. *Funkce $c_i : R^n \to R$, $i \in I$, jsou konvexní a funkce $c_i : R^n \to R$, $i \in E$, jsou lineární. Platí*

$$\bigcap_{i \in I_N} \mathcal{C}_i^{\diamond} \neq \emptyset$$

(relativní vnitřky množin $\mathcal{C}_i$, $i \in I_N$, mají neprázdný průnik).

Tento předpoklad se obvykle formuluje ve tvaru Slaterovy podmínky.

Předpoklad C1b. *(Slater) Funkce $c_i : R^n \to R$, $i \in I$, jsou konvexní a funkce $c_i : R^n \to R$, $i \in E$, jsou lineární. Existuje bod $x \in \mathcal{C}$ takový, že $c_i(x) < 0$ $\forall i \in I_N$.*

Poznámka 451. K formulaci podmínek optimality se používají lineární omezení

$$\begin{aligned} l_i(x+y) &= c_i(x) + a_i^T(x)y = a_i^T(x)y \leq 0, \quad i \in \bar{I}, \\ l_i(x+y) &= c_i(x) + a_i^T(x)y = a_i^T(x)y = 0, \quad i \in E, \end{aligned}$$

vzniklá linearizací aktivních omezení $c_i(x)$, $i \in I \cup E$, v bodě $x \in R^n$. Těm odpovídají tečné kužely

$$\begin{aligned} \mathcal{T}_{\mathcal{L}_i}(x) &= \mathcal{H}(a_i(x), 0) = \{y \in R^n : a_i^T(x)y \leq 0, \;\; i \in \bar{I}\}, \\ \mathcal{T}_{\mathcal{L}_i}(x) &= \mathcal{L}(a_i(x), 0) = \{y \in R^n : a_i^T(x)y = 0, \;\; i \in E\}, \end{aligned}$$

kde $\mathcal{H}(a_i(x), 0)$, $i \in \bar{I}$, a $\mathcal{L}(a_i(x), 0)$, $i \in E$, jsou poloprostory a nadroviny v R^n definované v oddílu 15.1 (přitom $\mathcal{H}(0,0) = R^n$ a $\mathcal{L}(0,0) = R^n$), a

$$\mathcal{T}_{\mathcal{L}}(x) = \bigcap_{i \in \bar{E}} \mathcal{T}_{\mathcal{L}_i}(x) = \{y \in R^n : a_i^T(x)y \leq 0,\, i \in \bar{I},\; a_i^T(x)y = 0,\, i \in E\}. \tag{1224}$$

Množinu $\mathcal{T}_{\mathcal{L}}(x)$ nazveme tečným kuželem přípustné množiny omezení linearizovaných v bodě x.

Poznámka 452. Pro další zjednodušení zápisu budeme často argument x vynechávat (množiny odpovídající bodu x^* budeme označovat hvězdičkou). Tuto konvenci jsme již použili u množin $\bar{I} = \bar{I}(x)$, $\bar{E} = \bar{E}(x)$. Budeme tedy psát $a_i = a_i(x) = \nabla c_i(x)$, $G_i = G_i(x) = \nabla^2 c_i(x)$ a také $a_i^* = a_i(x^*)$, $G_i^* = G_i(x^*)$ pro $i \in I \cup E$. Toto zjednodušení zápisu je užitečné zejména pro označování Lagrangeových multiplikátorů a jejich množin (definice 129 a poznámka 457).

Poznámka 453. K odvození prakticky použitelných podmínek optimality pro úlohy matematického programování je důležité aby ve vyšetřovaném bodě $x \in \mathcal{C}$ platilo $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$, kde $\mathcal{T}_{\mathcal{C}}(x)$ je tečný kužel množiny $\mathcal{C}$ v bodě x (definice 110). Zřejmě $\mathcal{T}_{\mathcal{C}}(x) \subset \mathcal{T}_{\mathcal{L}}(x)$, neboť pro libovolný vektor $s \in \mathcal{T}_{\mathcal{C}}(x)$ existují posloupnosti $s_k \to s$, $t_k \downarrow 0$, takové, že $x + t_k s_k \in \mathcal{C}$, neboli

$$c_i(x + t_k s_k) = t_k a_i^T s_k + o(t_k) = t_k a_i^T s + o(t_k) \leq 0, \quad i \in \bar{I},$$

$$c_i(x + t_k s_k) = t_k a_i^T s_k + o(t_k) = t_k a_i^T s + o(t_k) = 0, \quad i \in E$$

(používáme větu o střední hodnotě a poznámku 410), což v limitě dává $a_i^T s \leq 0$, $i \in \bar{I}$, $a_i^T s = 0$, $i \in E$, neboli $s \in \mathcal{T}_{\mathcal{L}}(x)$. Ukážeme, že v případě konvexních omezení splňujících Slaterovu podmínku platí $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$.

Lemma 132. *Je li funkce $c : R \to R$ konvexní, je množina $\mathcal{C} = \{x \in R^n : c(x) \leq 0\}$ konvexní. Je-li funkce c spojitě diferencovatelná v okolí bodu $x \in \mathcal{C}$, ve kterém $c(x) = 0$, platí $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$, kde $\mathcal{T}_{\mathcal{L}}(x) = \{y \in R^n : (\nabla c(x))^T y \leq 0\}$.*

Důkaz (a) Nechť $x_1 \in \mathcal{C}$ (takže $c(x_1) \leq 0$) a $x_2 \in \mathcal{C}$ (takže $c(x_2) \leq 0$). Jelikož funkce c je konvexní, platí $c(\lambda x_1 + (1-\lambda)x_2) \leq \lambda c(x_1) + (1-\lambda)c(x_2) \leq 0$ pro $0 \leq \lambda \leq 1$, takže $\lambda x_1 + (1-\lambda)x_2 \in \mathcal{C}$.

(b) Jestliže $\nabla c(x) = 0$, je bod x globálním minimem funkce c, takže $\mathcal{T}_{\mathcal{C}}(x) = R^n$ a jelikož také $\mathcal{T}_{\mathcal{L}}(x) = \mathcal{H}(0,0) = R^n$, platí $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$. Nechť $\nabla c(x) \neq 0$ a $y \in \mathcal{F}_{\mathcal{C}}(x)$. Pak podle definice 109 existuje číslo $\bar{t} > 0$ takové, že $x + ty \in \mathcal{C}$, pokud $0 \leq t \leq \bar{t}$. Jelikož funkce c je konvexní a spojitě diferencovatelná, můžeme podle věty 327 pro $x + ty \in \mathcal{C}$ (kde $0 < t \leq \bar{t}$), psát

$$0 \geq c(x + ty) \geq (\nabla c(x))^T ty,$$

takže $y \in \mathcal{T}_{\mathcal{L}}(x)$ a jelikož množina $\mathcal{T}_{\mathcal{L}}(x)$ je uzavřená, platí také $\mathcal{T}_{\mathcal{C}}(x) = \overline{\mathcal{F}_{\mathcal{C}}(x)} \subset \mathcal{T}_{\mathcal{L}}(x)$. Nechť naopak $y \in \mathcal{T}_{\mathcal{L}}(x)$ a $(\nabla c(x))^T y < 0$. Pak podle věty 1 existuje číslo $\bar{t}$ takové, že $c(x+ty) < c(x)$ a tedy $x + ts \in \mathcal{C}$, pokud $0 < t \leq \bar{t}$, což podle definice 109 znamená, že $y \in \mathcal{F}_{\mathcal{C}}(x)$. Jelikož obě množiny $\mathcal{T}_{\mathcal{L}}(x)$ a $\mathcal{T}_{\mathcal{C}}(x)$ jsou uzavřené platí $\mathcal{T}_{\mathcal{L}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$. □

Věta 364. *Je-li splněn předpoklad C1a (nebo C1b), je množina $\mathcal{C}$ určená vztahem (1221) konvexní. Jsou-li funkce c_i, $i \in \bar{E}$, spojitě diferencovatelné v okolí bodu $x \in \mathcal{C}$, platí $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$.*

Důkaz (a) Podle lemmatu 132 jsou množiny $\mathcal{C}_i$, $i \in I$, konvexní. Jelikož funkce c_i, $i \in E$, jsou lineární, jsou množiny $\mathcal{C}_i$, $i \in E$, afinní a tudíž konvexní. Podle (1223) a podle věty 282 je tedy množina $\mathcal{C}$ konvexní.

(b) Podle lemmatu 132 platí $\mathcal{T}_{\mathcal{C}_i}(x) = \mathcal{T}_{\mathcal{L}_i}(x)$ pro $i \in \bar{I}_N$ a zřejmě $\mathcal{T}_{\mathcal{C}_i}(x) = R^n$ pro $i \in I \setminus \bar{I}$. Jelikož lineární funkce se při linearizaci zachovávájí, platí $\mathcal{T}_{\mathcal{C}_i}(x) = \mathcal{T}_{\mathcal{L}_i}(x)$ pro $i \in E \cup \bar{I}_L$. Podle (1223), (1224) a věty 314 pak dostaneme

$$\mathcal{T}_{\mathcal{C}}(x) = \bigcap_{i \in \bar{I}} \mathcal{T}_{\mathcal{C}_i}(x) \cap \bigcap_{i \in E} \mathcal{T}_{\mathcal{C}_i}(x) = \bigcap_{i \in \bar{I}} \mathcal{T}_{\mathcal{L}_i}(x) \cap \bigcap_{i \in E} \mathcal{T}_{\mathcal{L}_i}(x) = \mathcal{T}_{\mathcal{L}}(x).$$

□

Věta 365. *Jsou-li splněny předpoklady věty 364, platí*

$$\mathcal{N}_{\mathcal{C}}(x) = \{z \in R^n : z = \sum_{i \in \bar{I}} u_i a_i(x) + \sum_{i \in E} u_i a_i(x), \;\; u_i \geq 0, \; i \in \bar{I}\}. \tag{1225}$$

Důkaz Podle poznámky 451 platí $\mathcal{T}_{\mathcal{L}_i}(x) = \mathcal{H}(a_i(x), 0)$, $i \in \bar{I}$, a $\mathcal{T}_{\mathcal{L}_i}(x) = \mathcal{L}(a_i(x), 0)$, $i \in E$, kde $\mathcal{L}(a_i(x), 0) = \mathcal{H}(a_i(x), 0) \cap \mathcal{H}(-a_i(x), 0)$. Můžeme tedy psát

$$\mathcal{T}_{\mathcal{L}}(x) = \bigcap_{i \in \bar{I}} \mathcal{H}(a_i(x), 0) \cap \bigcap_{i \in E} \mathcal{H}(a_i(x), 0) \cap \bigcap_{i \in E} \mathcal{H}(-a_i(x), 0).$$

Tečný kužel $\mathcal{T}_{\mathcal{L}}(x)$ je tedy polyedrálním kuželem a podle věty 310 platí

$$\begin{aligned} \mathcal{N}_{\mathcal{L}}(x) &= \operatorname{cone}(\operatorname{conv}\{a_i(x), \; i \in \bar{I}, \;\; a_i(x), \; i \in E, \; -a_i(x), \; i \in E\}) \\ &= \{z \in R^n : z = \sum_{i \in \bar{I}} u_i a_i(x) + \sum_{i \in E} u_i a_i(x), \;\; u_i \geq 0, \;\; i \in \bar{I}\}. \end{aligned}$$

Jelikož $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$, platí $\mathcal{N}_{\mathcal{C}}(x) = \mathcal{N}_{\mathcal{L}}(x)$. □

Poznámka 454. Podmínka $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$ se nazývá kvalifikační podmínkou pro omezení (constraint qualification). Je to nejslabší používaná kvalifikační podmínka. Později uvedeme silnější kvalifikační podmínky, které se snadněji ověřují a které rovnost $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$ implikují. Je podstatné, že konvexní omezení vyhovující předpokladu C1a (nebo C1b) kvalifikační podmínku $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$ splňují, takže není třeba nic ověřovat. Význam kvalifikační podmínky spočívá v tom, že její splnění umožňuje vyjádřit normálový kužel ve tvaru (1225).

Nyní zavedeme dva důležité pojmy používané při teoretickém vyšetřování úloh matematického programování a při konstrukci numerických metod používaných k jejich řešení.

Definice 131. *Funkci*

$$L(x,u) = F(x) + \sum_{i\in I\cup E} u_i c_i(x) \tag{1226}$$

nazveme Lagrangeovou funkcí úlohy matematického programování. Vektor $u = [u_1,\ldots,u_m]^T$ *(kde* $m = m_I + m_E$*) nazveme Lagrangeovým vektorem a jeho složky* u_i*,* $i \in I \cup E$*, Lagrangeovými multiplikátory. Je-li splněn předpoklad C3, označíme symbolem*

$$g(x,u) = g(x) + \sum_{i\in I\cup E} u_i a_i(x) \tag{1227}$$

gradient Lagrangeovy funkce. Je-li splněn předpoklad C4, označíme symbolem

$$G(x,u) = G(x) + \sum_{i\in I\cup E} u_i G_i(x) \tag{1228}$$

Hessovu matici Lagrangeovy funkce.

Definice 132. *Nechť jsou splněny předpoklady C1 a C3, kde* $\mathcal{C} \subset R^n$ *je množina určená vztahem (1221),* $x \in \mathcal{C}$ *a* $F : R^n \to R$ *je funkce spojitě diferencovatelná v okolí bodu* x*. Jestliže existují Lagrangeovy multiplikátory* u_i*,* $i \in I \cup E$*, kde* $u_i \geq 0$ *a* $u_i c_i(x) = 0$ *pro* $i \in I$*, takové, že*

$$g(x,u) = g(x) + \sum_{i\in I\cup E} u_i a_i(x) = 0, \tag{1229}$$

řekneme, že bod x *je KKT (Karush, Kuhn, Tucker) bodem úlohy matematického programování. Podmínky* $c_i(x) \leq 0$*,* $u_i \geq 0$ *a* $u_i c_i(x) = 0$*,* $i \in I$*, se nazývají podmínkami komplementarity.*

Poznámka 455. Jelikož $c_i(x) < 0$ a $u_i = 0$, pokud $i \in I \setminus \bar{I}$, můžeme vztah (1229) zapsat ve tvaru

$$g(x,u) = g(x) + \sum_{i\in \bar{I}\cup E} u_i a_i(x) = 0, \tag{1230}$$

kde $u_i \geq 0$, $i \in \bar{I}$.

Věta 366. *(Nutné podmínky prvního řádu) Nechť jsou splněny předpoklady C1 a C3, kde* $\mathcal{C} \subset R^n$ *je množina určená vztahem (1221),* $x^* \in \mathcal{C}$ *a* $F : R^n \to R$ *je funkce spojitě diferencovatelná v okolí bodu* x^**, ve kterém je splněna kvalifikační podmínka* $\mathcal{T}_{\mathcal{C}}(x^*) = \mathcal{T}_{\mathcal{L}}(x^*)$*. Pak je-li bod* x^* *lokálním minimem funkce* F *na množině* $\mathcal{C}$*, je nutně KKT bodem dané úlohy.*

Důkaz Je-li bod $x^* \in \mathcal{C}$ lokálním minimem funkce F na množině $\mathcal{C}$, musí podle věty 318 a poznámky 412 platit $0 \in g(x^*) + \mathcal{N}_{\mathcal{C}}(x^*)$, což s použitím (1225) dává

$$g(x^*) + \sum_{i\in \bar{I}^*} u_i^* a_i(x^*) + \sum_{i\in E} u_i^* a_i(x^*) = 0, \tag{1231}$$

kde $u_i^* \geq 0$ pro $i \in \bar{I}^* = \bar{I}(x^*)$. Použijeme-li dodatečnou podmínku $u_i^* c_i(x^*) = 0$ pro $i \in I$, platí $u_i^* = 0$ pro $i \in I \setminus \bar{I}^*$ a dostaneme (1229). □

Věta 366 říká, že je-li splněna podmínka $\mathcal{T}_{\mathcal{C}}(x^*) = \mathcal{T}_{\mathcal{L}}(x^*)$ a je-li bod $x^* \in \mathcal{C}$ lokálním řešením úlohy matematického programování, existují Lagrangeovy multiplikátory u_i^*, $i \in I \cup E$, takové, že $u_i^* \geq 0$, $u_i^* c_i(x^*) = 0$ pro $i \in I$ a platí (1229). Lagrangeovy multiplikátory však nemusí být určeny jednoznačně (platí-li například $a_i(x^*) = 0$ pro nějaký index $i \in \bar{I}$, můžeme zvolit multiplikátor $u_i^* \geq 0$ libovolně). Tato situace nenastane, jsou-li vektory $a_i(x^*)$, $i \in \bar{E}$, lineárně nezávislé.

Definice 133. *Nechť jsou splněny předpoklady C1 a C3, kde $\mathcal{C} \subset R^n$ je množina určená vztahem (1221), a nechť $x \in \mathcal{C}$. Jsou-li vektory $a_i(x)$, $i \in \bar{E}$, lineárně nezávislé, řekneme, že je splněna podmínka LICQ (linear independence constraint qualification).*

Poznámka 456. Jak ukážeme později, je LICQ nejsilnější používanou kvalifikační podmínkou implikující rovnost $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$ v případě obecných nelineárních omezení. V daný moment je důležité, že podmínka LICQ zaručuje jednoznačnost Lagrangeových multiplikátorů a jejich spojitou závislost na parametrech úlohy. Vztah (1229) můžeme zapsat ve tvaru $-g(x) = \bar{A}\ \bar{u}$, kde $\bar{A} = \bar{A}(x)$, a má-li matice $\bar{A}$ plnou hodnost, dostaneme $\bar{u} = -((\bar{A})^T\bar{A})^{-1}(\bar{A})^T g(x)$.

Poznámka 457. Je-li bod $x \in \mathcal{C}$ KKT bodem úlohy matematického programování, označíme symbolem $\mathcal{U} \subset R^m$ (kde $m = m_I + m_E$) množinu Lagrangeových vektorů $u = [u_1, \ldots, u_m]^T$, které splňují podmínku (1229). Tato množina je neprázdná ale nemusí být jednoprvková (není-li splněna podmínka LICQ).

Definice 134. *Nechť bod $x \in \mathcal{C}$ je KKT bodem úlohy matematického programování a $\mathcal{U}$ je množina Lagrangeových vektorů, pro které platí (1229). Nechť $i \in \bar{I}$. Jestliže existuje alespoň jeden vektor $u \in \mathcal{U}$ takový, že $u_i > 0$, řekneme, že omezení c_i je striktně aktivní v bodě x. Symbolem $\bar{I}_+$ označíme množinu indexů omezení striktně aktivních v bodě x a položíme $\bar{I}_0 = \bar{I} \setminus \bar{I}_+$, takže $\bar{I}_0 = \{i \in \bar{I} : u_i = 0 \ \ \forall u \in \mathcal{U}\}$. Dále označíme $\bar{E}_+ = \bar{I}_+ \cup E = \bar{E} \setminus \bar{I}_0$. Jestliže $\bar{I}_0 = \emptyset$ neboli $\bar{I}_+ = \bar{I}$ (všechny aktivní omezení jsou striktně aktivní), řekneme, že jsou splněny podmínky striktní komplementarity.*

Poznámka 458. Nechť $x \in \mathcal{C}$ je KKT bodem úlohy matematického programování. Z vyjádření (1229) je zřejmé, že rovnost (1229) zůstává zachována i tehdy odstraníme-li omezení s indexy $i \in \bar{I}_0$ (tato omezení jsou v bodě x nadbytečná a lze je při teoretickém vyšetřování vynechat). Jako příklad lze uvést úlohu kvadratického programování, kde $F(x) = x_1^2 + x_2^2$ a $\mathcal{C} = \{x \in R^2 : c(x) = x_2 \leq 0\}$. Řešením této úlohy je bod $x = 0$, ve kterém platí (1229) s Lagrangeovým multiplikátorem $u = 0$. Bod $x = 0$ je také globálním minimem funkce F na R^2.

Nyní zformulujeme podmínky optimality druhého řádu. Jsou-li všechna omezení lineární, je třeba (podobně jako v oddílu 1.2) vyšetřovat výraz $s^T G(x^*)s$ na množině $\tilde{\mathcal{S}}^* = \tilde{\mathcal{S}}(x^*)$, kde

$$\tilde{\mathcal{S}}(x) = \{s \in \mathcal{T}_{\mathcal{C}}(x) : s^T g(x) = 0\} \tag{1232}$$

(v oddílu 1.2 se tato množina rovnala R^n). V obecném případě, kdy jsou funkce c_i nelineární, je situace poněkud komplikovanější, neboť musíme vzít v úvahu i křivost ploch ohraničujících přípustnou množinu, neboli Hessovy matice $G_i(x^*)$, $i \in \bar{E}^*$. Pro tento účel zavedeme označení

$$\tilde{\mathcal{C}} = \tilde{\mathcal{C}}(x) = \{y \in R^n : c_i(y) = 0,\ i \in \bar{E}_+,\ c_i(y) \leq 0,\ i \in I \setminus \bar{I}_+\}, \tag{1233}$$

kde $\bar{E}_+ = \bar{E}_+(x)$ a $\bar{I}_+ = \bar{I}_+(x)$. Zřejmě $\tilde{\mathcal{C}} \subset \mathcal{C}$, množina $\tilde{\mathcal{C}}$ však není konvexní, nejsou-li omezení c_i, $i \in \bar{E}_+$, lineární. V dalším výkladu budeme používat tečné kužely $\mathcal{T}_{\tilde{\mathcal{C}}}(x)$ a $\mathcal{T}_{\tilde{\mathcal{L}}}(x)$. Zřejmě $\mathcal{T}_{\tilde{\mathcal{C}}}(x) \subset \mathcal{T}_{\mathcal{C}}(x)$, $\mathcal{T}_{\tilde{\mathcal{L}}}(x) \subset \mathcal{T}_{\mathcal{L}}(x)$ a podobně jako v poznámce 453 platí $\mathcal{T}_{\tilde{\mathcal{C}}}(x) \subset \mathcal{T}_{\tilde{\mathcal{L}}}(x)$.

Lemma 133. *Nechť jsou splněny předpoklady C1 a C3, kde $\mathcal{C} \subset R^n$ je množina určená vztahem (1221). Nechť $x \in \mathcal{C}$ a $F : R^n \to R$ je funkce spojitě diferencovatelná v okolí bodu x, ve kterém je splněna kvalifikační podmínka $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$. Pak je-li bod x KKT bodem, platí*

$$\tilde{\mathcal{S}}(x) = \mathcal{T}_{\tilde{\mathcal{L}}}(x) = \{s \in R^n :\ \ s^T a_i(x) = 0,\ \ i \in \bar{E}_+,\ \ s^T a_i(x) \leq 0,\ \ i \in \bar{I}_0\}. \tag{1234}$$

Důkaz Jelikož $\mathcal{T}_{\mathcal{C}}(x) = \mathcal{T}_{\mathcal{L}}(x)$, platí $s \in \mathcal{T}_{\mathcal{C}}(x)$ právě tehdy, když $s \in \mathcal{T}_{\mathcal{L}}(x)$, neboli

$$s^T a_i(x) \leq 0,\ \ i \in \bar{I}, \quad s^T a_i(x) = 0,\ \ i \in E. \tag{1235}$$

Je-li bod $x \in \mathcal{C}$ KKT bodem, platí (1230), takže $s^T g(x) = 0$ právě tehdy, když

$$\sum_{i \in \bar{I}} u_i s^T a_i(x) + \sum_{i \in E} u_i s^T a_i(x) = 0$$

pro libovolný vektor $u \in \mathcal{U}$. Jelikož $u_i \geq 0$ a platí (1235), můžeme psát $u_i s^T a_i(x) = 0 \ \forall i \in \bar{E}$. Pokud $i \in \bar{E}_+$, existuje vektor $u \in \mathcal{U}$ takový, že $u_i \neq 0$. Platí tedy $s^T a_i(x) = 0$, pokud $i \in \bar{E}_+$, a $s^T a_i(x) \leq 0$, pokud $i \in \bar{I}_0$. □

Věta 367. *(Nutné podmínky druhého řádu) Nechť jsou splněny předpoklady C1 a C4, kde $\mathcal{C} \subset R^n$ je množina určená vztahem (1221), $x^* \in \mathcal{C}$ a $F : R^n \to R$ je funkce dvakrát spojitě diferencovatelná v okolí bodu x^*, ve kterém je splněna kvalifikační podmínka $\mathcal{T}_{\mathcal{C}}(x^*) = \mathcal{T}_{\mathcal{L}}(x^*)$. Pak je-li bod x^* lokálním minimem funkce F na množině $\mathcal{C}$, je nutně KKT bodem dané úlohy (věta 366) a je-li splněna podmínka $\mathcal{T}_{\tilde{\mathcal{C}}}(x^*) = \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$, je pro libovolný vektor $u^* \in \mathcal{U}^*$ matice $G(x^*, u^*)$ positivně semidefinitní na podprostoru $\{s \in R^n : s^T a_i^* = 0, \ i \in \bar{E}_+^*\}$.*

Důkaz (a) Nechť $x^* \in \mathcal{C}$. Jelikož funkce c_i, $i \in I \setminus \bar{I}^*$, jsou spojité v okolí bodu x^*, který je lokálním minimem funkce F na množině $\mathcal{C}$, existuje číslo $\varepsilon > 0$ takové, že $c_i(x) < 0$ pro $i \in I \setminus \bar{I}^*$, pokud $x \in \mathcal{C} \cap \mathcal{B}(x^*, \varepsilon)$. Pak pro $x \in \tilde{\mathcal{C}}(x^*) \cap \mathcal{B}(x^*, \varepsilon)$ a $u^* \in \mathcal{U}^*$ platí

$$F(x) = F(x) + \sum_{i \in I \cup E} u_i^* c_i(x) = L(x, u^*), \tag{1236}$$

neboť $c_i(x) = 0$ pro $i \in \bar{E}_+^*$ a $u_i^* = 0$ pro $i \in (I \cup E) \setminus \bar{E}_+^*$.

(b) Nechť $s \in \tilde{\mathcal{S}}(x^*) = \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$. Jelikož $\mathcal{T}_{\tilde{\mathcal{C}}}(x^*) = \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$, existují posloupnosti $s_k \to s$ a $t_k \downarrow 0$ takové, že $x_k = x^* + t_k s_k \in \tilde{\mathcal{C}}(x^*) \cap \mathcal{B}(x^*, \varepsilon)$. Aplikujeme-li větu o střední hodnotě na funkci (1236), dostaneme

$$F(x_k) = L(x_k, u^*) = L(x^*, u^*) + t_k s_k^T g(x^*, u^*) + t_k^2 s_k^T G(x^* + \tilde{t}_k s_k, u^*) s_k,$$

kde $0 \leq \tilde{t}_k \leq t_k$. Jelikož $g(x^*, u^*) = 0$ a podle předpokladu platí $F(x_k) \geq F(x^*) = L(x^*, u^*)$ pro $x_k \in \mathcal{B}(x^*, \varepsilon)$, můžeme v limitě psát $s^T G(x^*, u^*) s \geq 0$.

(c) Dokázali jsme, že pro libovolný vektor $u^* \in \mathcal{U}^*$ platí $s^T G(x^*, u^*) s \geq 0$, pokud $s \in \tilde{\mathcal{S}}(x^*)$, neboli pokud $s^T a_i(x^*) = 0$, $i \in \bar{E}_+^*$, a $s^T a_i(x^*) \leq 0$, $i \in \bar{I}_0^*$. Nechť $i \in \bar{I}_0^*$. Jelikož hodnota výrazu $s^T G(x^*, u^*) s$ nezávisí na orientaci vektoru s, platí $s^T G(x^*, u^*) s \geq 0$, pokud $s^T a_i \leq 0$ i pokud $s^T a_i \geq 0$, takže omezení s indexem $i \in \bar{I}_0^*$ není třeba uvažovat. Musí tedy platit $s^T G(x^*, u^*) s \geq 0$, pokud $s^T a_i(x^*) = 0$ pro $i \in \bar{E}_+^*$. □

Podobně jako v případě nepodmíněné minimalizace, je podstatou postačujících podmínek druhého řádu záměna pozitivní semidefinitnosti za pozitivní definitnost. Situace je však příznivější, neboť není třeba předpokládat splnění kvalifikačních podmínek (ty jsou nahrazeny předpokladem, že bod x^* je KKT bodem úlohy matematického programování).

Věta 368. *(Postačující podmínky druhého řádu) Nechť jsou splněny předpoklady C1 a C4, kde $\mathcal{C} \subset R^n$ je množina určená vztahem (1221), $x^* \in \mathcal{C}$ a $F : R^n \to R$ je funkce dvakrát spojitě diferencovatelná v okolí bodu x^*. Pak je-li bod x^* KKT bodem (platí-li (1231) pro $u^* \in \mathcal{U}^*$) a jsou-li matice $G(x^*, u^*)$, $u^* \in \mathcal{U}^*$, pozitivně definitní na podprostoru $\{s \in R^n : s^T a_i^* = 0, \ i \in \bar{E}_+^*\}$, je bod x^* ryzím lokálním minimem funkce F na množině $\mathcal{C}$.*

Důkaz Hessova matice $G(x^*, u^*)$ je podle předpokladu pozitivně definitní na $\mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$, takže je splněna nerovnost $s^T G(x^*, u^*) s > 0$, pokud $s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$ a $\|s\| = 1$. Jelikož množina $\{s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*) : \|s\| = 1\}$ je kompaktní, nabývá na ní výraz $s^T G(x^*, u^*) s$ svého minima $\underline{\lambda} > 0$, neboli $s^T G(x^*, u^*) s \geq \underline{\lambda} > 0$, pokud $s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$ a $\|s\| = 1$. Máme dokázat, že existuje číslo $\varepsilon > 0$ takové, že $F(x) > F(x^*) \ \forall x \in \mathcal{C} \cap \mathcal{B}(x^*, \varepsilon) \setminus \{x^*\}$, což je totéž jako dokázat, že neexistuje posloupnost $x_k \in \mathcal{C} \setminus \{x^*\}$, $x_k \to x^*$ taková, že $F(x_k) \leq F(x^*)$ $\forall k \in N$. Předpokládejme naopak, že taková posloupnost existuje. Tuto posloupnost vyjádříme ve tvaru $x_k = x^* + t_k s_k$, kde $\|s_k\| = 1$ a $t_k \downarrow 0$. Bez újmy na obecnosti budeme předpokládat, že $s_k \to s$ (v opačném případě lze vybrat vhodnou podposloupnost). Pak $s \in \mathcal{T}_{\mathcal{C}}(x^*)$, $\|s\| = 1$ a podle věty o střední hodnotě platí

$$L(x_k, u^*) = L(x^*, u^*) + t_k s_k^T g(x^*, u^*) + o(t_k) = F(x^*) + o(t_k), \tag{1237}$$

neboť $L(x^*, u^*) = F(x^*)$ a $g(x^*, u^*) = 0$.

(a) Předpokládejme nejprve, že $s \in \mathcal{T}_{\mathcal{L}}(x^*) \setminus \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$. Pak existuje index $j \in \bar{I}_+$ a vektor $u^* \in \mathcal{U}^*$ tak, že $u_j^* s^T a_j^* \stackrel{\Delta}{=} -\delta_j < 0$ a podle (1237) platí

$$\begin{aligned} F(x_k) &= L(x_k, u^*) - \sum_{i \in \bar{I}_+} u_i^* c_i(x_k) \geq L(x_k, u^*) - u_j^* c_j(x_k) \\ &= F(x^*) - u_j^* c_j(x^*) - t_k u_j^* s_k^T a_j + o(t_k) = F(x^*) + t_k \delta_j + o(t_k) \end{aligned}$$

(používáme větu o střední hodnotě pro $c_j(x_k)$ a poznámku 410). Odtud plyne existence čísla $l \in N$ takového, že $|o(t_k)|/t_k < (1/2)\delta_j$ a tedy $F(x_k) \geq F(x^*) + (1/2)t_k\delta_j > F(x^*)\ \forall k \geq l$, což je ve sporu s předpokladem, že $F(x_k) \leq F(x^*)\ \forall k \in N$.

(b) Nechť nyní $s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$. Použijeme-li dva členy Taylorova rozvoje a vztah (1237), dostaneme

$$\begin{aligned} F(x_k) &= L(x_k, u^*) - \sum_{i \in \bar{I}_+} u_i^* c_i(x_k) \geq L(x_k, u^*) = L(x^*, u^*) + t_k s_k^T g(x^*, u^*) + t_k^2 s_k^T G(x^*, u^*) s_k + o(t_k^2) \\ &= F(x^*) + t_k^2 s_k^T G(x^*, u^*) s_k + o(t_k^2) \geq F(x^*) + t_k^2 \underline{\lambda} + o(t_k^2). \end{aligned}$$

Odtud plyne existence čísla $l \in N$ takového, že $|o(t_k^2)|/t_k^2 < (1/2)\underline{\lambda}$, takže $F(x_k) \geq F(x^*) + (1/2)t_k^2\underline{\lambda} > F(x^*)\ \forall k \geq l$, což opět je ve sporu s předpokladem, že $F(x_k) \leq F(x^*)\ \forall k \in N$. □

Podobně jako v případě nepodmíněné minimalizace se situace zjednoduší, jsou-li funkce F a c_i, $i \in I \cup E$ konvexní. To má význam zejména pro lineární a konvexní kvadratické programování.

Důsledek 47. *Nechť jsou splněny předpoklady C1 a C4, kde $\mathcal{C} \subset R^n$ je množina určená vztahem (1221), $x^* \in \mathcal{C}$ a $F : R^n \to R$ je funkce dvakrát spojitě diferencovatelná v okolí bodu x^*. Pak jsou-li funkce F a c_i, $i \in I \cup E$ konvexní v okolí bodu x^*, je tento bod lokálním minimem funkce F na množině $\mathcal{C}$ právě tehdy, je-li KKT bodem dané úlohy.*

Důkaz Jsou-li funkce F a c_i, $i \in I \cup E$ konvexní, jsou matice $G(x^*)$ a $G_i(x^*)$, $i \in I \cup E$ pozitivně semidefinitní a je-li bod x^* KKT bodem, jsou splněny nutné podmínky druhého řádu (věta 367). Nechť x^* je KKT bodem dané úlohy. V části (a) důkazu věty 368 jsme ukázali, že funkce F nemůže klesat ve směru $s \in \mathcal{T}_{\mathcal{L}}(x^*) \setminus \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$. Pokud $s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*) = \tilde{\mathcal{S}}(x^*)$, platí podle věty 323 (d) $F(x) \geq F(x^*)s^T g(x^*) = 0$, takže funkce F neklesá ani ve směru $s \in \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$. □

Poznámka 459. Důležitou vlastností Lagrangeových multiplikátorů je to, že udávají rychlost změny hodnoty účelové funkce, při změně omezení. Nechť $x \in \mathcal{C}$ a nechť aktivní omezení $c_k(x) = 0$ je změněno na $c_k(x) = \varepsilon$, kde $k \in \bar{E}$ a $\varepsilon > 0$.

V případě konvexního programování, kdy jsou funkce F a c_i, $i \in I$, konvexní a funkce c_i, $i \in E$, lineární, má velký teoretický a ve speciálních případech (lineární programování a konvexní kvadratické programování) i praktický význam teorie duality. Při vyšetřování duality budeme používat označení $c(x) = [c_I^T(x), c_E^T(x)]^T$, $A(x) = [A_I(x), A_E(x)]$ a $u = [u_I^T, u_E^T]^T$.

Definice 135. *Duální úlohou k (primární) úloze matematického programování (definice 128) navýváme maximalizaci Lagrangeovy funkce $L(x, u) = F(x) + u^T c(x) : R^{n+m} \to R$ na množině zadané omezeními $u_I \geq 0$ a $g(x, u) = g(x) + A(x)u = 0$.*

Věta 369. *Nechť bod $x^* \in \mathcal{C}$ je řešením primární úlohy matematického programování, kde jsou funkce F a c_i, $i \in I$, konvexní a funkce c_i, $i \in E$, lineární, a nechť $u^* \in \mathcal{U}^*$ je Lagrangeův vektor vyhovující podmínce (1231). Pak dvojice $(x^*, u^*) \in R^{n \times m}$ je řešením duální úlohy matematického programování a platí $F(x^*) = L(x^*, u^*)$.*

Důkaz Jsou-li funkce F a c_i, $i \in I$, konvexní a funkce c_i, $i \in E$, lineární a platí-li $u_I \geq 0$, je i Lagrangeova funkce $L(x, u) = F(x) + u^T c(x)$ konvexní. Je-li bod $x^* \in \mathcal{C}$ řešením primární úlohy a splňuje-li dvojice (x, u) omezení $u_I \geq 0$ a $g(x, u) = 0$, můžeme podle věty 323 (d) psát

$$L(x^*, u^*) = F(x^*) \geq F(x^*) + \sum_{i=1}^{m} u_i c_i(x^*) = L(x^*, u) \geq L(x, u) + (x^* - x)^T g(x, u) = L(x, u),$$

takže dvojice (x^*, u^*) je řešením duální úlohy. □

Věta 370. *Jsou-li funkce F a c_i, $i \in I$, konvexní a funkce c_i, $i \in E$, lineární a je-li bod $z \in R^n$ primárně přípustný (platí $z \in \mathcal{C}$) a dvojice $(x, u) \in R^{n \times m}$ duálně přípustná (platí $u_I \geq 0$ a $g(x, u) = 0$), je splněna nerovnost $F(z) \geq L(x, u)$.*

Důkaz Podle věty 323 (d) platí

$$F(z) - F(x) \geq (z - x)^T g(x) = -\sum_{i=1}^{m} (z - x)^T u_i a_i(x) \geq -\sum_{i=1}^{m} u_i (c_i(z) - c_i(x)) \geq \sum_{i=1}^{m} u_i c_i(x)$$

(neboť $u_I \geq 0$, $g(x, u) = 0$ a $c(z) \leq 0$), takže $F(z) \geq F(x) + u^T c(x) = L(x, u)$. □

Poznámka 460. *Z věty 370 plyne, že neexistuje-li optimální řešení primární úlohy (funkce F není zdola omezená na $\mathcal{C}$), neexistuje přípustné řešení duální úlohy a naopak.*

Věta 371. *Jsou-li splněny předpoklady věty 369, je dvojice (x^*, u^*) sedlovým bodem Lagrangeovy funkce.*

Důkaz Zřejmě $g(x^*, u^*) = 0$, takže dvojice (x^*, u^*) je stacionárním bodem Lagrangeovy funkce. Ukážeme, že

$$L(x^*, u) \leq L(x^*, u^*) \leq L(x, u^*),$$

pokud $x \in \mathcal{C}$ a $u_I \geq 0$. Jelikož Lagrangeova funkce je konvexní, můžeme podle věty 323 (d) psát

$$L(x, u^*) \geq L(x^*, u^*) + (x - x^*)^T g(x^*, u^*) = L(x^*, u^*)$$

(neboť $g(x^*, u^*) = 0$) a

$$L(x^*, u^*) = F(x^*) \geq F(x^*) + u_I^T c_I(x^*) = L(x^*, u)$$

(neboť $c_I(x^*) \leq 0$, $c_E(x^*) = 0$ a $u_I \geq 0$). □

19 Minimalizace s lineárními omezeními

V tomto oddílu budeme předpokládat, že všechny funkce $c_i(x)$, $i \in I \cup E$, jsou lineární, takže přípustná množina je definovaná předpisem

$$\mathcal{C} = \{x \in R^n : a_i^T x \leq \alpha_i,\ i \in I,\ a_i^T x = \alpha_i,\ i \in E\}. \tag{1238}$$

V tomto případě lze podmínky optimality značně zjednodušit

Věta 372. *(Nutné podmínky druhého řádu) Nechť $x^* \in \mathcal{C}$, kde $\mathcal{C} \in R^n$ je neprázdná množina určená vztahem (1238) a $F : R^n \to R$ je funkce dvakrát spojitě diferencovatelná v okolí bodu x^*. Pak je-li bod x^* lokálním minimem funkce F na množině $\mathcal{C}$, je nutně KKT bodem dané úlohy a Hessova matice $G(x^*)$ je positivně semidefinitní na podprostoru $\{s \in R^n : a_i^T s = 0,\ i \in \bar{E}_+^*\}$.*

Důkaz Jsou-li všechna omezení lineární platí $\mathcal{T}_{\mathcal{C}}(x^*) = \mathcal{T}_{\mathcal{L}}(x^*)$ i $\mathcal{T}_{\tilde{\mathcal{C}}}(x^*) = \mathcal{T}_{\tilde{\mathcal{L}}}(x^*)$, takže jsou splněny předpoklady věty 367. Navíc platí $G_i(x^*) = 0$, $i \in I \cup E$, takže $G(x^*, u^*) = G(x^*)\ \forall u^* \in \mathcal{U}^*$. Tvrzení věty 372 pak plyne z tvrzení věty 367. □

Věta 373. *(Postačující podmínky druhého řádu) Nechť $x^* \in \mathcal{C}$, kde $\mathcal{C} \in R^n$ je neprázdná množina určená vztahem (1238) a $F : R^n \to R$ je funkce dvakrát spojitě diferencovatelná v okolí bodu x^*. Pak je-li bod x^* KKT bodem a je-li Hessova matice $G(x^*)$, pozitivně definitní na podprostoru $\{s \in R^n : s^T a_i = 0,\ i \in \bar{E}_+^*\}$, je bod x^* ryzím lokálním minimem funkce F na množině $\mathcal{C}$.*

Důkaz Věta 373 je bezprostředním důsledkem věty 368. □

Úlohy s lineárními omezeními mají jistou výhodu v tom, že pokud startujeme z přípustného bodu není obtížné zůstat v přípustné množině. Stačí, aby směrový vektor ležel v přípustném kuželu, čili aby platilo $a_i^T s \leq 0$, $i \in \bar{I}$ a $a_i^T s = 0$, $i \in E$. Není tedy třeba používat různé pokutové funkce, ktreré slouží k penalizaci porušených omezení. Metody tohoto typu se nazývají metodami aktivních omezení.

Kromě metod aktivních omezení existují i další metody pro minimalizaci s lineárními omezeními. Velmi účinné, zejména pro lineární a kvadratické programování, jsou metody vnitřních bodů, které jsou studovány v oddílu **??**.

19.1 Minimalizace na lineární varietě

Chceme-li aplikovat metody aktivních omezení na úlohy s lineárními omezeními, je nutné použít efektivní metody umožňující nalézt minimum účelové funkce na lineární varietě zadané lineárními omezeními ve tvaru rovností, tedy minimalizovat funkci $F : R^n \to R$ na množině

$$\mathcal{C} = \{x \in R^n : a_i^T x = \alpha_i,\ 1 \leq i \leq m\}.$$

Pro tuto úlohu lze KKT podmínky vyjádřit pomocí soustavy n nelineárních a m lineárních rovnic

$$g(x) + Au = 0 \tag{1239}$$
$$A^T x = b \tag{1240}$$

v proměnných x a u (závislost na u je lineární), kde $g(x)$ je gradient funkce F v bodě x, $A = [a_1, \ldots, a_m]$ a $b = [\alpha_1, \ldots, \alpha_m]^T$. Tuto soustavu rovnic lze řešit pomocí Newtonovy metody, jejíž iterační krok má tvar $x_+ = x + \alpha s$, kde s je směrový vektor určený řešením soustavy $n + m$ lineárních rovnic

$$\begin{bmatrix} B & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} s \\ u \end{bmatrix} = - \begin{bmatrix} g(x) \\ 0 \end{bmatrix} \tag{1241}$$

a $\alpha > 0$ je délka kroku. Zde B je symetrická (obvykle pozitivně definitní) aproximace Hessovy matice $G(x)$ a u je získaná aproximace vektoru Lagrangeových multiplikátorů. Předpokládejme, že matice B je regulární

a položme $H = B^{-1}$. Eliminujeme-li ze soustavy rovnic (1241) vektor Lagrangeových multiplikátorů, dostaneme

$$u = -(A^T HA)^{-1} A^T Hg, \tag{1242}$$
$$s = -(H - HA(A^T HA)^{-1} A^T H)g \tag{1243}$$

(za předpokladu, že matice $A^T HA$ je regulární). K výpočtu vektorů u a s lze použít několik přístupů lišících se reprezentací použité lineární variety.

- Přímé řešení soustavy $n + m$ rovnic (1241) (které se nazývají rovnicemi sedlového bodu). Tento přístup se používá zejména tehdy, jsou-li matice A a B rozsáhlé a řídké.
- Použití matice projekce $P = I - HA(A^T HA)^{-1} A^T$ do ortogonálního doplňku podprostoru generovaného sloupci matice A (metody promítaných gradientů).
- Použití matice Z jejíž sloupce tvoří ortonormální bázi v ortogonálním doplňku podprostoru generovaného sloupci matice A (metody redukovaných gradientů).
- Použití matice S jejíž sloupce tvoří konjugovanou bázi v ortogonálním doplňku podprostoru generovaného sloupci matice A takovou, že $S^T BS = I$. Tento přístup vyžaduje, aby matice B byla pozitivně definitní a je výhodný zejména ve spojení s metodami s proměnnou metrikou v součinovém tvaru (oddíl 4.2).

V dalších úvahách budeme předpokládat, že matice A má lineárně nezávislé sloupce, platí $Z^T Z = I$, $A^T Z = 0$, $A^T S = 0$ a matice $[A, Z]$, $[A, S]$ jsou čtvercové a regulární. Nemá-li matice A lineárně nezávislé sloupce, je buď $\mathcal{C} = \emptyset$, nebo jsou některé rovnosti nadbytečné a lze je vynechat.

Věta 374. *Nechť $\hat{H} = PH = H - HA(A^T HA)^{-1} A^T H$. Je-li symetrická matice H pozitivně definitní, platí $\hat{H} = Z\tilde{H}Z^T$, kde $\tilde{H} = Z^T HZ$. Jestliže $S^T BS = I$, platí $\hat{H} = SS^T$. Rovnost $\hat{H} = SS^T$ je splněna pro symetrickou pozitivně definitní matici $H = SS^T + AA^T$.*

Důkaz (a) Sloupce matic A a BZ, kde $B = H^{-1}$, tvoří bázi v R^n, neboť z $Au + BZv = 0$ plyne

$$Z^T Au + Z^T BZv = Z^T BZv = 0 \quad \Rightarrow \quad v = 0,$$
$$A^T HAu + A^T Zv = A^T HAu = 0 \quad \Rightarrow \quad u = 0$$

(matice $A^T HA$ a $Z^T BZ$ jsou pozitivně definitní). Maticová rovnost $\hat{H} - Z\tilde{H}Z^T = 0$ platí právě tehdy, zůstane-li zachována po vynásobení zprava regulární maticí $[A, BZ]$. Jelikož $\hat{H}A = 0$ a $\hat{H}BZ = Z$, dostaneme

$$(\hat{H} - Z\tilde{H}Z^T)A = 0,$$
$$(\hat{H} - Z\tilde{H}Z^T)BZ = Z - ZZ^T HZZ^T BZ = 0,$$

neboť $B = H^{-1}$ a podle lemmatu 37 platí $Z^T BZ = (Z^T HZ)^{-1}$.

(b) Stejným způsobem jako v (a) lze ukázat, že sloupce matic A a BS tvoří bázi v R^n. Jelikož $\hat{H}A = 0$ a $\hat{H}BS = S$, dostaneme

$$(\hat{H} - SS^T)A = 0,$$
$$(\hat{H} - SS^T)BS = S - SS^T BS = 0$$

neboť $B = H^{-1}$ a podle předpokladu platí $S^T BS = I$.

(c) Nechť $H = SS^T + AA^T$. Jelikož sloupce matic A a S tvoří bázi v R^n, lze každý vektor $x \in R^n$ vyjádřit jednoznačně ve tvaru $x = Au + Sv$. Platí tedy

$$x^T Hx = (Au + Sv)^T (AA^T + SS^T)(Au + Sv) = u^T (A^T A)^2 u + v^T (S^T S)^2 v$$

a jelikož matice A^TA a S^TS jsou pozitivně definitní, platí $x^THx \geq 0$, přižemž $x^THx = 0$ právě tehdy, když $x = 0$ (kdy $u = 0$ a $v = 0$), takže matice H je pozitivně definitní. Dále platí

$$\begin{aligned}\hat{H} &= H - HA(A^THA)^{-1}A^TH \\ &= AA^T + SS^T - (AA^T + SS^T)A(A^T(AA^T + SS^T)A)^{-1}A^T(AA^T + SS^T) \\ &= AA^T + SS^T - AA^TA(A^TA)^{-2}A^TAA^T = AA^T + SS^T - AA^T = SS^T.\end{aligned}$$

□

Nejobecnějšími metodami pro minimalizaci na lineární varietě jsou metody redukovaných gradientů. Nechť $z \in \mathcal{C}$, takže $A^Tz = b$. Pak pokud $x \in \mathcal{C}$, můžeme psát $A^T(x - z) = 0$. Má-li matice A $m < n$ lineárně nezávislých sloupců a tvoří-li $n - m$ sloupců matice Z ortonormální bázi v ortogonálním doplňku podprostoru generovaného sloupci matice A, lze bod x jednoznačně vyjádřit ve tvaru $x = z + Z\tilde{x}$. Toto jednoznačné vyjádření ukazuje, že minimalizace funkce $F(x)$ na $\mathcal{C}$ je ekvivalentní minimalizaci funkce $\tilde{F}(\tilde{x}) = F(z + Z\tilde{x})$ na R^{n-m}. Přitom $\tilde{g}(\tilde{x}) = Z^Tg(x)$ a $\tilde{G}(\tilde{x}) = Z^TG(x)Z$. Jestliže $x_+ = x + \alpha s$, kde $x_+ \in \mathcal{C}$, můžeme psát $\alpha s = x_+ - x = Z(\tilde{x}_+ - \tilde{x}) = \alpha Z\tilde{s}$. To znamená, že $s = Z\tilde{s}$, kde $\tilde{s}$ je směrový vektor získaný pomocí redukovaného gradientu $\tilde{g}$ s případným použitím aprozimace $\tilde{B}$ redukované Hessovy matice $\tilde{G} = Z^TGZ$. V obecném případě se směrový vektor $\tilde{s}$ počítá podle vzorců uvedených v poznámce 138. Tedy $s = Z\tilde{s}$, kde například

$$\tilde{s} = -\tilde{g}, \qquad \tilde{s} = -\tilde{g} + \tilde{\beta}_-\tilde{s}_-, \qquad \tilde{s} = -\tilde{H}\tilde{g}, \qquad \tilde{s} = -(\tilde{G})^{-1}\tilde{g}$$

pro metodu největšího spádu, metodu sdružených gradientů, metodu s proměnnou metrikou a Newtonovu metodu. Koeficient $\tilde{\beta}$ se určuje podle vzorců uvedených v poznámce 65, kam dosazujeme vektory $\tilde{s}$, $\tilde{g} = Z^Tg$, $\tilde{g}_+ = Z^Tg_+$ a $\tilde{y} = \tilde{g}_+ - \tilde{g} = Z^Ty$ místo vektorů s, g, g_+ a $y = g_+ - g$. Matice $\tilde{H}$ se aktualizuje podle vzorců uvedených v poznámce 139, takže například pro metodu BFGS platí

$$\frac{1}{\gamma}\tilde{H}_+ = \tilde{H} + \left(\frac{\rho}{\gamma} + \frac{a}{b}\right)\frac{1}{b}\tilde{d}\tilde{d}^T - \frac{1}{b}\left(\tilde{H}\tilde{y}\tilde{d}^T + \tilde{d}(\tilde{H}\tilde{y})^T\right),$$

kde $a = \tilde{y}^T\tilde{H}\tilde{y}$, $b = \tilde{y}^T\tilde{d}$, $c = \tilde{d}^T\tilde{B}\tilde{d}$.

Metody redukovaných gradientů lze realizovat jako metody spádových směrů nebo jako metody s lokálně omezeným krokem. V případě metod spádových směrů se k výběru délky kroku používají Wolfeho podmínky (S2)–(S3), kde vystupují vektory $\tilde{s}$, $\tilde{g}$, $\tilde{g}_+$ místo vektorů s, g, g_+ (platí totiž $s^Tg = (Z\tilde{s})^Tg = \tilde{s}^TZ^Tg = \tilde{s}^T\tilde{g}$ a podobně $s^Tg_+ = \tilde{s}^T\tilde{g}_+$). V případě metod s lokálně omezeným krokem se snažíme přibližně řešit úlohu

$$\tilde{s}^* = \arg\min_{\|\tilde{s}\| \leq \Delta} \tilde{Q}(\tilde{s}),$$

kde

$$\tilde{Q}(\tilde{s}) = \tilde{g}^T\tilde{s} + \frac{1}{2}\tilde{s}^T\tilde{B}\tilde{s}.$$

Jelikož $Z^TZ = I$, platí $\|\tilde{s}\| = \|s\|$. Můžeme tedy použít definici 38, kde vystupuje vektor $\tilde{s}$ a kvadratická funkce $\tilde{Q}(\tilde{s})$ místo vektoru s a kvadratické funkce $Q(s)$. Jelikož minimalizaci na lineární varietě určené m rovnostmi převádíme na nepodmíněnou minimalizaci v R^{n-m} a v R^{n-m} používáme standardní optimalizační metody, zůstávají v platnosti věty o globální a superlineární konvergenci dokázané v předchozích oddílech.

Používáme-li k minimalizaci na lineární varietě metody s proměnnou metrikou, je úspornější pracovat s jedinou maticí S místo se dvěma maticemi Z a $\tilde{H}$. Směrový vektor se pak vypočte podle vzorců $s = S\tilde{s}$, $\tilde{s} = -S^Tg$ (poznámka 125) a matice S se aktualizuje podle vzorců uvedených v důsledku 13. Tento způsob však není zcela obecný, neboť ho nelze použít pro metody s lokálně omezeným krokem, kdy pracujeme s maticí $\hat{B} = (SS^T)^\dagger$.

Metody promítaných gradientů pracují s maticí

$$\hat{H} = H - HA(A^THA)^{-1}A^TH \tag{1244}$$

(takže $\hat{H} = P = I - A(A^TA)^{-1}A^T$, pokud $H = I$). Pak $s = -\hat{H}g$ (takže $s = -Pg$, pokud $H = I$), tedy například

$$s = -Pg, \qquad s = -Pg + \beta_- s_- \qquad s = -\hat{H}g$$

pro metodu největšího spádu, metodu sdružených gradientů a metodu s proměnnou metrikou (nebo Newtonovu metodu, pokud $\hat{H} = G^{-1} - G^{-1}A(A^TG^{-1}A)^{-1}A^TG^{-1}$). Koeficient β se určuje podle vzorců uvedených v poznámce 65, kam dosazujeme vektory Pg a Pg_+ místo vektorů g a g_+. Matice P je symetrická a idempotentní (platí $P^2 = P$), takže je maticí ortogonální projekce do ortogonálního doplňku podprostoru generovaného sloupci matice A. Používáme-li k minimalizaci na lineární varietě metody s proměnnou metrikou, máme dvě možnosti. Buď aktualizujeme matici $\hat{H}$ pomocí rekurentního vztahu

$$\hat{H}_+ = \gamma(\hat{H} + \hat{U}\hat{M}\hat{U}^T), \quad \hat{U} = [d, \hat{H}y], \quad d = -\alpha\hat{H}g$$

(metody s proměnnou vnější metrikou), nebo matici H pomocí rekurentního vztahu

$$H_+ = \gamma(H + UMU^T), \quad U = [d, Hy], \quad d = -\alpha(H - HA(A^THA)^{-1}A^TH)g$$

(metody s proměnnou vnitřní metrikou). Označíme-li $C = (A^THA)^{-1}$ a $C_+ = (A^TH_+A)^{-1}$, pak v prvním případě platí $\gamma C_+ = C$, neboť $A^T\hat{U} = 0$. Ve druhém případě lze použít tuto větu.

Věta 375. *Nechť $C = (A^THA)^{-1}$ a $C_+ = (A^TH_+A)^{-1}$, kde $H_+ = \gamma(H + UMU^T)$, $U = [d, Hy]$ a matice M je určena podle (282). Pak lze psát*

$$\gamma C_+ = C - \frac{m_{22}\, CA^THyy^THAC}{1 + m_{22}\, y^THACA^THy},$$

kde $m_{22} = (\eta - 1)/a$, takže pro metodu BFGS (kdy $\eta = 1$) platí $\gamma C_+ = C$.

Důkaz Jelikož $A^Td = -\alpha A^T((H - HA(A^THA)^{-1}A^TH)g = 0$, můžeme psát

$$A^TH_+A = \gamma\left(A^THA + [0, A^THy]\, M \begin{bmatrix} 0 \\ y^THA \end{bmatrix}\right) = \gamma(A^THA + m_{22}\, A^THyy^THA).$$

Označíme-li $u = A^THy$, $v = m_{22}\, u$ a použijeme-li Shermannův-Morrisonův vzorec (poznámka 106), dostaneme

$$\gamma C_+ = C - \frac{m_{22}\, Cuu^TC}{1 + m_{22}\, u^TCu} = C - \frac{m_{22}\, CA^THyy^THAC}{1 + m_{22}\, y^THACA^THy},$$

kde podle (282) platí $m_{22} = (\eta - 1)/a$. □

19.2 Změna lineární variety při přidání nebo ubrání aktivního omezení

Najdeme-li minimum na lineární varietě, je třeba ověřit, zda je tento bod řešením původní úlohy s lineárními omezeními, tedy jsou-li splněny KKT podmínky. K tomu je třeba určit vektor Lagrangeových multiplikátorů, buď podle vzorce (1242) nebo aplikací metody nejmenších čtverců na soustavu rovnic (1239), kde předpokládáme, že x se nemění. Ve druhém případě podle poznámky 268 platí

$$u = -(A^TA)^{-1}A^Tg, \tag{1245}$$

což je vzorec (1242) s $H = I$. Z těchto úvah plyne, že je třeba kromě matic uvedených v oddílu 19.1 uchovávat navíc matici $C = (A^THA)^{-1}$ (nebo trojúhelníkový rozklad $R^TR = A^THA$), kde $H = I$, používáme-li vzorec (1245).

Nejprve se zaměříme na metody promítaných gradientů, které reprezentují lineární varietu pomocí matice A, matice $C = (A^THA)^{-1}$ (nebo trojúhelníkového rozkladu $R^TR = A^THA$) a matice $\hat{H}$ definované vztahem (1244). V tomto případě lze psát

$$\begin{aligned} u &= -CA^THg = -R^{-1}(R^{-1})^TA^THg, \\ s &= -\hat{H}g. \end{aligned}$$

Přidáváme-li k aktivním omezení nové omezení s normálovým vektorem $a \notin \mathcal{L}(A)$, udává změnu reprezentace lineární variety tato věta.

Věta 376. *Nechť $A^+ = [A, a]$, kde $a \notin \mathcal{L}(A)$ a nechť $C^+ = ((A^+)^T H A^+)^{-1}$, $(R^+)^T R^+ = (A^+)^T H A^+$, $\hat{H}^+ = H - HA^+C^+(A^+)^T H$. Pak platí*

$$\begin{aligned} C^+ &= \begin{bmatrix} C + \frac{CA^THaa^THAC}{a^T\hat{H}a}, & -\frac{CA^THa}{a^T\hat{H}a} \\ -\frac{a^THAC}{a^T\hat{H}a}, & \frac{1}{a^T\hat{H}a} \end{bmatrix}, \\ R^+ &= \begin{bmatrix} R, & r \\ 0, & \rho \end{bmatrix}, \qquad \hat{H}^+ = \hat{H} - \frac{\hat{H}aa^T\hat{H}}{a^T\hat{H}a}, \end{aligned}$$

kde $R^T r = A^T H a$ a $\rho^2 = a^T H a - r^T r$.

Důkaz Je-li $A^+ = [A, a]$, můžeme po dosazení $B = A^T HA$, $b = A^T Ha$, $\beta = a^T Ha$ do (746) psát

$$C^+ = (A^+)^T H A^+ = \begin{bmatrix} A^THA, & A^THa \\ a^THA, & a^THa \end{bmatrix}^{-1} = \begin{bmatrix} C + \frac{CA^THaa^THAC}{a^T\hat{H}a}, & -\frac{CA^THa}{a^T\hat{H}a} \\ -\frac{a^THAC}{a^T\hat{H}a}, & \frac{1}{a^T\hat{H}a} \end{bmatrix}, \tag{1246}$$

kde $\hat{H} = H - HACA^TH$. Podobně použitím (745) dostaneme

$$\begin{aligned} \hat{H}^+ &= H - H[A, a] \begin{bmatrix} A^THA, & A^THa \\ a^THA, & a^THa \end{bmatrix}^{-1} \begin{bmatrix} A^T \\ a^T \end{bmatrix} H \\ &= H - H\left(ACA^T + \frac{(a - ACA^THa)(a - ACA^THa)^T}{a^T\hat{H}a}\right) H \\ &= \hat{H} - \frac{(H - HACA^TH)aa^T(H - HACA^TH)^T}{a^T\hat{H}a} = \hat{H} - \frac{\hat{H}aa^T\hat{H}}{a^T\hat{H}a} \end{aligned} \tag{1247}$$

Nechť $R^T R = A^T HA$, kde R je horní trojúhelníková matice. Položme

$$R^+ = \begin{bmatrix} R, & r \\ 0, & \rho \end{bmatrix}.$$

Pak lze psát

$$(R^+)^T R^+ = \begin{bmatrix} R^TR, & R^Tr \\ r^TR, & r^Tr + \rho^2 \end{bmatrix},$$

takže $(R^+)^T R^+ = (A^+)^T HA^+$ platí právě tehdy, když $R^T r = A^T Ha$ a $r^T r + \rho^2 = a^T Ha$ (r dostaneme řešením soustavy rovnic s dolní trojúhelníkovou maticí a $\rho^2 = a^T Ha - r^T r$). □

Poznamenejme, že matici $C = (A^THA)^{-1}$ (nebo trojúhelníkový rozklad $R^TR = A^THA$) používáme pouze v případě metod s proměnnou vnitřní metrikou, kdy známe matici H. V tomto případě nekonstruujeme explicitně matici $\hat{H}$ a vektor $s = -\hat{H}g$ počítáme podle vzorce $s = -Hg + HACA^THg$. V ostatních případech používáme matici $C = (A^TA)^{-1}$ (nebo trojúhelníkový rozklad $R^TR = A^TA$) a matici $\hat{H}$ (nebo matici P) a v příslušných vzorcích pokládáme $H = I$.

Ubíráme-li od aktivních omezení staré omezení s normálovým vektorem a, udává změnu reprezentace lineární variety tato věta.

Věta 377. *Nechť $AM = [A^-, a]$, kde M je nějaká permutační matice, a nechť $C^- = ((A^-)^T HA^-)^{-1}$, $(R^-)^T R^- = (A^-)^T HA^-$. Nechť*

$$M^TCM = \begin{bmatrix} \tilde{C} & \tilde{c} \\ \tilde{c}^T & \tilde{\gamma} \end{bmatrix}, \qquad QRM = \begin{bmatrix} \tilde{R} & \tilde{r} \\ 0 & \tilde{\rho} \end{bmatrix},$$

kde Q je ortogonální matice taková, že matice QRM je horní trojúhelníková. Pak platí

$$C^- = \tilde{C} - \frac{\tilde{c}\tilde{c}^T}{\tilde{\gamma}}, \qquad R^- = \tilde{R}.$$

Důkaz Vztah pro C^- plyne bezprostředně z vyjádření (1246) (kde píšeme A^-, C^- místo A, C). Jelikož matice QRM je horní trojúhelníková je i matice $R^- = \tilde{R}$ horní trojúhelníková a jelikož

$$\begin{bmatrix} \tilde{R}^T\tilde{R} & \tilde{R}^T\tilde{r} \\ \tilde{r}^T\tilde{R} & \tilde{r}^T\tilde{r}+\tilde{\rho}^2 \end{bmatrix} = MR^TQ^TQRM = MR^TRM = MA^THAM = \begin{bmatrix} (A^-)^THA^- & (A^-)^THa \\ a^THA^- & a^THa \end{bmatrix},$$

platí $(R^-)^TR^- = \tilde{R}^T\tilde{R} = (A^-)^THA^-$. □

Věta 377 neříká nic o tom, jak určit matici $\hat{H}^-$. Jednu z možností udává tato věta, ve které Z je matice jejíž sloupce tvoří ortonormální bázi v ortogonálním doplňku podprostoru generovaného sloupci matice A.

Věta 378. *Nechť jsou splněny předpoklady věty 377 a nechť $\hat{H} = Z\tilde{H}Z^T$, kde $\tilde{H} = Z^THZ$ (věta 374), a $P^- = I - A^-((A^-)^TA^-)^{-1}(A^-)^T$. Pak, položíme-li*

$$\hat{H}^- = \hat{H} + \frac{P^-aa^TP^-}{a^TP^-a}, \tag{1248}$$

kde $P^-a = a - A^-C^-(A^-)^Ta = a - A^-(R^-)^{-1}((R^-)^T)^{-1}(A^-)^Ta$, platí

$$\hat{H}^- = Z^-\begin{bmatrix} Z^THZ & 0 \\ 0 & 1 \end{bmatrix}(Z^-)^T = ZZ^THZZ^T + zz^T,$$

kde $Z^- = [Z, z]$ je matice jejíž sloupce tvoří ortonormální bázi v ortogonálním doplňku podprostoru generovaného sloupci matice A^-.

Důkaz Podle věty 374 lze matici $P^- = I - A^-((A^-)^TA^-)^{-1}(A^-)^T$ vyjádřit ve tvaru $P^- = Z^-(Z^-)^T$, což spolu s předpokladem $\hat{H} = ZZ^THZZ^T$ a vyjádřením (1248), dává

$$\hat{H}^- = [Z,z]\begin{bmatrix} Z^THZ & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} Z^T \\ z^T \end{bmatrix} + [Z,z]\frac{(Z^-)^Taa^TZ^-}{a^TZ^-(Z^-)^Ta}\begin{bmatrix} Z^T \\ z^T \end{bmatrix} = [Z,z]\begin{bmatrix} Z^THZ & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} Z^T \\ z^T \end{bmatrix},$$

neboť $a^TZ = 0$, takže $a^TZ^- = [0, a^Tz]$. □

Věta 378 dokládá vhodnost aktualizace (1248). Redukovaná matice $\tilde{H} = Z^THZ$ se ovroubí sloupcem a řádkem jednotkové matice, což sice změní matici H v definici matice $\hat{H}$, ale výsledná změna matice $\hat{H}$ má hodnost 1, takže se ztrácí pouze minimum informací.

Nyní se zaměříme na metody redukovaných gradientů, které reprezentují lineární varietu pomocí matic A a Z takových, že $[A, Z]$ je čtvercová regulární matice a platí $A^TZ = 0$, $Z^TZ = I$, pomocí trojúhelníkového rozkladu $R^TR = A^TA$ a pomocí redukované matice $\tilde{H} = Z^THZ$ (nebo trojúhelníkového rozkladu $LL^T = \tilde{B}$, kde $\tilde{B} = \tilde{H}^{-1}$). V tomto případě lze psát

$$\begin{aligned} u &= -R^{-1}(R^{-1})^TA^Tg, \\ s &= -Z\tilde{H}Z^Tg. \end{aligned}$$

Přidáváme-li k aktivním omezení nové omezení s normálovým vektorem $a \notin \mathcal{L}(A)$, udávají změnu reprezentace lineární variety tyto věty.

Věta 379. *Nechť $A^+ = [A, a]$, kde $a \notin \mathcal{L}(A)$. Nechť $\tilde{a} = Z^Ta$ a Q je ortogonální matice taková, že*

$$Q^T\tilde{a} = [0, \ldots, 0, \|\tilde{a}\|]^T \tag{1249}$$

Položme $ZQ = [Z^+, z]$. Pak $(A^+)^TZ^+ = 0$ a $(Z^+)^TZ^+ = I$. Předpokládejme, že $\tilde{H} = Z^THZ$. Pak matice $\tilde{H}^+ = (Z^+)^THZ^+$ vznikne z matice $Q^T\tilde{H}Q$ vyškrtnutím posledního řádku a posledního sloupce.

Důkaz Matice Q je ortogonální, takže $\|Q^T\tilde{a}\| = \|\tilde{a}\|$. Jelikož $A^TZ = 0$, takže $A^TZQ = 0$, platí $A^TZ^+ = 0$. Dále podle (1249) platí

$$[a^TZ^+, a^Tz] = a^TZQ = \tilde{a}^TQ = [0, \|\tilde{a}\|],$$

takže $a^T Z^+ = 0$, což dohromady dává $(A^+)^T Z^+ = 0$. Jelikož $Z^T Z = I$ a matice Q je ortogonální, můvzeme psát

$$\begin{bmatrix} (Z^+)^T Z^+ & (Z^+)^T z \\ z^T Z^+ & z^T z \end{bmatrix} = Q^T Z^T Z Q = Q^T Q = I,$$

takže $(Z^+)^T Z^+ = I$. Jestliže $\tilde{H} = Z^T H Z$ a $ZQ = [Z^+, z]$, můžeme psát

$$\begin{bmatrix} (Z^+)^T \\ z^T \end{bmatrix} H \left[Z^+, z\right] = Q Z^T H Z Q^T = Q^T \tilde{H} Q,$$

takže matice $\tilde{H}^+ = (Z^+)^T H Z^+$ je hlavní submaticí řádu $n - m - 1$ matice $Q^T \tilde{H} Q$. □

Věta 380. *Nechť jsou splněny předpoklady věty 379, přičemž ortogonální matice $Q = Q_{m-1,m} \dots Q_{23} Q_{12}$ je součinem Givensových matic elementárních rotací (poznámka 271). Nechť $\tilde{H} = LL^T$, kde L je dolní trojúhelníková matice řádu $n-m$, takže $Q^T L$ je dolní Hessenbergova matice řádu $n-m$. Nechť Givensovy matice elementárních rotací v součinu $\tilde{Q} = \tilde{Q}_{12} \tilde{Q}_{23} \dots \tilde{Q}_{m-1,m}$ jsou vybrány tak, že $Q^T L \tilde{Q}$ je dolní trojúhelníková matice (poznámka 272). Pak pro trojúhelníkovou matici L^+ řádu $n-m-1$, která vznikne z matice $Q^T L \tilde{Q}$ vyškrtnutím posledního řádku a posledního sloupce, platí $\tilde{H}^+ = (L^+)^T L^+$.*

Důkaz Tvrzení věty plyne z toho, že

$$Q^T \tilde{H} Q = Q^T L L^T Q = Q^T L \tilde{Q} \tilde{Q}^T L^T Q = \begin{bmatrix} L^+ & 0 \\ l^T & \lambda \end{bmatrix} \begin{bmatrix} (L^+)^T & l \\ 0 & \lambda \end{bmatrix} = \begin{bmatrix} L^+ (L^+)^T & L^+ l \\ l^T (L^+)^T & l^T l + \lambda \end{bmatrix}$$

a z toho, že matice $\tilde{H}^+$ vznikne z matice $Q^T \tilde{H} Q$ vyškrtnutím posledního řádku a posledního sloupce. □

Ubíráme-li od aktivních omezení staré omezení s normálovým vektorem a, udává změnu reprezentace lineární variety tato věta.

Věta 381. *Nechť $AM = [A^-, a]$, kde M je nějaká permutační matice. Nechť Q je ortogonální matice taková, že matice QRM je horní trojúhelníková. Nechť $Z^- = [Z, z]$, kde*

$$z = AM(QRM)^{-1}[0, \dots, 0, 1]^T \tag{1250}$$

Pak platí $(A^-)^T Z^- = 0$ a $(Z^-)^T Z^- = I$.

Důkaz (a) Jelikož $A^T Z = 0$, je též $(A^-)^T Z = 0$. Dále podle (1250) platí

$$[A^-, a]^T z = M A^T z = M A^T A M (QRM)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = M R^T Q^T Q R M (QRM)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = (QRM)^T \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

(permutační matice M je symetrická) a jelikož matice $(QRM)^T$ je dolní trojúhelníková, má vektor $[A^-, a]^T z$ nenulový pouze poslední prvek ($a^T z \neq 0$, neboť matice A, R a QRM mají lineárně nezávislé sloupce), takže $(A^-)^T z = 0$. Spojíme-li oba výsledky, dostaneme $(A^-)^T Z^- = 0$.

(b) Jelikož $A^T Z = 0$, plyne z (1250), že $z^T Z = 0$. Dále platí

$$\begin{aligned} z^T z &= [1,0](MR^TQ^T)^{-1} M A^T A M (QRM)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} \\ &= [0,1](MR^TQ^T)^{-1} M R^T Q^T Q R M (QRM)^{-1} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = [0,1] \begin{bmatrix} 0 \\ 1 \end{bmatrix} = 1. \end{aligned}$$

Spojíme-li oba výsledky a použijeme-li rovnost $Z^T Z = I$, dostaneme $(Z^-)^T Z^- = I$. □

Poznámka 461. Používáme-li redukované metody s proměnou metrikou, je třeba po ubrání aktivního omezení určit matici $\tilde{H}^-$, jejíž řád je o jednotku vyšší nežli řád matice $\tilde{H}$. Jelikož neznáme matici H, nedokážeme určit vektor Z^THz a číslo z^THz (v matici $\tilde{H}$ se kumulují pouze informace získané v podprostoru generovaném sloupci matice Z). V tomto případě lze postupovat podobně jako ve větě 378 a položit

$$\hat{H}^- = \begin{bmatrix} \hat{H} & 0 \\ 0 & 1 \end{bmatrix}, \qquad L^- = \begin{bmatrix} L & 0 \\ 0 & 1 \end{bmatrix}.$$

Nakonec se zaměříme na metody s proměnnou metrikou v součinovém tvaru, které reprezentují lineární varietu pomocí matic A a S takových, že $[A, S]$ je čtvercová regulární matice a platí $A^TS = 0$, $S^TBS = I$ a pomocí trojúhelníkového rozkladu $R^TR = A^TA$. V tomto případě lze psát

$$\begin{aligned} u &= -R^{-1}(R^{-1})^TA^Tg, \\ s &= -SS^Tg. \end{aligned}$$

Tak jako ve větě 374 budeme předpokládat, že $SS^T = \hat{H} = H - HA(A^THA)^{-1}A^TH$.

Přidáváme-li k aktivním omezení nové omezení s normálovým vektorem $a \notin \mathcal{L}(A)$, udává změnu reprezentace lineární variety tato věta.

Věta 382. *Nechť $A^+ = [A, a]$, kde $a \notin \mathcal{L}(A)$. Nechť $S = [\tilde{S}, s]$ a*

$$S^+ = \tilde{S} - \left(\frac{1 - \lambda a^Ts}{a^TSS^Ta}SS^Ta + \lambda s\right)a^T\tilde{S} \tag{1251}$$

kde λ je kořenem kvadratické rovnice

$$\lambda^2 a^T\tilde{S}\tilde{S}^Ta + 2\lambda a^Ts = 1 \tag{1252}$$

Pak $(A^+)^TS^+ = 0$ a platí

$$S^+(S^+)^T = H - HA^+((A^+)^THA^+)^{-1}(A^+)^TH \tag{1253}$$

Důkaz Jelikož matice $\tilde{S}$ vznikne z matice S vyškrtnutím sloupce s, mužeme psát $SS^T = \tilde{S}\tilde{S}^T + ss^T$. Nejprve ukážeme, že platí

$$S^+(S^+)^T = SS^T - \frac{SS^Taa^TSS^T}{a^TSS^Ta}, \tag{1254}$$

což spolu s (1247) dává $S^+(S^+)^T = \hat{H}^+ = H - HA^+((A^+)^THA^+)^{-1}(A^+)^TH$. Položme pro zjednodušení zápisu $\alpha = (1 - \lambda a^Ts)/a^TSS^Ta$ a $\omega = a^T\tilde{S}\tilde{S}^Ta = a^TSS^Ta - (a^Ts)^2$. Pak podle (1251) platí

$$\begin{aligned} S^+(S^+)^T &= (\tilde{S} - (\alpha SS^Ta + \lambda s)\,a^T\tilde{S})(\tilde{S} - (\alpha SS^Ta + \lambda s)\,a^T\tilde{S})^T \\ &= \tilde{S}\tilde{S}^T - (\alpha SS^Ta + \lambda s)\,a^T\tilde{S}\tilde{S}^T - \tilde{S}\tilde{S}^Ta(\alpha SS^Ta + \lambda s)^T \\ &\quad + (\alpha SS^Ta + \lambda s)\,a^T\tilde{S}\tilde{S}^Ta(\alpha SS^Ta + \lambda s)^T \\ &= SS^T - ss^T - (\alpha SS^Ta + \lambda s)\,a^TSS^T + (\alpha SS^Ta + \lambda s)a^Tss^T \\ &\quad - SS^Ta(\alpha SS^Ta + \lambda s)^T + ss^Ta(\alpha SS^Ta + \lambda s)^T + \omega(\alpha SS^Ta + \lambda s)(\alpha SS^Ta + \lambda s)^T \\ &= SS^T + (\omega\alpha^2 - 2\alpha)SS^Taa^TSS^T + (\alpha a^Ts - \lambda + \alpha\omega\lambda)SS^Tas^T \\ &\quad + (\alpha a^Ts - \lambda + \alpha\omega\lambda)sa^TSS^T + (2\lambda a^Ts + \omega\lambda^2 - 1)ss^T. \end{aligned}$$

Použijeme-li vztah (1252), dostaneme $2\lambda a^T s + \omega\lambda^2 - 1 = 0$,

$$\begin{aligned}
\alpha a^T s - \lambda + \alpha\omega\lambda &= \frac{a^T s\,(1 - \lambda a^T s) - \lambda(\omega + (a^T s)^2) + \lambda\,\omega(1 - \lambda a^T s)}{\omega + (a^T s)^2} \\
&= \frac{a^T s\,(1 - 2\lambda a^T s - \omega\lambda^2)}{\omega + (a^T s)^2} = 0\,, \\
\omega\alpha^2 - 2\alpha &= \frac{\omega(1 - \lambda a^T s)^2 - 2(1 - \lambda a^T s)(\omega + (a^T s)^2)}{a^T S S^T a(\omega + (a^T s)^2)} \\
&= \frac{\omega\lambda^2(a^T s)^2 + 2\lambda(a^T s)^3 - 2(a^T s)^2 - \omega}{a^T S S^T a(\omega + (a^T s)^2)} = -\frac{1}{a^T S S^T a}\,,
\end{aligned}$$

což po dosazení do předchozí rovnosti dává (1254). Z vyjádření (1253) plyne, že $(A^+)^T S^+ (S^+)^T A^+ = 0$ a jelikož A^+ a S^+ mají plnou hodnost, také $(A^+)^T S^+ = 0$. □

Ubíráme-li od aktivních omezení staré omezení s normálovým vektorem a, udává změnu reprezentace lineární variety tato věta.

Věta 383. *Nechť $AM = [A^-, a]$, kde M je nějaká permutační matice. Nechť Q je ortogonální matice taková, že QRM je horní trojúhelníková. Nechť $S^- = [S, s]$, kde*

$$s = AM(QRM)^{-1}[0, \ldots, 0, 1]^T \tag{1255}$$

Pak platí $(A^-)^T S^- = 0$ a

$$S^-(S^-)^T = H - HA^-((A^-)^T H A^-)^{-1}(A^-)^T H$$

Důkaz (a) Vztah $(A^-)^T S^- = 0$ se dokazuje postupem, který jsme použili v části (a) důkazu věty 381 (místo Z^-, Z a z píšeme S^-, S a s), neboť v této části důkazu se nepoužívá ortogonalita sloupců matice Z. Poznamenejme, že vztah (1255) implikuje ortogonalitu $S^T s = 0$ a nerovnost $a^T s \neq 0$.

(b) Jelikož $S^-(S^-)^T = SS^T + ss^T$ a $A^T S = 0$, můžeme psát $S^-(S^-)^T a = ss^T a$. Platí tedy

$$\begin{aligned}
S^-(S^-)^T &= SS^T + ss^T = SS^T + \frac{ss^T aa^T ss^T}{a^T ss^T a} \\
&= SS^T + \frac{S^-(S^-)^T aa^T S^-(S^-)^T}{a^T S^-(S^-)^T a},
\end{aligned}$$

což spolu s (1247) dává $S^-(S^-)^T = \hat{H}^- = H - HA^-((A^-)^T H A^-)^{-1}(A^-)^T H$. □

19.3 Metody aktivních omezení

Myšlenka metod aktivních omezení spočívá v postupné minimalizaci účelové funkce na lineárních varietách určených omezeními aktivními v průběžném bodě. Množiny indexů aktivních omezení se adaptivně mění tak, abychom zůstali v přípustné množině a aby docházelo ke snížení hodnoty účelové funkce. Abychom porozuměli metodám aktivních omezení uvedeme nejprve vzorový algoritmus spolu s potřebnými poznámkami.

Algoritmus 30. (metoda aktivních omezení)

Krok 1 Najdeme přípustný bod $x \in \mathcal{C}$, určíme množinu $\overline{E}(x)$ indexů omezení aktivních v bodě x a zvolíme reprezentaci lineární variety

$$\overline{\mathcal{L}}(x) = \bigcap_{i \in \overline{E}(x)} \mathcal{L}(a_i, \alpha_i).$$

Vypočteme hodnotu $F(x)$ a gradient $g(x)$ účelové funkce F v bodě x.

Krok 2 Způsobem popsaným oddílu 19.1 (minimalizací na lineární varietě $\overline{\mathcal{L}}(x)$) určíme směrový vektor $s \in R^n$ a vektor Lagrangeových multiplikátorů $u \in R^m$. Jestliže s požadovanou přesností platí $\|s\| = 0$ a $u \geq 0$ (míněno po složkách) ukončíme výpočet (získali jsme aproximaci KKT bodu).

Krok 3 Jestliže s dostatečnou přesností platí $\|s\| = 0$ a $u < 0$, určíme index $\ell \in \overline{E}(x)$ takový, že

$$u_\ell = \min_{i \in \overline{E}(x)} (u_i),$$

a ubereme omezení s indexem ℓ, čímž změníme množinu $\overline{E}(x)$, varietu $\overline{\mathcal{L}}(x)$ a její reprezentaci. Přejdeme na krok 2.

Krok 4 Jestliže s dostatečnou přesností platí $\|s\| > 0$, určíme délku kroku $0 < \alpha < \overline{\alpha}$ (například Armijovým výběrem), kde

$$\overline{\alpha} = \min_{i \notin \overline{E}(x), a_i^T s > 0} \frac{\alpha_i - a_i^T x}{a_i^T s}.$$

Vypočteme hodnotu $F(x)$ a gradient $g(x)$ účelové funkce F v bodě x.

Krok 5 Pokud $\alpha = \overline{\alpha}$, přidáme nová aktivní omezení, čímž změníme množinu $\overline{E}(x)$, varietu $\overline{\mathcal{L}}(x)$ a její reprezentaci.

Krok 6 Přejdeme na krok 2.

K realizaci tohoto algoritmu je třeba uvést několik poznámek.

(1) Určení počátečního přípustného bodu je speciální úlohou lineárního programování. Tato úloha je studována v oddílu **??**.

(2) V algoritmu jsou použity dva pojmy, požadovaná přesnost a dostatečná přesnost. Požadovaná přesnost je přesností s jakou chceme splnit KKT podmínky, zatímco dostatečná přesnost stanovuje, kdy je třeba ubírat aktivní omezení. V algoritmu je použita norma směrového vektoru, ale v praxi je výhodnější používat normu promítaného nebo redukovaného gradientu (podle typu zvolené reprezentace lineární variety). Tyto otázky, které ovlivnují globální konvergenci metody, jsou hlavní náplní tohoto oddílu.

(3) Pohybujeme-li se v lineární varietě, můžeme narazit na omezení, které bylo původně neaktivní a které musíme vzít do úvahy. Maximální délka kroku vypočtená v kroku 4 odpovídá nejbližšímu omezení, které se stává aktivním.

(4) V konečné aritmetice nelze počítat úplně přesně. Proto používáme mez ε_a a i-té omezení považujeme za neaktivní, pokud $a_i^T x - \alpha_i < \varepsilon_a \max(1, |\alpha_i|)$, za aktivní, pokud $|a_i^T x - \alpha_i| \leq \varepsilon_a \max(1, |\alpha_i|)$, a za porušené, pokud $a_i^T x - \alpha_i > \varepsilon_a \max(1, |\alpha_i|)$. Přitom KKT podmínky považujeme za splněné, není-li žádné omezení porušené a platí-li $\|g(x, u)\| \leq \varepsilon_x \max(1, \|g(x)\|)$ a $u_i \geq -\varepsilon_u \max(1, \|g(x)\|)$, když $i \in I$.

Učební texty

[T1] Z.Dostál, P.Beremlijski: Metody optimalizace. Učební text, Ostrava 2012.

[T2] J.Duintjer-Tebbens, I.Hnětynková, M.Plešinger, Z.Strakoš, P.Tichý: Analýza metod pro maticové výpočty. Matfyzpress Praha 2012.

[T3] M.Fiedler: Speciální matice a jejich použití v numerické matematice. SNTL, Praha 1981.

[T4] J.Kurzweil: Obyčejné diferenciální rovnice. SNTL Praha 1970.

[T5] L.Lukšan: Metody s proměnnou metrikou. Academia, Praha 1990.

[T6] J.Machalová, H.Netuka: Numerické metody nepodmíněné optimalizace. Učební text, Olomouc 2012.

[T7] J.Machalová, H.Netuka: Nelineární programování. Teorie a metody. Učební text, Olomouc 2012.

[T8] W.Rudin: Analýza v reálném a komplexním oboru. Academia, Praha 2003.

[T9] M.Tůma: Teorie grafů a soustavy lineárních algebraických rovnic. Učební text, Praha 2015.

[T10] E.Vitásek: Základy teorie numerických metod pro řešení diferenciálních rovnic. Academia, Praha 1994.

[T11] L.Zajíček: Vybrané úlohy z matematické analýzy pro 1. a 2. ročník. Matfyzpress, Praha 2005

Literatura

[1] M.Al-Baali: Descent property and global convergence of the Fletcher-Reeves method with inexact linesearch. IMA J. Numerical Analysis 5 (1985) 121-124.

[2] E.Anderson, Z.Bai, C.Bischof, S.Blackford, J.Demmel, J.Dongarra, J.Du Croz, A.Greenbaum, S.Hammarling, A.McKenney, D.Sorensen: LAPACK User's Guide. SIAM, Philadelphia 1999.

[3] N.Andrei: An unconstrained optimization test functions collection, Advanced Modeling and Optimization 10 (2008) 147-161.

[4] M.C.Biggs: Minimization algorithms making use of nonquadratic properties of the objective function. J. Inst. Math. Appl. 8 (1971) 315-327.

[5] M.C.Biggs: A note on minimization algorithms which make use of non-quadratic properties of the objective function. J. Inst. Maths. Appl. 12 (1973), 337-338.

[6] E.G.Birgin, J.M.Martinez: A spectral conjugate gradient method for unconstrained optimization. Applied Mathematics and Optimization 43 (2001) 117-128.

[7] J.F.Bonnans, J.C.Gilbert, C.Lemarechal, C.A.Sagastizabal: Numerical Optimization. Theoretical and Practical Aspects. Springer-Verlag Berlin, Heidelberg, 2006

[8] S.Boyd, L.Vandenberghe: Convex optimization. Cambridge University Press, Cambridge, 2006.

[9] K.W.Brodlie, A.R.Gourlay, J.Greenstadt: Rank-one and rank-two corrections to positive definite matrices expressed in product form. J. Inst. Maths. Applics. 11 (1973) 73-82.

[10] I.Bongartz, A.R.Conn, N.Gould, P.L.Toint: CUTE – constrained and unconstrained testing environment. ACM Transactions on Mathematical Software 21 (1995), 123-160.

[11] C.G.Broyden: A class of methods for solving nonlinear simultaneous equations. Mathematics of Computation 19 (1965) 577-593.

[12] C.G.Broyden: Quasi-Newton methods and their application to function minimization. Mathematics of Computation 21 (1967) 368-381.

[13] C.G.Broyden: The convergence of a class of double rank minimization algorithms. Part 1 – general considerations. Part 2 – the new algorithm. J. Institute of Mathematics and its Applications 6 (1970) 76-90, 222-231.

[14] A.Buckley: A combined conjugate-gradient quasi-Newton minimization algorithm. Mathematical Programming 15 (1978) 200-210.

[15] A.Buckley, A.LeNir: QN-like variable storage conjugate gradients. Mathematical Programming 27 (1983) 155-175.

[16] J.R.Bunch, B.N. Parlett: Direct methods for solving symmetric indefinite systems of linear equations. SIAM J. Numerical Analysis 8 (1971) 639-655.

[17] R.H.Byrd, D.C.Liu, J.Nocedal: On the behavior of Broyden's class of quasi-Newton methods. SIAM J. Optimization 2 (1992) 533-557.

[18] R.H.Byrd, J.Nocedal, R.B.Schnabel: Representation of quasi-Newton matrices and their use in limited memory methods. Mathematical Programming 63 (1994) 129-156.

[19] R.H.Byrd, J.Nocedal, Y.Yuan: Global convergence of a class of quasi-Newton methods on convex problems. SIAM J. Numerical Analysis 24 (1987) 1171-1190.

[20] W.Cheng: A two-term PRP-based descent method. Numerical Functional Analysis and Optimization 28 (2007) 1217-1230.

[21] W.Cheng: Spectral scaling BFGS method. J. Optimizaton Theory and Applications 146 (2010) 305-319.

[22] T.F.Coleman, B.S.Garbow, J.J.Moré: Software for estimating sparse Jacobian matrices. ACM Transactions on Mathematical Software 10 (1984) 329-345.

[23] T.F.Coleman, B.S.Garbow, J.J.Moré: Algorithm 618. Fortran subroutines for estimating sparse Jacobian matrices. ACM Transactions on Mathematical Software 10 (1984) 346-347.

[24] T.F.Coleman, B.S.Garbow, J.J.Moré: Software for estimating sparse Hessian matrices. ACM Transactions on Mathematical Software 11 (1985) 363-378.

[25] T.F.Coleman, B.S.Garbow, J.J.Moré: Algorithm 636. Fortran subroutines for estimating sparse Hessian matrices. ACM Transactions on Mathematical Software 11 (1985) 378-378.

[26] T.F.Coleman, J.J.Moré: Estimation of sparse Jacobian matrices and graph coloring problems. SIAM J. Numerical Analysis 20 (1983) 187-209.

[27] T.F.Coleman, J.J.Moré: Estimation of sparse Hessian matrices and graph coloring problems. Mathematical Programming 28 (1984) 243-270.

[28] Y.Dai: Nonlinear conjugate gradient methods. Preprint (2010) 1-36.

[29] Y.Dai, J.Han, G.Liu, D.Sun, H.Yin, Y.Yuan: Convergence properties of nonlinear conjugate gradient methods. SIAM J. Optimization 10 (1999) 345-358.

[30] Y.H.Dai, C.X.Kou: A nonlinear conjugate gradient algorithm with an optimal property and an improved Wolfe line search. SIAM J. on Optimization, 23 (2013) 296-320.

[31] Y.H.Dai, L.Z.Liao: New conjugacy conditions and related nonlinear conjugate gradient methods. Applied Mathematics and Optimization 43 (2001) 87-101.

[32] Y.Dai, J.M.Martinez, Y.Yuan: An increasing-angle property of the conjugate gradient method and the implementation of large-scale minimization algorithms with line searches. Numerical Linear Algebra with Applications 10 (2003) 323-334.

[33] Y.Dai, Y.Yuan: Convergence properties of the conjugate descent method. Advances in Mathematics 25 (1996) 552-562.

[34] Y.Dai, Y.Yuan: A nonlinear conjugate gradient method with a strong global convergence property. SIAM J. Optimization 10 (1999) 177-182.

[35] J.W.Daniel: The conjugate gradient method for linear and nonlinear operator equations. SIAM J. Numerical Analysis 4 (1967) 10-26.

[36] W.C.Davidon: Optimally conditioned optimization algorithms without line searches. Mathematical Programming 9 (1975) 1-30.

[37] W.C.Davidon: Conic approximations and collinear scalings for optimizers. SIAM J. Numerical Analysis 17 (1980), 268-281.

[38] W.C.Davidon: Variable metric method for minimization. SIAM J. Optimization, 1 (1991), 1-17.

[39] R.S.Dembo, T.Steihaug: Truncated Newton algorithms for large-scale optimization. Mathematical Programming 26 (1983) 190-212.

[40] J.E.Dennis, H.H.W.Mei: An unconstrained optimization algorithm which uses function and gradient values. Report No. TR-75-246. Dept. of Computer Science, Cornell University, 1975.

[41] J.E.Dennis, J.J.Moré: A characterization of superlinear convergence and its application to quasi-Newton methods. Mathematics of Computation 28 (1974) 549-560.

[42] J.E.Dennis, R.B.Schnabel: Least change secant updates for quasi-Newton methods. Report No. TR78-344, Dept. of Computer Sci., Cornell University, Ithaca, 1978.

[43] L.C.W.Dixon: Quasi-Newton algorithms generate identical points. Mathematical Programming 2 (1972) 383-387.

[44] S.C.Eisenstat, M.C.Gursky, M.H.Schultz, A.H.Sherman: The Yale matrix package 1, the symmetric codes. Report 114, Yale University Department of Computer Science, 1977.

[45] S.C.Eisenstat, M.C.Gursky, M.H.Schultz, A.H.Sherman: The Yale matrix package 2, the non-symmetric codes. Report 114, Yale University Department of Computer Science, 1977.

[46] G.Fasano: Planar-CG methods and matrix tridiagonalization in large scale unconstrained optimization. In: High Performance algorithms and Software for Nonlinear Optimization (G.DiPillo, A.Murli eds.) pp. 238-258, Kluwer, Academic Press, 2003.

bibitemfas04 G.Fasano: Conjugate gradient (CG)-type method for the solution of Newton's equation within optimization frameworks. Optimization Methods and Software19 (2004) 267-290.

[47] G.Fasano: Planar-conjugate gradient algorithm for large scale unconstrained optimization. Part 1 – Theory. Journal of Optimization Theory and Applications 125 (2005) 523-541.

[48] G.Fasano: Planar-conjugate gradient algorithm for large scale unconstrained optimization. Part 2 – Applications. Journal of Optimization Theory and Applications 125 (2005) 543-558.

[49] G.Fasano, S.Lucidi: A nonmonotone truncated NewtonKrylov method exploiting negative curvature directions, for large scale unconstrained optimization. Optimization Letters 3 (2009) 521-535.

[50] G.Fasano, M.Roma: Preconditioning Newton-Krylov methods in nonconvex large scale optimization. Computational Optimization and Applications 56 (2012) 253-290.

[51] G.Fasano, M.Roma: AINVK: a class of approximate inverse preconditioners based on Krylov-subspace methods, for large indefinite linear systems. Preprint, to appear.

[52] R.Fletcher: A new approach to variable metric algorithms. Computer J. 13 (1970) 317-322.

[53] R.Fletcher: Practical methods of optimization. Wiley, New York, 1987.

[54] R.Fletcher: A new variational result for quasi-Newton formulae. SIAM J. Optimization 1 (1991) 18-21.

[55] R.Fletcher: An optimal positive definite update for sparse Hessian matrices. SIAM J. Optimization 5 (1995) 192-218.

[56] R.Fletcher, M.J.D.Powell: A rapidly convergent descent method for minimization. Computer J. 6 (1963) 163-168.

[57] R.Fletcher, C.M.Reeves: Function minimization by conjugate gradients. Computer J. 7 (1964) 149-154.

[58] R.Fletcher, C.Xu: Hybrid methods for nonlinear least squares. IMA J. Numerical Analysis 7 (1987) 371-389.

[59] J.A.Ford, I.A.Moghrabi: Multi-step quasi-Newton methods for optimization. J. Computational and Applied Mathematics 50 (1994) 305-323.

[60] A.H.Gebremedhin, F.Manne, A.Pothen: What color is your Jacobian. Graph coloring for computing derivatives. SIAM Review 47 (2005) 629-705.

[61] A.H.Gebremedhin, a.Tarafdar, F.Manne, A.Pothen: New cyclic and star coloring algorithms with application to computing Hessians. SIAM J. Scientific Computing 29 (2007) 1042-1072.

[62] A.George, W.H.Liu: Computer Solution of Large Sparse Positive Definite Systems. Prentice Hall, Englewood Cliffs, New Jersey 1984.

[63] J.C.Gilbert, J.Nocedal: Global convergence properties of conjugate gradient methods for optimization. SIAM J. Optimization 2 (1992), 21-42.

[64] P.E.Gill, M.W.Leonard: Reduced-Hessian quasi-Newton methods for unconstrained optimization. SIAM J. Optimization, 12 (2001), 209-237.

[65] P.E.Gill, M.W.Leonard: Limited-memory reduced-Hessian methods for large-scale unconstrained optimization. SIAM J. Optimization 14 (2003), 380-401.

[66] P.E.Gill, W.Murray: Newton type methods for unconstrained and linearly constrained optimization. Mathematical Programming 7 (1974) 311-350.

[67] P.E.Gill, W.Murray, M.A.Saunders: Methods for computing and modifying LDV factors of a matrix. Mathematics of Computation 29 (1975) 1051-1077.

[68] D.Goldfarb: A family of variable metric algorithms derived by variational means. Math Comput. 24 (1970) 23-26.

[69] D.Goldfarb: Curvilinear path steplength algorithms for minimization which use directions of negative curvature. Mathematical Programming 18 (1980) 31-40.

[70] N.I.M.Gould, S.Lucidi, M.Roma, P.L.Toint: Exploiting negative curvature directions in linesearch methods for unconstrained optimization. Optimization Methods and Software 14 (2000) 75-98.

[71] S.Gratton, A.Sartenaer, J. Tshimanga: On a class of limited memory preconditioners for large scale linear systems with multiple right-hand sides. SIAM J. Optimization 21 (2011) 912-935.

[72] A.Greenbaum: Iterative Methods for Solving Linear Systems. SIAM, Philadelphia, 1997.

[73] J.Greenstadt: Variations on variable metric methods. Mathematics of Computation 24 (1970) 1-18.

[74] J.Greenstadt: Reminiscences on the development of the variational approach to Davidon's variable-metric method. Mathematical Programming 87 (2000) 265-280.

[75] A.Griewank, P.L.Toint: Partitioned variable metric updates for large-scale structured optimization problems. Numererische Mathematik 39 (1982) 119-137.

[76] A.Griewank, P.L.Toint: Local convergence analysis for partitioned quasi-Newton updates. Numerische Mathematik 39 (1982) 429-448.

[77] A.Griewank, A.Walther: Evaluating Derivatives. SIAM, Philadelphia, 2008.

[78] W.W.Hager, H.Zhang: A new conjugate gradient method with guaranted descent and an efficient line search. SIAM J. on Optimization, 16 (2005) 170-192.

[79] W.W.Hager, H.Zhang: Algorithm 851: CG–DESCENT, a conjugate gradient method with guaranted descent. ACM Transactions on Mathematical Software 32 (2006) 113-137.

[80] M.R.Hestenes: Conjugate Direction Methods. Springer-Verlag, Berlin 1980.

[81] M.R.Hestenes, C.M.Stiefel: Methods of conjugate gradient for solving linear systems. J. Research NBS 49 (1964) 409-436.

[82] N.J.Higham: Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia, 2002.

[83] S.Hoshino: A formulation of variable metric methods. J. Inst. Math. Appl. 10 (1972) 394-403.

[84] H.Y.Huang: Unified approach to quadratically convergent algorithms for function minimization. J. Optimizaton Theory and Applications 5 (1970) 405-423.

[85] H.Huang, S.Lin: A modified Wei-Yao-Liu conjugate gradient method for unconstrained optimization. Applie Mathematics and Computation 231 (2014) 179-186.

[86] J.Huschens: On the use of product structure in secant methods for nonlinear least squares. SIAM J. Optimization 4 (1994) 108-129.

[87] C.X.Kou, Y.H.Dai: A modified self-scaling memoryless Broyden-Fletcher-Goldfarb-Shanno method for unconstrained optimization. J. Optimization Theory and Applications xx (2013) xx-xx.

[88] K.Levenberg: A method for the solution of certain nonlinear problems in least squares. Quaterly Applied Mathematics 2 (1944) 164-168.

[89] D.C.Liu, J.Nocedal: On the limited memory BFGS method for large scale optimization. Mathematical Programming 45 (1989) 503-528.

[90] Y.Liu, C.Storey: Efficient generalized conjugate gradient algorithms. Part 1 – Theory. J. Optimization Theory and Applications 69 (1991) 129-137.

[91] D.G.Luenberger: Hyperbolic pairs in the method of conjugate gradients. SIAM J. Applied Mathematics 17 (1969) 1263-1267.

[92] L.Lukšan: Quasi-Newton Methods without Projections for Unconstrained Minimization. Kybernetika, Vol.18, 1982, No.4, pp.290-306.

[93] L.Lukšan: Quasi-Newton Methods without Projections for Linearly Constrained Minimization. Kybernetika 18 (1982) 307-319.

[94] L.Lukšan: Variable Metric Method with Limited Storage for Large Scale Unconstrained Minimization. Kybernetika 18 (1982) 517-528.

[95] L.Lukšan: Conjugate Direction Algorithms For Extended Conic Functions. Kybernetika 22 (1986) 31-46.

[96] L.Lukšan: Computational experience with improved variable metric methods for unconstrained minimization. Kybernetika 26 (1990) 415-431.

[97] L.Lukšan: Computational experience with improved conjugate gradient methods for unconstrained minimization. Kybernetika 28 (1992) 249-262.

[98] L.Lukšan: Variationally derived scalling and variable metric updates from the preconvex part of the Broyden family. J. Optimizaton Theory and Applications 73 (1992) 299-307.

[99] L.Lukšan: Computational experience with known variable metric updates. J. Optimizaton Theory and Applications 83 (1994) 27-47.

[100] L.Lukšan: Combined trust region methods for nonlinear least squares. Kybernetika 32 (1996) 121-138.

[101] L.Lukšan: Hybrid methods for large sparse nonlinear least squares. J. Optimizaton Theory and Applications 89 (1996) 575-595.

[102] L.Lukšan, C.Matonoha, J.Vlček: A shifted Steihaug-Toint method for computing trust-region step. Technical Report V-914. Prague, ICS AS CR, 2004.

[103] L.Lukšan, C.Matonoha, J.Vlček: On Lagrange multipliers of trust-region subproblems. BIT Numerical Analysis 48 (2008a) 763-768.

[104] L.Lukšan, C.Matonoha, J.Vlček: Computational experience with modified conjugate gradient methods for unconstrained optimization. Technical Report V-1038. Prague, ICS AS CR 2008.

[105] L.Lukšan, C.Matonoha, J.Vlček J.: Algorithm 896: LSA: Algorithms for Large-Scale Optimization. ACM Transactions on Mathematical Software 36 (2009) No. 3.

[106] L.Lukšan, C.Matonoha, J.Vlček: Sparse test problems for unconstrained optimization. Technical Report V-1064. Prague, ICS AS CR 2010.

[107] L.Lukšan, C.Matonoha, J.Vlček: Modified CUTE Problems for Sparse Unconstrained Optimization. Technical Report V-1081. Prague, ICS AS CR 2010.

[108] L.Lukšan, C.Matonoha, J.Vlček: Band preconditioners for the matrix free truncated Newton method. Technical Report V-1079. Prague, ICS AS CR 2010.

[109] L.Lukšan, E.Spedicato: Variable metric methods for unconstrained optimization and nonlinear least squares. Journal of Computational and Applied Mathematics 124 (2000) 61-93.

[110] Lukšan L., Tůma M., Vlček J., Ramešová N., Šiška M., Hartman J., Matonoha C.: UFO 2014. Interactive System for Universal Functional Optimization. Technical Report V-1218. Prague, ICS AS CR 2014.

[111] L.Lukšan, J.Vlček: Truncated trust region methods based on preconditioned iterative subalgorithms for large sparse systems of nonlinear equations. J. Optimization Theory and Applications 95 (1997) 637-658.

[112] L.Lukšan, J.Vlček: Subroutines for testing large sparse and partially separable unconstrained and equality constrained optimization problems. Technical Report V-767. Prague, ICS AS CR 1998.

[113] L.Lukšan, J.Vlček: Computational experience with globally convergent descent methods for large sparse systems of nonlinear equations. Optimization Methods and Software 8 (1998) 201-223.

[114] L.Lukšan, J.Vlček: Test problems for unconstrained optimization. Technical Report V-897. Prague, ICS AS CR, 2003.

[115] L.Lukšan, J.Vlček: Recursive form of general limited memory variable metric methods. Kybernetika 49 (2013) 224-235 .

[116] L.Lukšan, J.Vlček: Efficient tridiagonal preconditioner for the matrix-free truncated Newton method. Applied Mathematics and Computation 235 (2014) 394-407.

[117] D.W.Marquardt: An algorithm for least-squares estimation of nonlinear parameters: SIAM J. Applied Mathematics 11 (1963) 431-441.

[118] E.S.Marwil: Exploiting sparsity in Newton-like methods. Ph.D. Thesis, Cornell University, Ithaca 1978.

[119] H.Matthies, G.Strang: The solution of nonlinear finite element equations. International Journal for Numerical Methods in Engineering 14 (1979) 1613-1623.

[120] J.L.Morales, J.Nocedal: Automatic preconditioning by limited memory quasi-Newton updating. SIAM J. Optimization 10 (2000) 1079-1096.

[121] J.J.Moré, D.C.Sorensen: On the use of directions of negative curvature in a modified Newton method. Mathematical Programming 16 (1979) 31-40.

[122] J.J.Moré, D.C.Sorensen: Computing a trust region step. Report ANL-81-83, Argonne National Laboratory, 1981.

[123] S.G.Nash: Newton-type minimization via Lanczos method. SIAM J. Numerical Analysis 21 (1984) 770-788.

[124] S.G.Nash: Preconditioning of truncated-Newton methods. SIAM J. Scientific and Statistical Computation 6 (1985) 599-616.

[125] J.Nocedal: Updating quasi-Newton matrices with limited storage. Mathematics of Computation 35 (1980) 773-782.

[126] J.Nocedal, S.J.Wright: Numerical optimization. Springer-Verlag, New York, 2006.

[127] J.Nocedal, Y.Yuan: Analysis of a self-scaling quasi-Newton method. Mathematical Programming 61 (1993) 19-37.

[128] D.P.O'Leary: A discrete Newton algorithm for minimizing a function of many variables. Mathematical Programming 23 (1983) 20-33.

[129] S.S.Oren, D.G.Luenberger: Self scaling variable metric (SSVM) algorithms. Part 1 – criteria and sufficient condition for scaling a class of algorithms. Part 2 – implementation and experiments. Management Sci. 20 (1974) 845-862, 863-874.

[130] S.S.Oren, E. Spedicato: Optimal conditioning of self scaling variable metric algorithms. Mathematical Programming 10 (1976) 70-90.

[131] M.R.Osborne, L.P.Sun, A new approach to the symmetric rank-one updating algorithm. IMA J. of Numerical Analysis 19 (1999) 497-507.

[132] C.C.Paige, M.A.Saunders: Solution of sparse indefinite systems of linear equations. SIAM J. Numerical Analysis 12 (1975), 617-629.

[133] J.M.Perry: A class of conjugate gradient algorithms with a two-step variable-metric memory. Discussion Paper 269, Center for Mathematical Studies in Economics and Management Sciences, Northwestern University, Evanston, Illinois, 1977.

[134] E.Polak, G.Ribière: Note sur la convergence de directions conjugeés. Rev. Francaise Informat. Recherche Opertionelle, 3e Annee 16 (1969) 35-43.

[135] B.T.Polyak: The conjugate gradient method in extreme problems. USSR Computational Mathematics and Mathematical Physics 9 (1969) 94-112.

[136] M.J.D.Powell: Some global convergence properties of a variable metric algorithm for minimization without exact line searches. In: Nonlinear Programming (R.W.Cottle, C.E.Lemke eds.), SIAM, Philadelphia, 1976.

[137] M.J.D.Powell: On the global convergence of trust region algorithms for unconstrained minimization. Mathematical Programming 29 (1984) 297-303.

[138] M.J.D.Powell: Quadratic termination properties of Davidon's new variable metric algorithm. Mathematical Programming 12 (1977) 141-147.

[139] M.J.D.Powell, P.L.Toint: On the estimation of sparse Hessian matrices. SIAM J. Numerical Analysis 16 (1979) 1060-1074.

[140] M.Roma: Dynamic scaling based preconditioning for truncated Newton methods in large scale unconstrained optimization. Optimization Methods and Software 20 (2005) 693-713.

[141] D.F.Shanno: Conditioning of quasi-Newton methods for function minimization. Mathematics of Computation 24 (1970) 647-656.

[142] D.F.Shanno, K.J.Phua: Matrix conditioning and nonlinear optimization. Mathematical Programming 14 (1978) 144-160.

[143] D.Siegel: Updating of conjugate direction matrices using members of Broyden's family. Mathematical programming 60 (1993) 167-185.

[144] D.Siegel: Modifying the BFGS update by a new column scaling technique. Mathematical Programming66 (1993) 45-78.

[145] T.Steihaug: Local and superlinear convergence for truncated iterated projections methods. Mathematical Programming 27 (1983) 176-190.

[146] T.Steihaug: The conjugate gradient method and trust regions in large-scale optimization. SIAM J. Numerical Analysis 20 (1983) 626-637.

[147] T.Steihaug: Damped inexact quasi-Newton methods. Report MASC TR 81-3, Department of Mathematical Sciences, Rice University, Houston, Texas 1984.

[148] J.Stoer: On the convergence rate of imperfect minimization algorithms in Broydens β-clas. Mathematical programming 9 (1975) 313-335.

[149] J.Stoer: On the relation between quadratic termination and convergence properties of minimization algorithms. Numerische Mathematik 28 (1977) 343-366.

[150] W.Sun, Y.Yuan: Optimization theory and methods. Nonlinear programming. Springer-Verlag, New York, 2006.

[151] P.L.Toint: On sparse and symmetric matrix updating subject to a linear equation. Mathematics of Computation 31 (1977) 954-961.

[152] P.L.Toint: Towards an efficient sparsity exploiting Newton method for minimization. In: Sparse Matrices and Their Uses (I.S.Duff, ed.), Academic Press, London 1981, 57-88.

[153] P.L.Toint: Global convergence of the partitioned BFGS algorithm for convex partially separable optimization. Mathematical Programming 36 (1986) 290-306.

[154] M.Tůma: A note on direct methods for approximations of sparse Hessian matrices. Aplikace matematiky 33 (1988) 171-176.

[155] M.Tůma: Sparse fractioned variable metric updates. Report No. 497, Institute of Computer and Information Sciences, Czechoslovak Academy of Sciences, Prague 1991.

[156] J.Vlček, L.Lukšan: New variable metric methods for unconstrained minimization covering the large-scale case. Technical Report V-876. Prague, ICS AS CR, 2002.

[157] J.Vlček, L.Lukšan: Additional properties of shifted variable metric methods. Technical Report V-899. Prague, ICS AS CR, 2004.

[158] J.Vlček, L.Lukšan: Shifted limited-memory variable metric methods for large-scale unconstrained minimization. J. of Computational and Applied Mathematics, 186 (2006) 365-390.

[159] J.Vlček, L.Lukšan: New class of limited-memory variationally-derived variable metric methods. Technical Report V-973. Prague, ICS AS CR, 2006.

[160] J.Vlček, L.Lukšan: Limited-memory projective variable metric methods for unconstrained minimization. Technical Report V-1036. Prague, ICS AS CR 2008.

[161] J.Vlček, L.Lukšan: Transformations enabling to construct limited-memory Broyden class methods. Technical Report V-1037. Prague, ICS AS CR 2008.

[162] J.Vlček, L.Lukšan: A conjugate directions approach to improve the limited-memory BFGS method. Technical Report V-1120. Prague, ICS AS CR 2011.

[163] J.Vlček, L.Lukšan: Modifications of the limited-memory BNS method for better satisfaction of previous quasi-Newton conditions. Technical Report V-1127. Prague, ICS AS CR 2011.

[164] J.Vlček, L.Lukšan: A conjugate directions approach to improve the limited-memory BFGS method. Applied Mathematics and Computation 219 (2012) 800-809.

[165] J.Vlček, L.Lukšan: Generalizations of the limited-memory BFGS method based on quasi-product form of update. Journal of Computational and Applied Mathematics 241 (2013) 116-129.

[166] J.Vlček, L.Lukšan: A modified limited-memory BNS method for unconstrained minimization based on the conjugate directions idea. Techical Report V-1203, Prague, ICS AS CR, 2014

[167] Z.Wei, S.Yao, L.Liu: The convergence properties of some new conjugate gradient methods. Applied Mathematics and Computation 183 (2006) 1341-1350.

[168] H.Yabe, T.Takahashi: Factorized quasi-Newton methods for nonlinear least squares problems. Mathematical Programming 51 (1991) 75-100.

[169] N.Yamashita: Sparse quasi-Newton updates with positive definite matrix completion. Mathematical Programming 115 (2008) 1-30.

[170] G.H.Yu, L.Guan and W.Chen: Spectral conjugate gradient methods with sucient descent property for large-scale unconstrained optimization. Optimization Methods and Software 23 (2008) 275-293.

[171] G.Yu, Y.Zhao, Z.Wei: A descent nonlinear conjugate gradient method for large-scale unconstrained optimization. Applied Mathematics and Computation 187 (2007) 636-643.

[172] J.Z.Zhang, N.Y.Deng, L.H.Chen: New quasi-Newton equation and related methods for unconstrained optimization. J. Optimization Theory and Applications, Vol.102, 1999, pp.147-167.

[173] L.Zhang, W.Zhou: Two descent hybrid conjugate gradient methods for optimization. Journal of Computational and Applied Mathematics 216 (2008) 251-264.

[174] L.Zhang, W.Zhou, D.Li: Global convergence of a modified Fletcher-Reeves conjugate gradient method with Armijo-type line search. Numerische Mathematik 104 (2006) 561-572.

[175] L.Zhang, W.Zhou, D.Li: A descent modified Polak-Ribiere-Polyak conjugate gradient method and its global convergence. IMA J. of Numerical Analysis 26 (2006) 629-640.