národní
úložiště
šedé
literatury

Numerické optimalizační metody. Nepodmíněná minimalizace

Lukšan, Ladislav
2009

Dostupný z http://www.nusl.cz/ntk/nusl-41638

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

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

Datum stažení: 05.06.2024

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. 1058

Prosinec 2009

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. 1058

Prosinec 2009

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. AVOZ10300504

Contents

1 Úvod **4**
- 1.1 Základní pojmy . . . 4
- 1.2 Podmínky optimality . . . 7
- 1.3 Základní pojmy z teorie konvergence . . . 9
- 1.4 Základní optimalizační metody . . . 14

2 Metody spádových směrů **16**
- 2.1 Základní vlastnosti metod spádových směrů . . . 16
- 2.2 Globální konvergence . . . 19
- 2.3 Asymptotická rychlost konvergence . . . 26
- 2.4 Výběr délky kroku . . . 34
- 2.5 Nemonotonní metody spádových směrů . . . 37

3 Metody sdružených gradientů **41**
- 3.1 Základní vlastnosti metod sdružených gradientů . . . 41
- 3.2 Globální konvergence . . . 43
- 3.3 Přerušované metody sdružených gradientů . . . 47
- 3.4 Asymptotická rychlost konvergence . . . 48
- 3.5 Modifikace a implementace metod sdružených gradientů . . . 56
- 3.6 Předpodmíněná metoda sdružených gradientů pro řešení soustav lineárních rovnic . . . 63

4 Metody s proměnnou metrikou **70**
- 4.1 Základní vlastnosti metod s proměnnou metrikou . . . 70
- 4.2 Součinový tvar metod s proměnnou metrikou . . . 82
- 4.3 Variační odvození metod s proměnnou metrikou . . . 95
- 4.4 Výběr parametrů (škálování a korekce) . . . 103
- 4.5 Globální konvergence . . . 112
- 4.6 Asymptotická rychlost konvergence . . . 114
- 4.7 Aktualizace trojúhelníkového rozkladu . . . 120
- 4.8 Modifikace a implementace metod s proměnnou metrikou . . . 123

5 Metody s lokálně omezeným krokem **132**
- 5.1 Základní vlastnosti metod s lokálně omezeným krokem . . . 132
- 5.2 Metody s optimálním lokálně omezeným krokem . . . 142
- 5.3 Newtonova metoda s lokálně omezeným krokem . . . 144
- 5.4 Nemonotonní metody s lokálně omezeným krokem . . . 147
- 5.5 Kombinované metody s lokálně omezeným krokem . . . 149

6 Výpočet lokálně omezeného kroku **153**
- 6.1 Výpočet optimálního lokálně omezeného kroku . . . 153
- 6.2 Využití směru největšího spádu (metody psí nohy) . . . 156
- 6.3 Nepřesné metody s lokálně omezeným krokem . . . 160
- 6.4 Použití symetrické Lanczosovy metody . . . 162
- 6.5 Posunuté nepřesné metody s lokálně omezeným krokem . . . 166
- 6.6 Maticové rozklady pro symetrické indefinitní matice . . . 169

7 Metody pro minimalizaci součtu čtverců **173**
- 7.1 Gaussova–Newtonova metoda . . . 177
- 7.2 Použití kvazinewtonovských aktualizací . . . 178
- 7.3 Řešení lineární úlohy nejmenších čtverců . . . 184

8 Metody pro rozsáhlé husté úlohy **189**
8.1 Metody s proměnnou metrikou s omezenou pamětí . . . 189
8.2 Metody redukovaných Hessiánů s omezenou pamětí . . . 207
8.3 Posunuté metody s proměnnou metrikou s omezenou pamětí . . . 213
8.4 Diferenční verze Newtonovy metody pro husté úlohy . . . 221
8.5 Numerické porovnání . . . 235

9 Metody pro rozsáhlé řídké úlohy **236**
9.1 Diferenční verze Newtonovy metody pro řídké úlohy . . . 236
9.2 Metody s proměnnou metrikou pro řídké úlohy . . . 240

10 Metody pro rozsáhlé separovatelné úlohy **247**
10.1 Diferenční verze Newtonovy metody pro separovatelné úlohy . . . 247
10.2 Metody s proměnnou metrikou pro separovatelné úlohy . . . 249
10.3 Modifikace Gaussovy–Newtonovy metody pro řídký součet čtverců . . . 250
10.4 Iterační řešení rozsáhlých lineárních úloh nejmenších čtverců . . . 252

11 Metody pro řešení soustav nelineárních rovnic **258**
11.1 Základní vlastnosti metod pro řešení soustav nelineárních rovnic . . . 258
11.2 Metody spádových směrů . . . 260
11.3 Metody s lokálně omezeným krokem . . . 266
11.4 Newtonova metoda . . . 271
11.5 Kvazinewtonovské metody . . . 273
11.6 Nemonotonní kvazinewtonovské metody . . . 277
11.7 Sdružené kvazinewtonovské metody . . . 280
11.8 Tenzorové metody . . . 283
11.9 Aktualizace ortogonálního rozkladu . . . 288

12 Metody pro rozsáhlé soustavy nelineárních rovnic **289**
12.1 Kvazinewtonovské metody s omezenou pamětí . . . 289
12.2 Diferenční verze Newtonovy metody pro husté úlohy . . . 292
12.3 Diferenční verze Newtonovy metody pro řídké úlohy . . . 293
12.4 Kvazinewtonovské metody pro řídké úlohy . . . 294
12.5 Sdružené kvazinewtonovské metody pro řídké úlohy . . . 298
12.6 Metody založené na aktualizaci nesymetrického trojúhelníkového rozkladu . . . 301
12.7 Nedokonalé diferenční verze Newtonovy metody . . . 302
12.8 Iterační řešení systémů lineárních rovnic s nesymetrickou maticí . . . 302
12.9 Metody s lokálně omezeným krokem . . . 313

13 Optimalizace dynamických systémů **315**
13.1 Přímý výpočet gradientu . . . 316
13.2 Zpětný výpočet gradientů . . . 316
13.3 Přímý výpočet Hessovy matice . . . 317
13.4 Přímá aproximace Hessovy matice (součet čtverců) . . . 318

14 Automatické derivování **318**

15 Základy nehladké analýzy **320**
15.1 Konvexní množiny . . . 320
15.2 Konvexní funkce . . . 333
15.3 Lipschitzovské funkce . . . 341
15.4 Lipschitzovská zobrazení . . . 350
15.5 Polohladká zobrazení . . . 356

16 Metody pro řešení soustav nehladkých rovnic **362**
16.1 Newtonova metoda 362
16.2 Aplikace nehladkých rovnic 367

17 Metody pro nehladkou optimalizaci **369**
17.1 Svazkové metody 369

1 Úvod

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, jsou uváděny bez důkazu. Mnoho chybějících důkazů lze nalézt v knize: L.Lukšan, Metody s proměnnou metrikou, Academia, Praha 1991.

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. Při vyšetřování konvergence optimalizačních metod budeme často používat tyto předpoklady.

Předpoklad 1 *Funkce $F : \mathcal{D}_F \to R$ je zdola omezená, takže existuje konstanta $\underline{F}$ taková, že*

$$F(x) \geq \underline{F} \quad \forall x \in \mathcal{D}_F. \tag{F1}$$

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 2 *Množina*

$$\mathcal{D}_F(\overline{F}) = \{x \in \mathcal{D}_F : F(x) \leq \overline{F}\} \tag{F2}$$

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

Poznámka 2 Předpoklad (F2) je opět logický. Je-li funkce $F : \mathcal{D}_F \to R$ 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 12). 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 $\{x \in R^n : |x| \leq 1\}$). 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 generované optimalizační metodou.

Předpoklad 3 *Funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$, kde $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$ je otevřená množina, má lipschitzovské první derivace na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$\|g(x_2) - g(x_1)\| \leq \overline{G}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{F3}$$

Poznámka 3 V (F3) je někdy výhodné předpokládat že množina $\mathcal{D}$ je konvexní (definice 60). Je to proto, že při vyšetřování optimalizační metody generující posloupnost x_i, $i \in N$, potřebujeme, aby (F3) platilo 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 (a nemá lipschitzovské první derivace) na úsečce spojující body x_1 a x_2. Předpoklad konvexity množiny $\mathcal{D}$ vylučuje případy, kdy funkce F má póly v $\operatorname{conv} \mathcal{D}_F$. Poznamenejme, že optimalizační metoda většinou funguje i pro funkce s póly v $\operatorname{conv} \mathcal{D}_F$ (neboť generuje posloupnost bodů ležících v $\mathcal{D}_F(\overline{F})$), může se ale stát, že některý mezilehlý bod leží v blízkosti pólu a výpočet skončí na selhání počítačové aritmetiky. Poznamenejme, že pokud je splněna podmínka (F2), budeme předpokládat, že množina $\mathcal{D}$ je omezená.

Předpoklad 4 *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$ je otevřená množina, má omezené druhé derivace na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$|d^T G(x) d| \leq \overline{G}\|d\|^2 \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{F4}$$

Podmínka (F4) je ekvivalentní podmínce $\|G(x)\| \leq \overline{G}$ $\forall x \in \mathcal{D}$.

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

Předpoklad 5 *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$ je otevřená množina, je stejnoměrně konvexní na $\mathcal{D}$, takže existuje konstanta $\underline{G} > 0$ taková, že*

$$d^T G(x) d \geq \underline{G}\|d\|^2 \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{F5}$$

Předpoklad 6 *Funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$, kde $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$ je otevřená konvexní množina, má lipschitzovské druhé derivace na $\mathcal{D}$, takže existuje konstanta $\overline{L} > 0$ taková, že*

$$\|G(x_2) - G(x_1)\| \leq \overline{L}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{F6}$$

Poznámka 5 Podmínky (F4)–(F6) jsou často zbytečně silné. Studujeme-li chování iteračního procesu v okolí minima $x^* \in R^n$, stačí předpokládat, že funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná v nějakém okolí bodu x^* a platí ($\overline{\text{F4}}$)–($\overline{\text{F6}}$) (místo $\mathcal{D}$ používáme $\mathcal{B}(x^*, \varepsilon) = \{x \in R^n :\| x - x^* \| < \varepsilon\}$).

Předpoklad 7 *Funkce $F : \mathcal{D}_F \to R$ je dvakrát spojiě diferencovatelná v okolí bodu $x^* \in R^n$. Pak pro libovolnou konstantu $0 < \overline{\lambda}(G(x^*)) < \overline{G}$ existuje číslo $\varepsilon > 0$ takové, že*

$$|d^T G(x) d| \leq \overline{G}\|d\|^2 \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{$\overline{\text{F4}}$}$$

Předpoklad 8 *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*

$$d^T G(x) d \geq \underline{G} \|d\|^2 \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{$\overline{\text{F5}}$}$$

Předpoklad 9 *Funkce $F : \mathcal{D}_F \to R$ je dvakrát spojitě diferencovatelná v okolí bodu $x^* \in R^n$ a existuje konstanta $\overline{L}$ a číslo $\varepsilon > 0$ tak, že platí*

$$\|G(x) - G(x^*)\| \leq \overline{L} \|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{$\overline{\text{F6}}$}$$

Poznámka 6 Jestliže $x_i \to x^*$, existuje k danému číslu $\varepsilon > 0$ index $k \in N$ takový, že $x_i \in \mathcal{B}(x^*, \varepsilon)$ pokud $i \geq k$. Pak, omezíme-li se na $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$, podmínka ($\overline{\text{F4}}$) implikuje (F4) a podmínka ($\overline{\text{F5}}$) implikuje (F5). Abychom nemuseli stále ověřovat zda pro daný index $i \in N$ již platí $x_i \in \mathcal{B}(x^*, \varepsilon)$, budeme často používat podmínky (F4) a (F5) (dokonce s $\mathcal{D} = R^n$) místo ($\overline{\text{F4}}$) a ($\overline{\text{F5}}$). Tím se formálně zjednoduší a zpřehlední většina důkazů aniž by došlo k újmě na obecnosti.

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 $x + d \in R^n$.

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 $0 \leq \tilde{\lambda} \leq 1$).

Použijeme-li tvrzení 1 a (F3), dostaneme

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

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 $0 \leq \tilde{\lambda} \leq 1$).

Použijeme-li tvrzení 2 a (F4), dostaneme

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

Použijeme-li tvrzení 2 a (F5), dostaneme

$$F(x + d) - F(x) \geq d^T g(x) + \frac{1}{2} \underline{G} \|d\|^2. \tag{3}$$

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 (F3) nebo tvrzení 3 a (F4), dostaneme

$$\|g(x+d)-g(x)\| \le \overline{G}\|d\|, \tag{4}$$

$$d^T(g(x+d)-g(x)) \le \overline{G}\|d\|^2. \tag{5}$$

Použijeme-li tvrzení 3 a (F5), dostaneme

$$\|g(x+d)-g(x)\| \ge \underline{G}\|d\|, \tag{6}$$

$$d^T(g(x+d)-g(x)) \ge \underline{G}\|d\|^2. \tag{7}$$

Důkaz posledních dvou nerovností:

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

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

Uvedeme nyní bez komentáře dvě definice, které budeme často používat.

Definice 1 *Řekneme, že funkce $F : \mathcal{D}_F \to R$ je konvexní na množině $C \subset \mathcal{D}_F \subset R^n$, jestliže*

$$F\left(\sum_{i=1}^m \lambda_i x_i\right) \le \sum_{i=1}^m \lambda_i F(x_i),$$

pokud $m \ge 1$, $x_i \in C$, $\lambda_i \ge 0$, $1 \le i \le m$ a

$$\sum_{i=1}^m \lambda_i = 1.$$

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

Definice 2 *Řekneme, že funkce $F : \mathcal{D}_F \to R$ je lipschitzovská na množině $C \subset \mathcal{D}_F \subset R^n$, jestliže existuje konstanta $\overline{L} > 0$ taková, že*

$$|F(x_2) - F(x_1)| \le \overline{L}\|x_2 - x_1\|,$$

pokud $x_1 \in C$ a $x_2 \in C$.

Poznámka 7 Konvexní a lipschitzovské finkce jsou podrobně studovány v kapitole 15. Zde pouze připomeneme, že dvakrát spojitě diferencovatelná funkce $F \in \mathcal{C}^2 : C \to R$ je konvexní na množině C, je-li její Hessova matice pozitivně semidefinitní na C.

1.2 Podmínky optimality

Definice 3 *Řekneme, že bod $x^* \in R^n$ je lokálním minimem funkce $F : \mathcal{D}_F \to R$, jestliže existuje číslo $\varepsilon > 0$ takové, že*

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

Jestliže navíc $F(x^*) < F(x)$ pokud $x^* \ne x$, 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 8 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 1). 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 $4x \leq 4x(2 + \cos(1/x)) \leq 12x$, má funkce $F'(x)$ na 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 1 *(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 $g^* = g(x^*) \neq 0$. Jelikož $F \in \mathcal{C}^1$ na $\mathcal{B}(x^*, \varepsilon)$, existuje číslo $\overline{\alpha} > 0$, takové, že $x^* - \alpha g^* \in \mathcal{B}(x^*, \varepsilon)$ a $(g^*)^T g(x^* - \alpha g^*) \geq (g^*)^T g^*/2$, pokud $0 \leq \alpha \leq \overline{\alpha}$ (plyne to ze spojitosti gradientu $g(x^* - \alpha g^*)$). Nechť $0 < \alpha \leq \overline{\alpha}$. Pak podle věty o střední hodnotě platí $F(x^* - \alpha g^*) = F(x^*) - \alpha (g^*)^T g(x^* - \tilde{\alpha} g^*)$, kde $0 \leq \tilde{\alpha} \leq \alpha \leq \overline{\alpha}$, takže

$$F(x^* - \alpha g^*) = F(x^*) - \alpha (g^*)^T g(x^* - \tilde{\alpha} g^*) \leq F(x^*) - \alpha (g^*)^T g^*/2 < F(x^*),$$

což je ve sporu s definicí 3.

(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, takže $\lambda^* < 0$, kde λ^* je nejmenší vlastní číslo matice $G(x^*)$. Nechť v^* je vlastní vektor matice $G(x^*)$ příslušný vlastnímu číslu λ^*. Jelikož $F \in \mathcal{C}^2$ na $\mathcal{B}(x^*, \varepsilon)$, existuje číslo $\overline{\alpha} > 0$, takové, že $x^* + \alpha v^* \in \mathcal{B}(x^*, \varepsilon)$ a $(v^*)^T G(x^* + \alpha v^*) v^* \leq \lambda^* (v^*)^T v^*/2$, pokud $0 \leq \alpha \leq \overline{\alpha}$ (plyne to ze spojitosti Hessovy matice $G(x^* + \alpha v^*)$). Nechť $0 < \alpha \leq \overline{\alpha}$. Pak podle věty o střední hodnotě platí $F(x^* + \alpha v^*) = F(x^*) + \alpha^2 (v^*)^T G(x^* + \tilde{\alpha} v^*) v^*/2$ (neboť $g(x^*) = 0$), kde $0 \leq \tilde{\alpha} \leq \alpha \leq \overline{\alpha}$, takže

$$F(x^* + \alpha v^*) = F(x^*) + \frac{\alpha^2}{2} (v^*)^T G(x^* + \tilde{\alpha} v^*) v^* \leq F(x^*) + \frac{\alpha^2}{4} \lambda^* (v^*)^T v^* < F(x^*),$$

což je ve sporu s definicí 3. □

Věta 2 *(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 $v \in R^n$. Jelikož $F \in \mathcal{C}^2$ na $\mathcal{B}(x^*, \varepsilon)$, existuje číslo $\overline{\alpha} > 0$ takové, že $x^* + \alpha v \in \mathcal{B}(x^*, \varepsilon)$ a $v^T G(x^* + \alpha v)v \geq \lambda^* v^T v/2$, pokud $0 \leq \alpha \leq \overline{\alpha}$ (plyne to ze spojitosti Hessovy matice $G(x^* + \alpha v)$). Nechť $0 < \alpha \leq \overline{\alpha}$. Pak podle věty o střední hodnotě platí $F(x^* + \alpha v) = F(x^*) + (\alpha^2/2)v^T G(x^* + \tilde{\alpha} v)v$ (neboť $g(x^*) = 0$), kde $0 \leq \tilde{\alpha} \leq \alpha \leq \overline{\alpha}$, takže

$$F(x^* + \alpha v) = F(x^*) + \frac{\alpha^2}{2} v^T G(x^* + \tilde{\alpha} v)v \geq F(x^*) + \frac{\alpha^2}{4} \lambda^* v^T v > F(x^*).$$

Jelikož $v \in R^n$ a $0 < \alpha \leq \overline{\alpha}$ lze vybrat libovolně, je bod x^* je ostrým lokálním minimem funkce F. Dále platí

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

takže $g(x^* + \alpha v) \neq 0$. Jelikož $v \in R^n$ a $0 < \alpha \leq \overline{\alpha}$ lze vybrat libovolně, je bod x^* izolovaným lokálním minimem funkce F. □

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

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

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

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

$$c_i(x) \leq 0, \quad \lambda_i \geq 0, \quad \lambda_i c_i(x) = 0, \quad i \in I.$$

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 $c_i(x) \leq 0$, $i \in I$, $c_i(x) = 0$, $i \in E$.

Podmínky pro vázané extrémy jsou studovány v druhé části práce.

1.3 Základní pojmy z teorie konvergence

Nyní se budeme zabývat vlastnostmi konvergentních posloupností. V duchu poznámky 6 budeme předpokládat, že platí (F4) a (F5) s $\mathcal{D} = R^n$.

Definice 4 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů. Jestliže pro libovolné $\varepsilon > 0$ existuje index $k \in N$ tak, že $x_i \in \mathcal{B}(x^*, \varepsilon)$ $\forall i \geq k$, 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 $k \geq 0$. Jestliže $\|u_i\|/\|v_i\|^k \to 0$, budeme psát $u_i = o(\|v_i\|^k)$. Jestliže existuje konstanta $C > 0$ taková, že $\|u_i\| \leq C\|v_i\|^k$ $\forall i \in N$, 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 $0 < \underline{c} \leq \overline{c} < \infty$ takové, že

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

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 3 *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 $e_i = x_i - x^*$, $i \in N$. 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 $0 \leq \tilde{\lambda} \leq 1$ a

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

Ze spojitosti druhých derivací plyne $\|G(x_i + \tilde{\lambda} d_i) - G(x_i)\| \leq \|G(x_i + \tilde{\lambda} d_i) - G^*\| + \|G(x_i) - G^*\| \to 0$, 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

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

Ze spojitosti druhých derivací plyne opět $\max_{0 \leq \lambda \leq 1} \|G(x_i + \lambda d_i) - G(x_i)\| \to 0$. 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 splněna podmínka (F6), dostaneme silnější odhady.

Věta 4 *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 podmínce (F6). Označme $e_i = x_i - x^*$, $i \in N$. 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 3. Vztahy typu $\|G(x_i + \lambda d_i) - G(x_i)\| \to 0$ se nahradí odhady typu $\|G(x_i + \lambda d_i) - G(x_i)\| \leq \overline{L} \|\lambda d_i\|$. □

Definice 5 *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 5 *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) R-lineárně právě tehdy, jestliže existují index $k \in N$ a čísla $M_k > 0$ a $0 < q < 1$, tak že*

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

$\forall i \geq k$.

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

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

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

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

$\forall i \geq k$, neboli

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

$\forall i \geq k$. Zvolme

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

Pak platí

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

□

Poznámka 11 Výraz použitý v definici 5 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 6 *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 7 *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, jestliže existuje index $k \in N$ a konstanta $0 < q < 1$ tak, že*

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

Definice 8 *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 6 *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 5). 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 9 *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 existuje číslo $m \in N$ takové, že

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

Poznámka 14 m-kroková Q-superlineární konvergence implikuje cyklickou m-krokově Q-superlineární konvergenci.

Ukážeme nyní, že cyklická m-krokově 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, $1 \le i \le m$, jsou nezáporná čísla. Pak platí*

$$\left(\prod_{i=1}^{m} \xi_i\right)^{\frac{1}{m}} \le \frac{1}{m} \sum_{i=1}^{m} \xi_i. \tag{8}$$

Důkaz Jelikož logaritmická funkce je konkávní (definice 1), platí

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

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

Věta 7 *Nechť $x_i \to x^*$ cyklicky m-krokově Q-superlineárně a nechť existuje konstanta $C > 0$ taková, že $\|e_{i+1}\| \le C\|e_i\|$ $\forall i \in N$. Pak $x_i \to x^*$ R-superlineárně.*

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

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

(používáme nerovnost (8) a tvrzení lemmatu 1, podle kterého platí $\frac{1}{k}\sum_{j=1}^{k} o(1) = o(1)$). Můžeme tedy psát

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

□

Poznámka 15 Předpoklady věty 7 jsou splněny pro cyklicky přerušovanou metodu sdružených gradientů s asymptoticky přesným výběrem délky kroku (věta 28).

Definice 10 *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*

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

Definice 11 *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*

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

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

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

kde směrový vektor $s_i \in R^n$ se určuje pomocí hodnot x_j, F_j, g_j, G_j, $1 \le j \le i$, 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 12 *Řekneme, že základní optimalizační metoda je globálně konvergentní, jestliže pro libovolný počáteční vektor $x_1 \in R^n$ platí*

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

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:

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

Nevýhody:

- Metoda největšího spádu vyžaduje přesný výběr délky kroku.
- 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:

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

Nevýhody:

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

2 Metody spádových směrů

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

$$\cos\theta_i = -\frac{s_i^T g_i}{\|s_i\|\|g_i\|} \tag{9}$$

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 spádových směrových vektorů.

Definice 13 *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 $C_1 > 1$ a $\overline{C} \geq 0$ jsou vhodné konstanty.

Poznámka 16 Definice dostatečné spádovosti směrových vektorů se může zdát dosti umělá. Nicméně je tato definice často velmi užitečná pro důkazy globální konvergence (věta 150). Použití podmínky dostatečné spádovosti se často nazývá principem omezeného znehodnocení. Poznamenejme, že z rekurentních nerovností použitých v (S1c) plyne

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

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 13).

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

Věta 8 *Nechť $B_i s_i = -g_i$ $\forall i \in N$, kde B_i, $i \in N$, je posloupnost symetrických pozitivně definitních matic. Pak platí*

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

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 17 Podle věty 37 platí stejný odhad, 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.

Další významnou součástí metod spádových směrů je výběr délky kroku, na který je třeba klást řadu omezení.

Definice 14 *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*

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

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

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

Poznámka 18 Při vyšetřování globální konvergence vystačíme s nerovnostmi $0 < \varepsilon_1 < \varepsilon_2 < 1$. Pro zaručení superlineární konvergence (věta 16) 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 19 Existují různé varianty zobecněné Wolfeho podmínky. Pokud $\varepsilon_3 = \infty$ (takže druhá nerovnost v (S3a) odpadne), dostaneme slabou Wolfeho podmínku. Pokud $\varepsilon_3 = \varepsilon_2$, dostaneme silnou Wolfeho podmínku. Někdy je třeba aby platilo $\varepsilon_3 = 0$ (věta 25). Položíme-li v silné Wolfeho podmínce $\varepsilon_2 = 0$, dostaneme přesný výběr délky kroku, kdy $s_i^T g_{i+1} = 0$. Obvykle se přesným výběrem délky kroku rozumí nalezení lokálního minima funkce $F(x_i + \alpha s_i)$ s nejmenší hodnotou parametru α.

Poznámka 20 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 $\underline{\alpha}\|g_i\|/\|s_i\| \leq \alpha_i^1 \leq \overline{\alpha}\|g_i\|/\|s_i\|$, 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 $i \in N$.

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

Definice 15 *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je metodou spádových směrů, jestliže směrové vektory $s_i \in R^n$, $i \in N$, splňují podmínku (S1a) a délky kroku $\alpha_i > 0$, $i \in N$, splňují podmínku (S2a) a některou z podmínek (S3). 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 21 Při realizaci metod sdružených gradientů odvozených z metody největšího spádu se používá silná Wolfeho podmínka (S3a) 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 (S3a) 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 (S3b) 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 $0 < \varepsilon_1 < 1/2$.

Poznámka 22 Metoda 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 $\varepsilon_0 = 1$.

Lemma 3 *(Konzistence) Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje podmínky (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

$$F(x_i + \alpha s_i) - F_i \leq \varepsilon_1 \alpha s_i^T g_i \qquad \forall\, 0 \leq \alpha \leq \tilde{\alpha}. \tag{10}$$

Jelikož $s_i^T g_i < 0$, platí $\tilde{\alpha}_i > 0$. Podle (F1) platí $F(x_i + \tilde{\alpha}_i s_i) \geq \underline{F}$, což podle (10) 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

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

$$s_i^T g(x_i + \tilde{\alpha}_i s_i) \geq \varepsilon_1 s_i^T g_i,$$

takže délka kroku $\alpha_i = \tilde{\alpha}_i$ splňuje podmínky (S2a), (S3a) s $\varepsilon_3 = \infty$ a (S3b) (neboť $0 < \varepsilon_1 < \varepsilon_2 < 1$). Tudíž slabá Wolfeho podmínka i Goldsteinova podmínka jsou konzistentní. Rovnost plyne ze spojitosti funkce $F : \mathcal{D} \to R$, nerovnost dokážeme sporem. Předpokládejme, že $s_i^T g(x_i + \tilde{\alpha}_i s_i) = \varepsilon s_i^T g_i$ pro nějaké číslo $\varepsilon > \varepsilon_1$. Jelikož nerovnost (10) implikuje $F(x_i + \tilde{\alpha}_i s_i) < F(x_i)$, platí $x_i + \tilde{\alpha}_i s_i \in \mathcal{D}_F(\overline{F}) \subset \mathcal{D}$ a 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 (F3) a (1) 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 $\lambda(\varepsilon_1 - \varepsilon) s_i^T g_i / \overline{G}\|s_i\| 2 < \delta$. Pak pro

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

dostaneme

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

což je spor, neboť $\tilde{\alpha}_i$ je největší číslo splňující podmínku (10).

(b) Nechť $s_i^T g(x_i + \tilde{\alpha}_i s_i) > 0$. 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 $s_i^T g(x_i + \alpha_i s_i) = 0$. Podle (10) tato délka kroku splňuje podmínku (S2a). Existuje tedy délka kroku $0 < \alpha_i \leq \tilde{\alpha}_i$ splňující podmínku (S2a), taková, že $s_i^T g(x_i + \tilde{\alpha}_i s_i) \leq 0$. Pak je podmínka (S3a) splněna pro libovolnou hodnotu $\varepsilon_3 \geq 0$ a 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. □

2.2 Globální konvergence

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

Lemma 4 *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje podmínky (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*

$$\alpha_i \geq -\frac{\varepsilon_4 s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{\varepsilon_4 \cos\theta_i \|g_i\|}{\overline{G}\|s_i\|} \tag{11}$$

a

$$F_{i+1} - F_i \leq -\frac{\varepsilon_1\varepsilon_4 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_4}{\overline{G}} \cos^2\theta_i \|g_i\|^2. \tag{12}$$

Důkaz Nerovnost (12) plyne bezprostředně z nerovnosti (11), 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 (11).

(a) Předpokládejme nejprve, že $0 < \alpha_i \leq \tilde{\alpha}_i$, kde $\tilde{\alpha}_i > 0$ je hodnota použitá v důkazu lemmatu 3. Pak $x + \alpha s_i \in \mathcal{D}\ \forall\, 0 < \alpha \leq \alpha_i$, takže lze použít předpoklad (F3). Platí-li (S3a), můžeme podle (5) psát

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

neboli

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

takže platí (11) s $\varepsilon_4 = 1 - \varepsilon_2 > 0$. Platí-li (S3b), můžeme s použitím odhadu (1) psát

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

neboli

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

takže platí (11) s $\varepsilon_4 = 1 - \varepsilon_2 > 0$.

(b) Nechť $\alpha_i \geq \tilde{\alpha}_i$. Protoze hodnota $\tilde{\alpha}_i > 0$ spluje podmínku (S3b) s $\varepsilon_2 = \varepsilon_1$ (důkaz lemmatu 3), platí

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

takže platí (11) s $\varepsilon_4 = 1 - \varepsilon_1 > 0$.

(c) Používáme-li Armijův výběr délky kroku, pak buď $\alpha_i = \alpha_i^1$, takže platí (11) s $\varepsilon_4 = \underline{\alpha}\overline{G} > 0$, nebo $\alpha_i \geq \tilde{\alpha}_i$, takže platí (11) s $\varepsilon_4 = 1 - \varepsilon_1 > 0$, nebo $\alpha_i \geq \underline{\beta}\tilde{\alpha}_i$ takže platí (11) s $\varepsilon_4 = \underline{\beta}(1-\varepsilon_1) > 0$. □

Poznámka 23 Jsou-li splněny předpoklady lemmatu 4, platí

$$\sum_{i=1}^{\infty} \frac{(s_i^T g_i)^2}{\|s_i\|^2} < \infty. \tag{13}$$

To plyne bezprostředně z (12), neboť

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

a výraz na levé straně je konečný (podrobnější argumentaci lze nalézt v důkazu věty 9).

Poznámka 24 V některých případech, například u metod sdružených gradientů, lze místo nerovnosti (S1b) dokázat nerovnost

$$-s_i^T g_i \geq \tilde{\varepsilon}_0 \|g_i\|^2, \tag{14}$$

kde $\tilde{\varepsilon}_0 > 0$. Pak podle (13) platí

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} < \infty. \tag{15}$$

Věta 9 *(Globální konvergence) Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak metoda spádových směrů, pro kterou platí*

$$\sum_{i=1}^{\infty} \cos^2\theta_i = \infty \tag{16}$$

je globálně konvergentní.

Důkaz Použijeme-li (12), 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 (12) je posloupnost F_i, $i \in N$ klesající a podle (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 25 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 $i \to \infty$. Metoda největšího spádu je tedy globálně konvergentní, přičemž $\|g_i\| \to 0$ pro $i \to \infty$.

Poznámka 26 Podle věty 8 a věty 9 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

$$\sum_{i=1}^{\infty} \frac{1}{\kappa_i} = \infty, \tag{17}$$

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 27 Podmínka 16 je splněna například tehdy, existuje-li konstanta $c \leq 1$ taková, že platí buď

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

nebo

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

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

Poznámka 28 Větu 9 lze použít ke globalizaci metod spádových směrů pomocí restartování. Restartováním rozumíme přerušení a nové nastartování iteračního procesu. Při novém nastartování iteračního procesu obvykle platí $s_i = -g_i$, takže je splněna podmínka (S1b). 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 9 a metoda spádových směrů je globálně konvergentní.

Ukážeme, že metoda dostatečně spádových směrů je globálně konvergentní

Věta 10 *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje podmínky (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 (F1), (S2a), definici čísla $\cos\theta_i$ a definici dostatečné spádovosti (nerovnost z posnámky 16), dostaneme (podobně jako v důkazu věty 9)

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

Nyní můžeme využít známou implikaci

$$\sum_{i=1}^{\infty} z_i < \infty \Rightarrow \prod_{i=1}^{\infty}(1 - z_i) > 0,$$

která platí pokud $0 < z_i < 1\ \forall i \in N$. Tato implikace ve spojení s předchozí nerovností dává

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

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

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

$\forall k \in N$, 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 $\cos\theta_k \geq 1/C_k \geq \underline{C}/C_1\ \forall k \in N$. To je však spor, neboť stejnoměrná metoda spádových směrů je podle poznámky 25 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 28.

Věta 11 *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 $F \in \mathcal{C}^1 : \mathcal{D} \to R$ podmínky (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

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

Pak

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

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 25 implikuje $\|g_i\| \to 0$. To je však ve sporu s předpokladem, že $\|g_i\| \geq \underline{\varepsilon}\ \forall i \in N$. □

Poznámka 29 Metody s proměnnou metrikou používají směrové vektory $s_i = -H_i g_i$, $i \in N$. Věta 65 tvrdí, že metody s proměnnou metrikou jsou globálné konvergentní, splňuje-li funkce $F \in \mathcal{C}^2 : R^n \to R$ podmínky (F1), (F4), (F5). Bez požadavku stejnoměrné konvexity (F5) tato věta neplatí. Věta 11 dává návod, jak lze metody s proměnnou metrikou korigovat tak, aby byly globálně konvergentní i tehdy, jsou-li splněny pouze podmínky (F1) a (F3). Poznamenejme, že číslo $\sigma > 0$ je obvykle velmi malé, například $\sigma = 10^{-12}$.

Je-li metoda spádových směru globálně konvergentní (definice 12), nemusí ještě platit $x_i \to x^*$. Splňuje-li funkce $F : \mathcal{D}_F \to R$ podmínku (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 2), pak platí $x_i \to x^*$.

Věta 12 *Nechť funkce $F \in \mathcal{C}^1 : \mathcal{D} \to R$ splňuje podmínky (F1)–(F3) 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 2, platí $x_i \to x^*$.*

Důkaz Protože bod $x^* \in \mathcal{D}$ vyhovuje předpokladům věty 2, 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 (2)–(4) a (3)–(6) platí

$$F - F^* \leq \frac{1}{2}\overline{G}\|x - x^*\|^2, \tag{18}$$

$$F - F^* \geq \frac{1}{2}\underline{G}\|x - x^*\|^2, \tag{19}$$

$$\|g\| \leq \overline{G}\|x - x^*\|, \tag{20}$$

$$\|g\| \geq \underline{G}\|x - x^*\|, \tag{21}$$

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

$$F_i - F^* < \frac{\underline{G}}{2}\varepsilon^2 \left(1 + \frac{\overline{G}^2}{2\varepsilon_0\varepsilon_1\underline{G}^2}\right)^{-2} \tag{22}$$

$\forall i \geq l$. 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 (18) a (21) 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 (S2a) dává

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

Použijeme-li tuto nerovnost spolu s (20), dostaneme

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \alpha_k\|s_k\| \leq \|x_k - x^*\| + \frac{\overline{G}}{2\varepsilon_0\varepsilon_1\underline{G}^2}\|g_k\| \leq \left(1 + \frac{\overline{G}^2}{2\varepsilon_0\varepsilon_1\underline{G}^2}\right)\|x_k - x^*\| \tag{23}$$

a podle (19) a (22) 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 (23) 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 30 Věta 12 vyžaduje stejnoměrnou spádovost směrových vektorů. Abychom dostali podobný výsledek v obecném případě (kdy platí pouze (S1a)), 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 požadovaná hodnota vychází větší).

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 2). Abychom nemuseli stále ověřovat zda pro daný index $i \in N$ již platí $x_i \in \mathcal{B}(x^*, \varepsilon)$, nahradíme předpoklady věty 2 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 13 *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 podmínkám (F4) a (F5). Pak jsou-li směrové vektory dostatečně spádové, jsou též stejnoměrně spádové (existuje číslo $\varepsilon_0 > 0$ takové že platí (S1b)).*

Důkaz Použijeme-li (S2a), definici čísla $\cos\theta_i$ a definici dostatečné spádovosti (nerovnost z poznámky 16), můžeme (podobně jako v důkazu věty 10) psát

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

$\forall i \in N$. Z druhé strany nerovnosti (18) a (21) implikují, že

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

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

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

Postupujeme-li stejným způsobem jako v důkazu věty 10, dokážeme že existuje číslo $\varepsilon_0 = \underline{C}/C_1 > 0$ takové, že $\cos\theta_i \geq \varepsilon_0 \ \forall i \in N$. □

Jsou li splněny předpoklady věty 13, 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 27.

Věta 14 *(Lineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů takovou, že*

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

kde $0 < \underline{c} \leq 1$. 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 podmínkám (F4) a (F5). Pak platí

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

kde $q = \sqrt{1 - \underline{c}\varepsilon_1\varepsilon_4\underline{G}/\overline{G}}$ (ε_4 je číslo z lemmatu 4).

Důkaz Podle (3) platí $F^* - F \geq g^T(x^* - x)$, což po úpravě dává $F - F^* \leq g^T(x - x^*) \leq \|g\|\|x - x^*\|$ a použijeme-li (21), dostaneme

$$\|g\|^2 \geq \underline{G}(F - F^*). \tag{24}$$

Podle (12) tedy platí

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

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

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

(používáme nerovnost mezi geometrickým a aritmetickým průměrem) (8), což s použitím (18) a (19) dává

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

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

Poznámka 31 Z monotonie posloupnosti F_i, $i \in N$, a z nerovností (18), (19) plyne, že $\|e_{i+1}\| = O(\|e_i\|)$. Z $\|e_{i+1}\| = O(\|e_i\|)$ plyne $\|d_i\| = \|e_{i+1}-e_i\| \le \|e_i\| + \|e_{i+1}\| = O(\|e_i\|)$.

Poznámka 32 Podle věty 14 a poznámky 31 platí

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

a také

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

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 16 *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 5 *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 podmínkám (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 3 platí

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

což s použitím (F4) a (F5) dává $g_i \sim e_i$, takže podle poznámky 31 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 3, 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 $\|d_i\|^2/d_i^T g_i \sim \|d_i\|^2/\|d_i\|\|g_i\| \sim 1$), takže

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

Podle věty 3 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 z (F4) a (F5) plyne $d_i^T G^* d_i \sim \|d_i\|^2$ 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 $(1 + o(1))^2 = 1 + o(1)$). □

Lemma 6 *Nechť B je symetrická pozitivně definitní (SPD) matice. Jestliže vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce*

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

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

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

Jestliže vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce

$$\frac{(u^T v)^2}{u^T u v^T v} \geq 1 - \varepsilon^2,$$

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

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

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

Důkaz (a) Nechť vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce $(u^T v)^2/(u^T u v^T v) \leq \varepsilon^2$. 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 $(u^T v)^2/(u^T u v^T v) \leq \varepsilon^2$ a z $\|v\| = 1$ pak plyne

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

a

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

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

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

Jelikož $V^TBV = M^TW^TBWM$, můžeme psát

$$\kappa(V^TBV) \leq \kappa(M^TM)\kappa(W^TBW).$$

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

$$\frac{x^TW^TBWx}{x^Tx} = \frac{x^TW^TBWx}{x^TW^TWx} = \frac{y^TBy}{y^Ty},$$

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

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

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

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

takže $\underline{\lambda}(M^TM) = 1 - |\alpha|$, $\overline{\lambda}(M^TM) = 1 + |\alpha|$ a

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

Můžeme tedy psát

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

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

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

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 $(x-1)/(x+1)$ je pro kladná x rostoucí).

(b) Nechť vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce $(u^Tv)^2/(u^Tuv^Tv) \geq 1 - \varepsilon^2$. Položme $w = BHv$, kde

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

Pak platí

$$u^T w = u^T B(B^{-1} - u(u^T Bu)^{-1}u^T)v = u^T v - u^T Bu(u^T Bu)^{-1}u^T v = 0,$$

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 platí

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

a

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

takže

$$\frac{(v^T w)^2}{w^T w v^T v} = 1 - \frac{(v^T u)^2}{u^T u v^T v} \leq \varepsilon^2$$

a použijeme-li (a), 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

$$\begin{aligned} w^T B^{-1} w &= w^T B^{-1} BHv = w^T B^{-1} v - w^T u(u^T Bu)^{-1} u^T v \\ &= w^T B^{-1} v = v^T HBB^{-1} v = v^T Hv \end{aligned}$$

a

$$v^T Hv = v^T B^{-1} v - (u^T v)^2 (u^T Bu)^{-1},$$

takže

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

□

Věta 15 *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 podmínkám (F4) a (F5). Pak platí*

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

Důkaz Podle věty 3 platí

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

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

takže s použitím (F4) a (F5) a toho, že $\|g_i\| \sim \|e_i\|$, dostaneme

$$e_i = (G^*)^{-1} g_i (1 + o(1))$$

a

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

Použijeme-li lemma 5 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 6 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

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

K libovolnému číslu $q, \hat{q} < q < 1$, tedy existuje index $k \in N$ tak, že

$$\frac{F_{i+1} - F^*}{F_i - F^*} \leq q^2$$

$\forall i \geq k$. Můžeme tedy postupovat stejně jako v důkazu věty 13, takže

$$\frac{F_i - F^*}{F_k - F^*} \leq q^{2(i-k)}$$

a

$$\frac{\|x_i - x^*\|}{\|x_k - x^*\|} \leq \sqrt{\frac{\overline{G}}{\underline{G}}} q^{i-k}$$

a podle věty 5 platí

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

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

Poznámka 33 Pro metodu největšího spádu je $\varepsilon_0 = 1$, takže

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

Poznámka 34 Používáme-li směrové vektory $s_i = -H_i g_i$, platí

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

kde $R_i = (G^*)^{1/2} H_i (G^*)^{1/2}$, neboť matice R_i mají stejná vlastní čísla jako matice $H_i^{-1/2} G^* H_i^{-1/2}$ a položíme-li $z_i = H_i^{-1/2} g_i$, můžeme stejně jako v důkazu věty 15 psát

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

a použitím lemmatu 6 dostaneme

$$\frac{F_{i+1}-F^*}{F_i - F^*} \le \left(\frac{\kappa(R_i)-1}{\kappa(R_i)+1}\right)^2 (1+o(1)).$$

Poznámka 35 Asymptoticky přesný výběr délky kroku dostaneme, vybíráme-li délku kroku pomocí kvadratické nebo kubické interpolace (věta 19).

Nyní se budeme zabývat superlineární konvergencí metod spádových směrů. Budeme používat označení

$$\omega_i = \frac{B_i s_i + g_i}{\|g_i\|}, \qquad \vartheta_i = \frac{(B_i - G_i)s_i}{\|s_i\|} \tag{25}$$

a budeme předpokládat, že konstanty ε_1 a ε_2 v podmínkách (S2)–(S3) vyhovují nerovnostem uvedeným v poznámce 18, tedy že platí $0 < \varepsilon_1 < 1/2$ a v případě podmínky (S3b) též $1/2 < \varepsilon_2 < 1$.

Věta 16 *(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 pozitivne definitní). Nechť $\omega_i \to 0$, $\vartheta_i \to 0$, neboli*

$$\lim_{i\to\infty} \frac{\|B_i s_i + g_i\|}{\|g_i\|} = 0, \qquad \lim_{i\to\infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0, \tag{26}$$

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

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

$$\underline{G}\|s_i\| \le \|g_i\| \le \overline{G}\|s_i\|$$

$\forall i \ge k_1$. Použijeme-li označení (25), můžeme psát

$$G_i s_i = (B_i s_i + g_i) - (B_i - G_i)s_i - g_i = \omega_i \|g_i\| - \vartheta_i \|s_i\| - g_i,$$

takže

$$(\overline{\lambda}(G_i) + \|\vartheta_i\|)\|s_i\| \ge (1 - \|\omega_i\|)\|g_i\|,$$

$$(\underline{\lambda}(G_i) - \|\vartheta_i\|)\|s_i\| \le (1 + \|\omega_i\|)\|g_i\|$$

a jelikož $\|\vartheta_i\| \to 0$ a $\|\omega_i\| \to 0$ (podle (26) a $\overline{\lambda}(G_i) \to \overline{\lambda}(G^*) < \overline{G}$, $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, existuje index $k_1 \in N$ takový, že

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

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

$\forall i \geq k_1$, což implikuje dokazovanou nerovnost.

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

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

a jelikož $\|\omega_i\| \to 0$ a $\|\vartheta_i\| \to 0$ (podle (26) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, existuje index $k_2 \geq k_1$ takový, že $-s_i^T g_i \geq (\underline{G}/\overline{G})\|s_i\|\|g_i\|$ $\forall i \geq k_2$.

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

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

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 $i \geq k_2$, takže podle (26) a (a) platí $|\eta_i| \to 0$. Nyní použijeme větu 3, 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 (S2a) a (S3a), (S3b) (s $\alpha_i = 1$) platí $\forall i \geq k$.

(d) Superlineární konvergence. Použijeme-li větu 3, podle které

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

a předchozí výsledky, dostaneme $x_{i+1} = x_i + s_i$ $\forall i \geq k$ a $\|s_i\| \sim \|g_i\| \to 0$. Můžeme tedy psát

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

takže podle (26) a (a) platí

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

□

Nyní ukážeme, že metodu spádových směrů, pro kterou jsou veličiny ω_i a ϑ_i dostatečně malé, lze považovat za nepřesnou Newtonovu metodu, pro kterou platí $\|G_is_i+g_i\|/\|g_i\|\le\overline{\omega}'$, kde $\overline{\omega}'<1$. Abychom zjednodušili argumentaci, budeme ve smyslu poznámky 6 předpokládat, že funkce $F\in\mathcal{C}^2:\mathcal{D}\to R$ vyhovuje podmínkám (F4) a (F5).

Věta 17 *Nechť $x_i\in R^n$, $i\in N$, je posloupnost generovaná metodou spádových směrů 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 podmínkám (F4) a (F5). Nechť $\|\omega_i\|\le\overline{\omega}$ a $\|\vartheta_i\|\le\overline{\vartheta}$ $\forall i\in N$, kde $\overline{\omega}<1$ a $\overline{\vartheta}<(1-\overline{\omega})\,\underline{G}/2$. Pak pro $i\in N$ platí*

$$\frac{\|G_is_i+g_i\|}{\|g_i\|}\le\overline{\omega}'<1,\qquad \overline{\omega}'=\frac{\overline{\omega}+\overline{\vartheta}/\underline{G}}{1-\overline{\vartheta}/\underline{G}}.$$

Důkaz Jelikož $B_is_i=(B_is_i+g_i)-g_i$, můžeme psát $\|B_is_i\|\le(1+\|\omega_i\|)\|g_i\|$, takže dostaneme

$$(1+\|\omega_i\|)\|g_i\|\ge\|B_is_i\|\ge\|G_is_i\|-\|(B_i-G_i)s_i\|\ge(\underline{G}-\|\vartheta_i\|)\|s_i\|, \tag{27}$$

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

$$\|G_is_i+g_i\|\le\|B_is_i+g_i\|+\|(B_i-G_i)s_i\|$$

dostaneme $\|\omega_i'\|\le\|\omega_i\|+\|\vartheta_i\|\|s_i\|/\|g_i\|$, což spolu s (27) dává

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

neboť $\|\omega_i\|\le\overline{\omega}<1$, $\|\vartheta_i\|\le\overline{\vartheta}$ a funkce $(\overline{\omega}+t)/(1-t)$ je pro $0\le t<1$ rostoucí. Jelikož $\overline{\vartheta}/\underline{G}<(1-\overline{\omega})/2$, platí $\overline{\omega}'<(\overline{\omega}+(1-\overline{\omega})/2)/(1-(1-\overline{\omega})/2)=1$. □

Poznámka 36 Dosavadní úvahy nám dovolují určit asymptotickou rychlost konvergence nepřesné Newtonovy metody. Tak jako v části (d) důkazu věty 16 můžeme psát

$$\frac{\|g_{i+1}\|}{\|g_i\|}\le\frac{\|G_is_i+g_i\|}{\|g_i\|}+o(\|s_i\|)/\|g_i\|=\|\omega_i'\|+o(\|s_i\|)/\|g_i\|,$$

takže pokud $\|\omega_i'\|\le\overline{\omega}'$, platí

$$\lim_{i\to\infty}\frac{\|g_{i+1}\|}{\|g_i\|}\le\overline{\omega}'.$$

Určujeme-li směrový vektor dostatečně přesně (takže číslo $\overline{\omega}'$ je malé), konverguje nepřesná Newtonova metoda velmi rychle. Pokud $\vartheta_i=0$ (takže $\omega_i'=\omega_i$), stačí dostatečně přesně řešit soustavu lineárních rovnic $G_is_i+g_i=0$, což však může klást velké nároky na výběr vhodné metody (je-li matice G_i špatně podmíněná, potřebují iterační metody k nalezení tohoto řešení velký počet iteračních kroků). Pokud $\vartheta_i\ne0$, je situace ještě složitější. Chceme-li v tomto případě dosáhnout asymptotickou rychlost konvergence $\overline{\omega}'$, je třeba, aby platilo $\overline{\omega}=\lambda\,\overline{\omega}'$ a $\overline{\vartheta}/\underline{G}=(1-\lambda)\,\overline{\omega}'/(1+\overline{\omega}')$, kde $0<\lambda<1$.

Poznámka 37 Věta 16 udává postačující podmínky pro superlineární konvergenci metody spádových směrů. Tyto podmínky však nejsou nutné. Platí-li například $B_i s_i + g_i = 0$, $B_i = \beta_i G_i$ a $x_{i+1} = x_i + \alpha_i s_i$, kde $\alpha_i = \beta_i \; \forall i \in N$, konverguje tato metoda superlineárně, ale není splněna druhá z podmínek (26). Potíž je v tom, že čísla β_i nemusí být explicitně známa, takže není jasné jak volit α_i. Dá se ukázat (důkaz není obtížný a podobá se části (d) důkazu věty 16), že metoda spádových směrů je superlineárně konvergentní právě tehdy pokud

$$\lim_{i\to\infty} \frac{\|d_i - G_i s_i\|}{\|s_i\|} = 0,$$

kde $d_i = x_{i+1} - x_i = \alpha_i s_i$. Význam věty 16 spočívá v tom, že udává prakticky ověřitelé postačující podmínky pro superlineární konvergenci.

2.4 Výběr délky kroku

Nyní se budeme zabývat implementací metod spádových směrů. Popíšeme nejprve algoritmus pro výběr délky kroku, 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 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 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 $\overline{\alpha} := \alpha$ a přejdeme na krok 4. V opačném případě položíme $\underline{\alpha} := \alpha$ a přejdeme na krok 4.

Poznámka 38 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 18 *Jsou-li splněny podmínky (F1) a (F3) najde algoritmus 1 délku kroku vyhovující podmínkám (S2a) a (S3b) 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 (F1) (podobně jako v důkazu lemmatu 3) plyne

$$\underline{\alpha}_j \le \frac{\underline{F} - \varphi(0)}{\varepsilon_1\varphi'(0)}. \tag{28}$$

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 $\varphi'(\underline{\alpha}_j) < \varepsilon_2\varphi'(0)$. Nechť 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 3 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 (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

$$\overline{\alpha}_j - \underline{\alpha}_j > \frac{\varepsilon_1 - \varepsilon_2}{\overline{G}\|s\|^2}\varphi'(0). \tag{29}$$

Ve druhé fázi algoritmu, upravujeme interval tak, že $\overline{\alpha}_{j+1} - \underline{\alpha}_{j+1} \leq \max(1-\beta_1, \beta_2)(\overline{\alpha}_j - \underline{\alpha}_j)$. 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 podmínkám (F4) a (F5), můžeme předchozí tvrzení podstatně zesílit (budeme to potřebovat pro důkaz asymptotické přesnosti výběru délky kroku).

Lemma 7 *Uvažujme algoritmus 1 s počáteční délkou kroku $\delta_1\|g\|/\|s\| \leq \alpha_1 \leq \delta_2\|g\|/\|s\|$, 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 podmínkám (F4) a (F5). Pak existují konstanty c_1 a c_2 nezávislé na vnějším indexu takové, že*

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

$\forall j \in N$. 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 18.
(a) V první fázi algoritmu využijeme toho, že vzhledem k (F5) můžeme $\underline{F}$ nahradit F^* v (28), což s použitím nerovnosti (24) a nerovnosti (S1b) (zapsané ve tvaru $-\varphi'(0) \geq \varepsilon_0\|s\|\|g\|$) dává

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

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

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

Můžeme tedy položit $c_1 = \delta_1 \min(1, \gamma_1 - 1)$ a $c_2 = \gamma_2/(\varepsilon_0\varepsilon_1\underline{G})$. Jelikož $\overline{\alpha}_j \geq \gamma_1^{j-1}\alpha_1 \geq \gamma_1^{j-1}\delta_1\|g\|/\|s\|$ a $\gamma_1 > 1$, existuje index $k_1 \in N$ takový, že $\overline{\alpha}_j \geq c_2\|g\|/\|s\|$ $\forall j \geq k_1$, takže první fáze skončí po nejvýše k_1 krocích.
(b) Ve druhé fázi algoritmu se již $\overline{\alpha}_j$ nezvětšuje, takže s použitím (29) a (a) můžeme psát

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

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 $j \geq j_2$ platí

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

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

Poznámka 39 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):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{2(1 - A)}. \tag{30}$$

Kvadratická interpolace (dvě derivace):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{1 - B}. \tag{31}$$

Kubická interpolace:

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{D + \sqrt{D^2 - 3C}}, \tag{32}$$

kde

$$C = (B - 1) - 2(A - 1),$$

$$D = (B - 1) - 3(A - 1).$$

Věta 19 *Nechť jsou splněny předpoklady lemmatu 7. Pak je-li délka kroku v algoritmu 1 spočtena podle (30) nebo (31) nebo (32), 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 z (F4) a (F5) plyne $s_i^T G^* s_i \sim \|s_i\|^2$ a z (S1b) plyne $s_i^T g_i \sim \|s_i\|\|g_i\|$, platí $\alpha_i^* \sim \|g_i\|/\|s_i\|$, kde

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

Podle lemmatu 7 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 z $\underline{\alpha}_i < \overline{\alpha}_i = O(\|g_i\|/\|s_i\|)$ a z $\|g_i\| \sim \|e_i\|$ ((F4), (F5) a věta 3) 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 3 dostaneme úpravou výrazů A, B uvedených v poznámce 39

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

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 (32) (důkaz pro (30) a (31) 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 (32), 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 40 Je-li kromě podmínek (F4) a (F5) splněna i podmínka (F6), můžeme místo věty 3 použít větu 4 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 41 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 17 *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 zobecněnou 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 42 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 18 *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 (S3). 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 43 *Pro každou nemonotonní metodou spádových směrů platí rovnice (11) z lemmatu 4, neboť $\overline{F}_i \geq F_i$ $\forall i \in N$. Rovnici (12) je třeba nahradit vztahem*

$$F_{i+1} - \overline{F}_i \leq -\frac{\varepsilon_1 \varepsilon_4 (s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1 \varepsilon_4}{\overline{G}} \cos^2 \theta_i \|g_i\|^2. \tag{33}$$

Nejprve vyšetříme jednoduchou nemonotonní metodu spádových směrů, pro kterou platí

$$\overline{F}_i = \max\{F_j : i - \min(m, i) + 1 \leq j \leq i\}, \tag{34}$$

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

Věta 20 *(Globální konvergence metody (34)) Nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak nemonotonní metoda stejnoměrně spádových směrů definovaná vztahem (34) je globálně konvergentní.*

Důkaz (a) Z (S2b) a (34) vyplývá, že posloupnost $\overline{F}_i$, $i \in N$, je nerostoucí. Platí totiž

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

Použijeme-li navíc (S1b) a (33) vidíme, že pro libovolné indexy $k \in N$ a $1 \leq j \leq m$ platí

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

takže

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

kde

$$\|g_{j(k)}\| = \min_{1\le j\le m} \|g_{mk+j}\|.$$

(b) Podle (a) platí

$$\frac{\varepsilon_0\varepsilon_1\varepsilon_4}{\overline{G}} \sum_{k=1}^{\infty} \|g_{j(k)}\|^2 \le \sum_{k=1}^{\infty} (\overline{F}_{mk} - \overline{F}_{m(k+1)}) = \overline{F}_m - \lim_{k\to\infty} \overline{F}_{m(k+1)} \le \overline{F}_1 - \underline{F} < \infty,$$

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

$$\overline{n}_{i+1} = \lambda \overline{n}_i + 1, \qquad \overline{F}_{i+1} = \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \tag{35}$$

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

Poznámka 44 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 $1 \le \overline{n}_i \le i$ a

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

pro $i \in N$, neboť z (S2b) a (35) plyne

$$F_{i+1} = \frac{\lambda \overline{n}_i + 1}{\overline{n}_{i+1}} F_{i+1} \le \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \le \frac{\lambda \overline{n}_i + 1}{\overline{n}_{i+1}} \overline{F}_i = \overline{F}_i$$

a funkce

$$\overline{F}_i(\lambda) = \frac{\lambda \overline{n}_{i-1} \overline{F}_{i-1} + F_i}{\lambda \overline{n}_{i-1} + 1}$$

je pro $F_i \le \overline{F}_{i-1}$ neklesající.

Věta 21 *(Globální konvergence metody (35)) Nechť funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak nemonotonní metoda spádových směrů definovaná rekurentními vztahy (35) je globálně konvergentní, pokud*

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

Důkaz Podle (33) platí

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

což spolu s (35) dává

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

Jelikož podle (F1) a poznámky 44 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 44 platí $\overline{n}_{i+1} \leq i + 1 \leq 2i$. Z poslední nerovnosti dostaneme dokazované tvrzení postupem uvedeným v důkazu věty 9. □

Poznámka 45 Podmínka použitá ve větě 21 je mnohem silnější než podmínka vystupující ve větě 9. 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 46 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 47 jsou-li splněny podmínky pro superlineární konvergenci (26) 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 16 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.

3 Metody sdružených gradientů

3.1 Základní vlastnosti metod sdružených gradientů

Definice 19 *Řekneme, že metoda spádových směrů (definice 15) je metodou sdružených gradientů, jestliže*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{36}$$

kde parametr β_i se vybírá tak, aby směrové vektory s_i, $1 \leq i \leq n$, byly sdružené (nebo G–ortogonální, podmínka (40)), aplikujeme-li tuto metodu na ryze konvexní kvadratickou funkci

$$Q(x) = \frac{1}{2}(x - x^*)^T G(x - x^*)$$

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 Q 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 (36) dostaneme rovnici $\beta_i y_i^T s_i - y_i^T g_{i+1} = 0$ neboli

$$\beta_i = \frac{y_i^T g_{i+1}}{y_i^T s_i}. \tag{37}$$

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 Q po konečném počtu kroků (je-li výběr délky kroku přesný).

Věta 22 *(Kvadratické ukončení) Nechť $Q : R^n \to R$ je ryze konvexní kvadratická funkce a x_i, $i \in N$, je posloupnost generovaná metodou sdružených gradientů (36)–(37) s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0 \ \forall i \in N$). Pak existuje index $m \leq n$ tak, že $g_{m+1} = 0$ a $x_{m+1} = x^*$.*

Důkaz Předpokládejme, že $g_i \neq 0 \ \forall 1 \leq i \leq n$ (není-li tato podmínka splněna, platí $g_{m+1} = 0$ a $x_{m+1} = x^*$ pro nějaký index $m < n$). Dokážeme indukcí, že $s_i \neq 0$, $\alpha_i \neq 0$, $1 \leq i \leq n$, a že pro $1 \leq j < i \leq n+1$ platí

$$s_j^T g_i = 0, \tag{38}$$

$$g_j^T g_i = 0, \tag{39}$$

$$s_j^T G s_i = 0, \tag{40}$$

$$s_j^T y_i = y_j^T s_i = 0. \tag{41}$$

Rovnosti (40) a (41) jsou ekvivalentní, neboť pro kvadratickou funkci $Q(x)$ platí $y_i = g_{i+1} - g_i = G(x_{i+1} - x_i) = Gd_i = \alpha_i G s_i$ a $\alpha_i \neq 0$ podle indukčního předpokladu. Z (39) plyne, že nenulové gradienty g_i, $1 \leq i \leq n$, jsou vzájemně ortogonální, tudíž lineárně nezávislé, takže nutně $g_{n+1} = 0$ a $x_{n+1} = x^*$. Indukční předpoklad je triviálně splněn v prvním iteračním kroku, neboť $s_1^T g_1 = -g_1^T g_1 < 0$ takže $s_1 \neq 0$ a $\alpha_1 \neq 0$ a dále není co dokazovat. Indukční krok:
(a) Nechť $i \leq n$. Podle indukčních předpokladů (38) a (41) platí:

$$s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = 0,$$

$1 \leq j < i$. Z přesného výběru délky kroku plyne $s_i^T g_{i+1} = 0$. Je tedy $s_j^T g_{i+1} = 0$, $1 \leq j \leq i$.
(b) Nechť $i \leq n$. Z (36) plyne

$$\begin{aligned} g_1 &= -s_1, \\ g_j &= -s_j + \beta_{j-1} s_{j-1}, \qquad 1 < j \leq i, \end{aligned}$$

takže podle (a) platí

$$\begin{aligned} g_1^T g_{i+1} &= -s_1^T g_{i+1} = 0, \\ g_j^T g_{i+1} &= -s_j^T g_{i+1} + \beta_{j-1} s_{j-1}^T g_{i+1} = 0, \qquad 1 < j \le i. \end{aligned}$$

(c) Nechť $i < n$. Z (37) a (a) dostaneme

$$s_{i+1}^T g_{i+1} = -g_{i+1}^T g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T s_i} s_i^T g_{i+1} = -g_{i+1}^T g_{i+1} < 0,$$

takže $s_{i+1} \neq 0$ a $\alpha_{i+1} \neq 0$. Z (37) a (b) dostaneme

$$y_j^T s_{i+1} = -y_j^T g_{i+1} + \beta_i y_j^T s_i = -y_j^T g_{i+1} = -(g_{j+1} - g_j)^T g_{i+1} = 0,$$

$1 \le j < i$ neboť podle předpokladu (41) platí $y_j^T s_i = 0$, $1 \le j < i$. Dále podle (37) 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 $s_j^T G s_{i+1} = 0$, $1 \le j \le i$. □

Poznámka 48 Z rovností $s_j^T g_{i+1} = 0$, $1 \le j \le i$, vyplývá, že bod x_{i+1} realizuje minimum ryze konvexní kvadratické funkce Q na podprostoru generovaném vektory s_j, $1 \le j \le i$.

Poznámka 49 Používáme-li přesný výběr délky kroku, můžeme podle (36) psát

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i = -g_i^T s_i = g_i^T g_i - \beta_{i-1} g_i^T s_{i-1} = g_i^T g_i.$$

Je-li navíc minimalizovaná funkce kvadratická, platí (39), 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 (37) můžeme použít tři různé jmenovatele a dva různé čitatele, aniž bychom porušili platnost věty 22. Dostaneme tak šest základních metod sdružených gradientů.

$$\beta_i^{HS} = \frac{y_i^T g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PR} = \frac{y_i^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{LS} = \frac{y_i^T g_{i+1}}{|g_i^T s_i|} \tag{42}$$

(HS – Hestenes a Stiefel, PR – Polak a Ribiére, LS – Liu a Storey),

$$\beta_i^{DY} = \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \quad \beta_i^{FR} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{CD} = \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|} \tag{43}$$

(DY – Dai a Yuan, FR – Fletcher a Reeves, CD – conjugate descent). Tyto metody můžeme rozdělit do dvou 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 patřící do téže skupiny se svými vlastnostmi příliš neliší.

Poznámka 50 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

$$s_1 = -H g_1 \quad \text{a} \quad s_{i+1} = -H g_{i+1} + \beta_i s_i \quad \text{pro} \quad i \in N,$$

kde

$$\beta_i^{PHS} = \frac{y_i^T H g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PPR} = \frac{y_i^T H g_{i+1}}{g_i^T H g_i}, \quad \beta_i^{PLS} = \frac{y_i^T H g_{i+1}}{|g_i^T s_i|},$$

nebo

$$\beta_i^{PDY} = \frac{g_{i+1}^T H g_{i+1}}{y_i^T s_i}, \quad \beta_i^{PFR} = \frac{g_{i+1}^T H g_{i+1}}{g_i^T H g_i}, \quad \beta_i^{PCD} = \frac{g_{i+1}^T H g_{i+1}}{|g_i^T s_i|}.$$

Metoda, která používá tyto vzorce se nazývá předpodmíněnou metodou sdružených gradientů. Pro tuto metodu platí všechny věty, které jsme zatím dokázali (splňuje-li funkce $F(x)$ podmínky (F1)–(F3), případně (F5)–(F6), splňuje tyto podmínky 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 (38), (40), (41) zůstanou beze změny, ale místo (39) platí

$$g_j^T H g_i = 0,$$

$1 \leq j < i \leq n+1$.

3.2 Globální konvergence

Jak již bylo poznamenáno (poznámka 49), jsou metody (43) 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 23 *(Globální konvergence metody DY). Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak metoda sdružených gradientů (36) s výběrem délky kroku splňujícím slabou Wolfeho podmínku (S2a) a (S3a), kde $0 < \varepsilon_1 < \varepsilon_2 < 1$ a $\varepsilon_3 = \infty$, je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \frac{g_{i+1}^T g_{i+1}}{y_i^T s_i}, \tag{44}$$

kde $-(1-\varepsilon_2)/(1+\varepsilon_2) \leq \lambda_i \leq 1 \; \forall i \in N$.

Důkaz (a) Dokážeme nejprve, že

$$|\beta_i| \leq \frac{g_{i+1}^T s_{i+1}}{g_i^T s_i} \tag{45}$$

$\forall i \in N$. Použijeme-li (36) 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 (S3b) 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 (45) 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 (45) $g_{i+1}^T s_{i+1} \leq |\beta_i| g_i^T s_i < 0$, pokud $\beta_i \neq 0$. Jestliže $\beta_i = 0$, dostaneme podle (36) $g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1} < 0$.
(b) Zapíšeme-li (36) ve tvaru $s_{i+1} + g_{i+1} = \beta_i s_i$, dostaneme umocněním, převedením dvou členů na pravou stranu a použitím nerovnosti (45) vztah

$$\|s_{i+1}\|^2 = \beta_i^2 \|s_i\|^2 - 2g_{i+1}^T s_{i+1} - \|g_{i+1}\|^2 \leq \left(\frac{g_{i+1}^T s_{i+1}}{g_i^T s_i} \right)^2 \|s_i\|^2 - 2g_{i+1}^T s_{i+1} - \|g_{i+1}\|^2,$$

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 (13) uvedenou v poznámce 23. □

Nyní se budeme zabývat důkazem globální konvergence metody FR. Opět budeme vyšetřovat poněkud širší třídu metod sdružených gradientů.

Věta 24 *(Globální konvergence metody FR). Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak metoda sdružených gradientů (36) s výběrem délky kroku splňujícím silnou Wolfeho podmínku (S2a) a (S3a), kde $0 < \varepsilon_1 < \varepsilon_2 < 1/2$ a $\varepsilon_3 = \varepsilon_2$, je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \frac{\|g_{i+1}\|^2}{\|g_i\|^2}, \tag{46}$$

kde $|\lambda_i| \leq 1$ $\forall i \in N$.

Důkaz (a) (Al-Baali) Dokážeme indukcí nerovnost

$$-1 - \frac{\varepsilon_2}{1-\varepsilon_2} \leq \frac{g_i^T s_i}{\|g_i\|^2} \leq -1 + \frac{\varepsilon_2}{1-\varepsilon_2} < 0. \tag{47}$$

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 (36) 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 (levá 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 (14) s $\tilde{\varepsilon}_0 = (1-2\varepsilon_2)/(1-\varepsilon_2)$, což podle poznámky 24 implikuje nerovnost (15).

(b) Použijeme-li levou část podmínky (S3a) a levou část nerovnosti (47), 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 (36), 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 $|\lambda_i| \leq 1$. dosazováním dostaneme 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 (15) uvedenou v poznámce 24. □

Poznámka 51 Věta 24 vyžaduje silnější předpoklady než věta 23. Je třeba, aby byla splněna silná Wolfeho podmínka a aby navíc platilo $\varepsilon_2 < 1/2$. Samotnou nerovnost (47) však můžeme zobecnit tak, že ji lze použít i za poněkud slabších předpokladů. Jestliže $|\lambda_i| \leq \overline{\varepsilon}_2/\varepsilon_2$, kde $0 < \varepsilon_2 < \overline{\varepsilon}_2 < 1/2$, můžeme psát

$$-1 - \frac{\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2} \leq \frac{g_i^T s_i}{\|g_i\|^2} \leq -1 + \frac{\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2} < 0. \tag{48}$$

Pokud $\varepsilon_2 \approx 1/10$ (což je doporučená hodnota) a $\overline{\varepsilon}_2 \approx 1/2$, platí tato nerovnost i pro $|\lambda_i| \approx 5$ (λ_i je koeficient v (46)).

Poznámka 52 V předpokladech věty 24 můžeme silnou Wolfeho podmínku (S2a) a (S3a) nahradit zobecněnou Wolfeho podmínkou (S2a) a (S3e), kde $0 < \varepsilon_1 < \varepsilon_2 < 1$ a $0 < \varepsilon_3 < 1/2$. Pro metodu FR to nemá žádný praktický význam. Zobecněná Wolfeho podmínka je však podstatná pro důkaz globální konvergence metody CD.

Věta 25 *(Globální konvergence metody CD). Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak metoda sdružených gradientů (36) s výběrem délky kroku splňujícím zobecněnou Wolfeho podmínku (S2a) a (S3a), kde $0 < \varepsilon_1 < \varepsilon_2 < 1$ a $\varepsilon_3 = 0$, je globálně konvergentní, pokud*

$$\beta_i = \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|}, \tag{49}$$

kde $0 \le \lambda_i \le 1$ $\forall i \in N$.

Důkaz (a) Použijeme-li (36), (49) a (S3a) s $\varepsilon_3 = 0$ (takže $g_{i+1}^T s_i \le 0$), dostaneme

$$-g_{i+1}^T s_{i+1} = g_{i+1}^T g_{i+1} - \lambda_i \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|} g_{i+1}^T s_i \ge g_{i+1}^T g_{i+1},$$

takže směrové vektory s_i, $i \in N$, jsou spádové a platí (14) s $\tilde{\varepsilon}_0 = 1$, což podle poznámky 24 implikuje nerovnost (15).

(b) Použijeme-li vztahy (37), (49), podmínku $|\lambda_i| \le 1$ a nerovnost $g_{i+1}^T s_i \le 0$, dostaneme

$$\begin{aligned}
\|s_{i+1}\|^2 &= \left(-g_{i+1} + \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right)^T \left(-g_{i+1} + \lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right) \\
&= \|g_{i+1}\|^2 - 2\lambda_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} g_{i+1}^T s_i + \lambda_i^2 \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2 \\
&\le \|g_{i+1}\|^2 + 2\varepsilon_2 \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2,
\end{aligned}$$

neboli

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \le \frac{1 + 2\varepsilon_2}{\|g_{i+1}\|^2} + \frac{\|s_i\|^2}{\|g_i\|^4}.$$

Předpokládejme, že neplatí

$$\liminf_{i \to \infty} \|g_i\| = 0.$$

Pak existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \ge \underline{\varepsilon}$ $\forall i \in N$, takže z předchozí nerovnosti plyne

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \le \frac{1 + 2\varepsilon_2}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \le \frac{1 + 2\varepsilon_2}{\underline{\varepsilon}^2}(i+1)$$

(předpokládáme bez újmy na obecnosti, že $\underline{\varepsilon}^2 \|s_1\|^2 / \|g_1\|^4 \le 1 + 2\varepsilon_2$). Můžeme tedy psát

$$\sum_{i=1}^{\infty} \frac{\|g_i\|^4}{\|s_i\|^2} \ge \frac{\underline{\varepsilon}^2}{1 + 2\varepsilon_2} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

což je ve sporu s nerovností (15) uvedenou v poznámce 24. □

Poznámka 53 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.

Poznámka 54 Jak již bylo zmíněno v poznámce 49, dávají metody (42) lepší praktické výsledky než metody (43). Vlastnosti metod (42) lze zlepšit tím, že vyloučíme záporné hodnoty, takže

$$\beta_i^{HS+} = \max(0, \beta_i^{HS}), \quad \beta_i^{PR+} = \max(0, \beta_i^{PR}), \quad \beta_i^{LS+} = \max(0, \beta_i^{LS}). \tag{50}$$

Metody (42) lze též kombinovat s metodami (43). Tyto kombinované metody používají vztahy

$$\begin{aligned} \beta_i^{HSC} &= \max(0, \min(\beta_i^{HS}, \beta_i^{DY})), \\ \beta_i^{PRC} &= \max(0, \min(\beta_i^{PR}, \beta_i^{FR})), \\ \beta_i^{LSC} &= \max(0, \min(\beta_i^{LS}, \beta_i^{CD})). \end{aligned} \tag{51}$$

Kladné hodnoty se používají proto, aby se předešlo možnému zacyklení. Z předchozích vět je zřejmé, že kombinované metody HSC, PRC, LSC jsou globálně konvergentní za stejných podmínek jako metody DY, FR a CD. Navíc jsou efektivní pro praktické výpočty.

Poznámka 55 Globální konvergenci metod HS, PR a LS lze zajistit pomocí restartování. Pokud neplatí $-g_{i+1}^T s_{i+1} \geq \varepsilon_0 \|g_{i+1}\| \|s_{i+1}\|$, kde $\varepsilon_0 > 0$, pokládáme $s_{i+1} = -g_{i+1}$ (což odpovídá hodnotě $\beta_i = 0$). Podle poznámky 28 je restartovaná metoda globálně konvergentní. Zvolíme li číslo ε_0 dostatečně malé, dochází k restartování pouze sporadicky a takto upravené metody HS, PR a LS jsou velmi efektivní.

3.3 Přerušované metody sdružených gradientů

Poznámka 56 Metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce po nejvýše n krocích (věta 22). 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 22, je třeba iterační proces přerušit ($s_{n+1} = -g_{n+1}$). V dalších úvahách se budeme zabývat cyklicky přerušovanými metodami sdružených gradientů.

Definice 20 *Řekneme, že základní optimalizační metoda je cyklicky přerušovanou metodou sdružených gradientů, jestliže $s_i = -g_i$ pro $i \in M$ a jestliže platí některý ze vzorců (42)–(43) pro $i \notin M$, kde $M = \{l \in N : l = nk + 1, k \in N\}$.*

Poznámka 57 Definice 20 je jistou idealizací. Ve skutečnosti může dojít k přerušení iteračního procesu dříve než po n krocích. V tomto případě lze množinu M posunout. Pro naše úvahy je podstatné, že k přerušení dojde nejpozději po n krocích.

Nejprve ukážeme, že cyklicky přerušovaná metoda sdružených gradientů, kde parametr β_i se vybírá tak, aby byla splněna nerovnost z poznámky 51, je metodou stejnoměrně spádových směrů a platí $s_i \sim g_i$.

Věta 26 *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 (S2a) a (S3a) s $\varepsilon_3 = \varepsilon_2$, přičemž platí*

$$0 < \beta_i \leq \frac{\overline{\varepsilon}_2}{\varepsilon_2} \frac{\|g_{i+1}\|^2}{\|g_i\|^2}, \tag{52}$$

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 podmínkám (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 (52), můžeme použít nerovnost (48) uvedenou v poznámce 51.
(a) Zřejmě $\|e_i\| = O(\|e_{i-1}\|)$ (poznámka 31) a $\|g_{i-1}\| \sim \|e_{i-1}\|$ (věta 3), takže $\|g_i\| = O(\|g_{i-1}\|)$. Existuje tedy konstanta $c < \infty$ tak, 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 (52) 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 (48) (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 (48) a Schwarzovy nerovnosti dostaneme

$$\|s_i\|\|g_i\| \geq -s_i^T g_i \geq \frac{1-2\overline{\varepsilon}_2}{1-\overline{\varepsilon}_2}\|g_i\|^2,$$

což dává $\|s_i\| \geq \underline{c}\|g_i\|$. Jelikož z (a) plyne $\|s_i\| \leq \overline{c}\|g_i\|$, platí $s_i \sim g_i$. □

3.4 Asymptotická rychlost konvergence

Nyní budeme vyšetřovat cyklicky přerušované metody sdružených gradientů s asymptoticky přesným výběrem délky kroku. Budeme předpokládat, že $e_i \neq 0$ a $g_i \neq 0$ $\forall i \in N$, neboť v opačném případě iterační proces končí ve stacionárním bodě. Dále budeme předpokládat, že

$$\|e_i\| \sim \|e_l\| \quad \forall l = nk + 1 \in M, \quad \forall l \leq i < l + n.$$

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 31 je $\|e_i\| = O(\|e_l\|)$), takže rychlost konvergence je vyšší než lineární (tato úvaha je precizována ve větě 28).

Věta 27 *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 podmínkám (F4) a (F5). Nechť $\|e_i\| \sim \|e_l\|$ $\forall l \in M$, $\forall l \leq i < l + n$. Pak pro $i \in N$ platí*

$$\beta_i = O(1), \tag{53}$$

$$s_i \sim g_i, \quad \alpha_i \sim 1, \tag{54}$$

$$-s_i^T g_i = g_i^T g_i(1 + o(1)). \tag{55}$$

Důkaz (Pro metodu Hestenese a Stiefela (42)). Poznamenejme, že předpokláme, že $e_i \neq 0$ a $g_i \neq 0$ pro $l \leq i < l+n$. Důkaz věty provedeme indukcí. Dokážeme navíc, že pro $l \leq j < i < l+n$ platí

$$s_j^T g_i = o(\|e_l\|^2), \tag{56}$$

$$g_j^T g_i = o(\|e_l\|^2), \tag{57}$$

$$s_j^T G^* s_i = o(\|e_l\|^2), \tag{58}$$

$$s_j^T y_i = y_j^T s_i = o(\|e_l\|^2), \tag{59}$$

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 asymptotické přesnosti výběru délky kroku plyne, že $\alpha_l \sim \|g_l\|/\|s_l\|$ (lemma 5), což spolu s $s_l \sim g_l$ dává $\alpha_l \sim 1$. Dále není co dokazovat (platí $\beta_{l-1} = 0 = O(1)$ a vztah pro β_l je dokázán v (c)). Nechť $l \leq i < l+n-1$.
(a) Podle indukčních předpokladů (56) a (59) 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 asymptotické přesnosti výběru délky kroku a z (55) 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 $l \leq j \leq i$.
(b) Zřejmě

$$\begin{aligned} g_l &= -s_l, \\ g_j &= -s_j + \beta_{j-1} s_{j-1} \qquad \forall l < j \leq i, \end{aligned}$$

takže podle (a) a (53) platí

$$\begin{aligned} g_l^T g_{i+1} &= -s_l^T g_{i+1} = o(\|e_l\|^2), \\ g_j^T g_{i+1} &= -s_j^T g_{i+1} + \beta_{j-1} s_{j-1}^T g_{i+1} = o(\|e_l\|^2) \qquad \forall l < j \leq i. \end{aligned}$$

(c) Protože $g_{i+1} \sim e_{i+1} \sim e_l$, můžeme podle (b) 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 asymptotické přesnosti výběru délky kroku a z (55) 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

$$\beta_i = \frac{y_i^T g_{i+1}}{y_i^T s_i} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} \frac{(1+o(1))}{(1+o(1))} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}(1+o(1)) = O(1),$$

neboť z $g_{i+1} \sim e_{i+1} \sim e_l$ a $g_i \sim e_i \sim e_l$ plyne $g_{i+1} \sim g_i$.
(d) Podle (53) a (54) platí

$$\|s_{i+1}\| \leq \|g_{i+1}\| + \|\beta_i s_i\| = \|g_{i+1}\| + O(1)\|s_i\| = \|g_{i+1}\| + O(\|g_i\|) = O(\|g_{i+1}\|)$$

a z asymptotické přesnosti výběru délky kroku a z (53) a (55) plyne, že

$$\begin{aligned} s_{i+1}^T s_{i+1} &= (-g_{i+1} + \beta_i s_i)^T(-g_{i+1} + \beta_i s_i) \geq g_{i+1}^T g_{i+1} - 2\beta_i g_{i+1}^T s_i = g_{i+1}^T g_{i+1} - g_i^T s_i o(1) \\ &= g_{i+1}^T g_{i+1} + g_i^T g_i(1+o(1))o(1) = g_{i+1}^T g_{i+1}(1+o(1)) \end{aligned}$$

(používáme relaci $g_{i+1} \sim g_i$). Spojením obou nerovností dostaneme $s_{i+1} \sim g_{i+1}$. Z asymptotické přesnosti výběru délky kroku plyne, že $\alpha_{i+1} \sim \|g_{i+1}\|/\|s_{i+1}\|$ (lemma 5), což spolu s $s_{i+1} \sim g_{i+1}$ dává $\alpha_{i+1} \sim 1$.

(e) Z asymptotické přesnosti výběru délky kroku a z (53) a (55) 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 $g_{i+1} \sim g_i$).
(f) Použijeme-li (53), (59) 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 $1 \leq j < i$ a podle (42) 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 $1 \leq j \leq i$. Použijeme-li větu 3, můžeme pro $1 \leq j \leq i$ psát

$$y_j = g_{j+1} - g_j = G^* d_j + o(\|d_j\|) = \alpha_j G^* s_j + o(\|e_l\|),$$

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 (54) 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 3, dostaneme

$$y_{i+1} = g_{i+2} - g_{i+1} = G^* d_{i+1} + o(\|d_{i+1}\|) = \alpha_{i+1} G^* s_{i+1} + o(\|e_l\|),$$

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 (54) platí $s_j \sim g_j \sim e_l$ a podle (d) je $\alpha_{i+1} \sim 1$. □

Poznámka 58 Je-li kromě podmínek (F4) a (F5) splněna i podmínka (F6), můžeme místo věty 3 použít větu 4 a tudíž místo $o(1)$ a $o(\|e_l\|^2)$ psát $O(\|e_l\|)$ a $O(\|e_l\|^3)$.

Poznámka 59 Podle věty 27 (vztah (53)) existuje pro cyklicky přerušovanou metodu sdružených gradientů s asymptoticky přesným výběrem délky kroku index $\underline{l} \in M$ takový, že nerovnost uvedená v poznámce 54 je splněna pro $i \geq \underline{l}$. Pak již nedochází k přerušování iteračního procesu vlivem porušení této nerovnosti a platí beze zbytku definice 20. Abychom zjednodušili některé úvahy, budeme od této chvíle předpokládat, že $\underline{l} = 1$ (v opačném případě lze posunout indexy aniž by se změnilo asymptotické chování uvažované posloupnosti).

Definice 21 *Při vyšetřování asymptotického chování metod sdružených gradientů budeme porovnávat dva iterační procesy, původní iterační proces*

$$x_{i+1} = x_i + \alpha_i s_i, \quad i \in N,$$

použitý pro minimalizaci funkce $F(x)$, a referenční iterační proces

$$\bar{x}_{i+1} = \bar{x}_i + \bar{\alpha}_i \bar{s}_i, \quad i \in N,$$

použitý pro minimalizaci kvadratické funkce

$$Q(x) = F(x^*) + \frac{1}{2}(x - x^*)^T G^* (x - x^*),$$

která má v bodě x^* stejnou hodnotu, gradient a Hessovu matici jako funkce F. Veličiny spjaté s původním 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 procesem budeme označovat symboly s pruhem $\bar{x}_i$, $\bar{g}_i$, $\bar{\alpha}_i$, $\bar{s}_i$, $\bar{e}_i = \bar{x}_i - x^*$, $\bar{d}_i = \bar{x}_{i+1} - \bar{x}_i = \alpha_i \bar{s}_i$, $\bar{y}_i = \bar{g}_{i+1} - g_i$. Oba procesy budeme cyklicky startovat v bodech $x_l \in R^n$, $l = nk + 1 \in M$, $k \in N$ tak, že $\bar{x}_l = x_l$, $\bar{e}_l = e_l$.

Lemma 8 *Nechť jsou splněny předpoklady věty 27. Nechť $\bar{x}_i \in R^n$, $i \in N$, je referenční posloupnost z definice 21 získaná metodou sdružených gradientů s přesným výběrem délky kroku aplikovanou na kvadratickou funkci $Q(x)$ a odstartovanou v bodě x_l. Pak $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \leq i \leq l+n$.*

Důkaz Poznamenejme, že z $e_i - \bar{e}_i = o(\|e_l\|)$ a $e_i \sim e_l$ plyne $\bar{e}_i = e_i + o(\|e_l\|)) \sim e_l = \bar{e}_l$, takže referenční posloupnost $\bar{x}_i \in R^n$, $i \in N$, vyhovuje předpokladům věty 27. Pro $l \leq i < l+n$ tedy platí $\bar{\beta}_i = O(1)$, $\bar{\alpha}_i \sim 1$ a $\bar{s}_i \sim \bar{g}_i \sim \bar{e}_i \sim \bar{e}_l = e_l$. Důkaz věty provedeme indukcí. Dokážeme navíc, že pro $i \in N$ platí

$$\bar{\alpha}_i = \alpha_i(1+o(1)), \quad \bar{\beta}_i = \beta_i(1+o(1)),$$

$$\bar{e}_i = e_i(1+o(1)), \quad \bar{g}_i = g_i(1+o(1))$$

a

$$\bar{s}_i = s_i(1+o(1)).$$

Na začátku cyklu platí $\bar{e}_l = e_l = e_l(1+o(1))$ a použijeme-li větu 3, dostaneme $\bar{g}_l = g_l + o(\|e_l\|) = g_l(1+o(1))$, což spolu s $s_l = -g_l$ dává $\bar{s}_l = s_l(1+o(1))$. Použijeme-li lemma 5, můžeme psát

$$\bar{\alpha}_l = -\frac{\bar{s}_l^T \bar{g}_l}{\bar{s}_l^T G^* \bar{s}_i} = -\frac{s_l^T g_l (1+o(1))^2}{s_l^T G^* s_l (1+o(1))^2} = \alpha_l(1+o(1)),$$

neboť $\bar{\alpha}_l = -\bar{s}_l^T \bar{g}_l / \bar{s}_l^T G^* \bar{s}_l$ realizuje pro kvadratickou funkci $Q(x)$ přesný výběr délky kroku a z asymptotické přesnosti výběru délky kroku plyne $-s_l^T g_l / s_l^T G^* s_l = \alpha_l(1+o(1))$. Dále není co dokazovat (platí $\bar{\beta}_{l-1} = 0$ a vztah pro $\bar{\beta}_l$ je dokázán v (b)). Nechť $l \leq i < l+n-1$.
(a) Jelikož podle věty 27 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 psát

$$\bar{e}_{i+1} = \bar{e}_i + \bar{\alpha}_i \bar{s}_i = e_i(1+o(1)) + \alpha_i s_i (1+o(1))^2 = e_{i+1} + o(\|e_l\|) = e_{i+1}(1+o(1))$$

a použijeme-li větu 3, dostaneme

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

Platí tedy

$$\bar{g}_{i+1} = \bar{g}_i + \bar{\alpha}_i G^* \bar{s}_i = g_i(1+o(1)) + \alpha_i G^* s_i (1+o(1))^2 = g_i + \alpha_i G^* s_i + o(\|e_l\|) = g_{i+1}(1+o(1)).$$

(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 (F4) a (F5) plyne

$$y_i = \int_0^1 G(x_i + t d_i) d_i \mathrm{d}t \sim d_i = \alpha_i s_i \sim e_l.$$

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

$$\begin{aligned} \bar{s}_{i+1} &= -\bar{g}_{i+1} + \bar{\beta}_i \bar{s}_i = -g_{i+1}(1+o(1)) + \beta_i s_i (1+o(1))^2 \\ &= -g_{i+1} + \beta_i s_i + o(\|e_l\|) = s_{i+1}(1+o(1)). \end{aligned}$$

(d) Podle lemmatu 5 a indukčních předpokladů platí

$$\bar{\alpha}_{i+1} = -\frac{\bar{s}_{i+1}^T \bar{g}_{i+1}}{\bar{s}_{i+1}^T G^* \bar{s}_{i+1}} = -\frac{s_{i+1}^T g_{i+1}(1+o(1))^2}{s_{i+1}^T G^* s_{i+1}(1+o(1))^2} = -\frac{s_{i+1}^T g_{i+1}}{s_{i+1}^T G^* s_{i+1}}(1+o(1)) = \alpha_{i+1}(1+o(1)),$$

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 $-s_{i+1}^T g_{i+1} / s_{i+1}^T G^* s_{i+1} = \alpha_{i+1}(1 + o(1))$.
(e) Nechť $l \le i < l + n$. Pak podle indukčních předpokladů platí

$$\bar{e}_{i+1} = \bar{e}_i + \bar{\alpha}_i \bar{s}_i = e_i(1 + o(1)) + \alpha_i s_i (1 + o(1))^2 = e_i + \alpha_i s_i + o(\|e_l\|) = e_{i+1} + o(\|e_l\|),$$

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 60 Tvrzení lemmatu 8 platí, pokud $\|e_i\| \sim \|e_l\|$ pro $l \le i < l + n$. Jestliže $\|e_i\| \sim \|e_l\|$ pouze pro $l \le i < l + m$, kde $m < n$, můžeme psát $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \le i \le l + m$ (plyne to z induktivní povahy důkazu).

Věta 28 *(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 podmínkám (F4) a (F5). Pak platí*

$$\lim_{l \overset{M}{\to} \infty} \frac{\|x_{l+n} - x^*\|}{\|x_l - x^*\|} = 0.$$

Důkaz Ve větě 28 mlčky předpokládáme, že k přerušování iteračního procesu dochází vždy po n krocích (poznámka 59). Podle věty 22 víme, že metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce $Q(x)$ po $m \le n$ krocích. Mohou nastat dva případy.
(a) Jestliže $\|e_i\| = o(\|e_l\|)$ pro nějaký index $l < i < l + m$, pak z $e_{l+m} = O(\|e_i\|)$ (poznámka 31), plyne $e_{l+m} = o(\|e_l\|)$.
(b) Protože referenční metoda sdružených gradientů najde minimum kvadratické funkce $Q(x)$ po m krocích, platí $\|\bar{e}_{l+m}\| = 0$. Použijeme-li tvrzení lemmatu 8 (které podle poznámky 60 platí pro $l \le i \le l + m$), dostaneme

$$\|e_{l+m}\| \le \|\bar{e}_{l+m}\| + \|e_{l+m} - \bar{e}_{l+m}\| = o(\|e_l\|).$$

Jelikož $e_{l+n} = O(\|e_{l+m}\|)$ (poznámka 31), v obou případech platí $\|e_{l+n}\| = o(\|e_l\|)$, což dává tvrzení věty. □

Poznámka 61 Podle věty 7 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 9 *Nechť jsou splněny předpoklady věty 22. Nechť $g_i \neq 0$ pro nějaký index $1 \le i \le n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \le \max_{1 \le k \le n} P_i^2(\lambda_k), \tag{60}$$

kde $P_i(\lambda)$ je libovolný polynom stupně i takový, že $P_i(0) = 1$, a λ_k, $1 \le k \le n$, jsou vlastní čísla matice G seřazená vzestupně.

Důkaz (a) Dokážeme indukcí, že pro $1 \le j \le i$ platí $g_j \in \mathcal{K}_j$ a $s_j \in \mathcal{K}_j$, kde

$$\mathcal{K}_j = \operatorname{span}\{g_1, Gg_1, \ldots, G^{j-1} g_1\}$$

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 $g_j = g_{j-1} + \alpha_{j-1} G s_{j-1}$ a protože podle indukčního předpokladu platí

$$g_{j-1} \in \mathcal{K}_{j-1}, \quad Gs_{j-1} \in \operatorname{span}(Gg_1, G^2 g_1, \ldots, G^{j-1} g_1) \subset \mathcal{K}_j,$$

dostaneme $g_j \in \mathcal{K}_j$. Podle (36) 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

$$\begin{aligned} x_{i+1} - x^* &= x_1 - x^* + \sum_{j=1}^{i} \alpha_j s_j = x_1 - x^* + p_{i-1}^*(G) g_1 = \\ &= x_1 - x^* + p_{i-1}^*(G) G (x_1 - x) = (I + G p_{i-1}^*(G))(x_1 - x^*), \end{aligned}$$

kde $p_{i-1}^*(G)$ je nějaký polynom stupně $i-1$ v G (matice $p_{i-1}^*(G)$ a G komutují). Označme $P_i^* = I + G p_{i-1}^*$, takže P_i^* je polynom stupně i a $P_i^*(0) = 1$. Jelikož podle poznámky 22 bod $x_{i+1} = x_1 + P_i^*(G)(x_1 - x)$ realizuje minimum ryze konvexní kvadratické funkce Q na $\mathcal{K}_i$, platí

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &= \frac{1}{2}(x_{i+1} - x^*)^T G (x_{i+1} - x^*) = \frac{1}{2}(x_1 - x^*)^T P_i^*(G) G P_i^*(G)(x_1 - x^*) \leq \\ &\leq \frac{1}{2}(x_1 - x^*)^T P_i(G) G P_i(G)(x_1 - x^*) \end{aligned}$$

pro libovolný polynom P_i stupně i takový, že $P_i(0) = 0$. Nechť λ_k a v_k $1 \leq k \leq n$ jsou vlastní čísla (nezáporná) a vlastní vektory (ortonormální) 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

$$\begin{aligned} Q(x_{i+1}) - Q(x^*) &\leq \frac{1}{2}(x_1 - x^*)^T P_i(G) G P_i(G)(x_1 - x^*) = \\ &= \frac{1}{2} \left(\sum_{k=1}^{n} P_i(\lambda_k) \gamma_k v_k \right)^T G \left(\sum_{k=1}^{n} P_i(\lambda_k) \gamma_k v_k \right) = \\ &= \frac{1}{2} \sum_{k=1}^{n} P_i^2(\lambda_k) \gamma_k^2 \lambda_k \leq \frac{1}{2} \max_{1 \leq k \leq n} P_i^2(\lambda_k) \sum_{k=1}^{n} \gamma_k^2 \lambda_k. \end{aligned}$$

Po vydělení dostaneme

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \max_{1 \leq k \leq n} P_i^2(\lambda_k).$$

□

Věta 29 *Nechť jsou splněny předpoklady věty 22. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \left(\frac{\lambda_{m+1-i} - \lambda_1}{\lambda_{m+1-i} + \lambda_1} \right)^2, \tag{61}$$

kde λ_k, $1 \leq k \leq m$, jsou různá vlastní čísla matice G seřazená vzestupně.

Důkaz Podle lemmatu 9 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 $(\lambda_1 + \lambda_{m+1-i})/2$ a λ_{m+1-j}, $1 \leq j \leq i-1$. Tento polynom stupně i má i reálných kořenů, takže jeho kořeny a stacionární body se střídají. Nebudeme provádět podrobnou analýzu tohoto polynomu, spokojíme se pouze s konstatováním, že v intervalu $\lambda_1 \leq \lambda \leq \lambda_{m+1-i}$ (v okolí prvního kořenu) platí

$$|P_i(\lambda)| \leq \left| 1 - \frac{2\lambda}{\lambda_1 + \lambda_{m+1-i}} \right|.$$

Výraz na pravé straně této nerovnosti nabývá v intervalu $\lambda_1 \leq \lambda \leq \lambda_{m+1-i}$ maxima pro $\lambda = \lambda_1$ nebo $\lambda = \lambda_{m+1-i}$. 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 (60) dává (61). □

Důsledek 1 *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 (61) $i = m$, 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 30 *Nechť jsou splněny předpoklady věty 22. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq 4 \left(\frac{\sqrt{\kappa(G)} - 1}{\sqrt{\kappa(G)} + 1} \right)^{2i}. \tag{62}$$

Důkaz Podle lemmatu 9 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 hodnotu 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

$$\begin{aligned} \frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} &\leq \left(\max_{1 \leq k \leq n} |P_i(\lambda_k)|\right)^2 \leq \left(\frac{1}{T_i\left(\frac{\lambda_n + \lambda_1}{\lambda_n - \lambda_1}\right)}\right)^2 \leq \\ &\leq 4\left(\frac{\sqrt{\lambda_n} - \sqrt{\lambda_1}}{\sqrt{\lambda_n} + \sqrt{\lambda_1}}\right)^{2i} = 4\left(\frac{\sqrt{\kappa(G)} - 1}{\sqrt{\kappa(G)} + 1}\right)^{2i}. \end{aligned}$$

□

Poznámka 62 Použijeme-li odhad (62) spolu s (c) a (d) z důkazu věty 12, 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 63 Větu 30 lze snadno zobecnit tak, aby platila pro předpodmíněnou metodu sdružených gradientů. Podle poznámky 50 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 31 *(Asymptotický odhad) Nechť jsou splněny předpoklady věty 27. 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 31. Pak podle poznámky 62 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 8, 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 64 Věta 31 se týká pouze vnitřních iterací každého cyklu. Celkově je cyklicky přerušovaná metoda sdružených gradientů s přesným výběrem délky kroku R-superlineárně konvergentní (poznámka 61).

Poznámka 65 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

3.5 Modifikace a implementace metod sdružených gradientů

Abychom zlepšili účinnost metod sdružených gradientů můžeme vztah (36) 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 platnost věty 22 zůstane zachována. Jednou z možností je nahradit vztah (36) předpisem

$$s_1 = -g_1 \quad a \quad s_{i+1} = -\left(1 + \beta_i \frac{g_{i+1}^T s_i}{g_{i+1}^T g_{i+1}}\right) g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{63}$$

kde β_i je některá z hodnot (42) nebo (43).

Věta 32 *Pro modifikovanou metodu sdružených gradientů (63) platí věta 22. Navíc pro $i \in N$ platí*

$$-g_{i+1}^T s_{i+1} = g_{i+1}^T g_{i+1}. \tag{64}$$

Důkaz Provádíme-li přesný výběr délky kroku, platí $g_{i+1}^T s_i = 0$, takže (63) přejde na (36) a platnost věty 22 zůstane zachována. Vynásobíme-li vztah (63) skalárně vektorem g_{i+1}, dostaneme rovnost (64). □

Dosadíme-li hodnotu β_i^{CD} do (63), dostaneme $s_{i+1} = -\vartheta_i^{CD} g_{i+1} + \beta_i^{CD} s_i$, kde $\vartheta_i^{CD} = -y_i^T s_i / g_i^T s_i$. Podobným způsobem lze modifikovat i metodu FR. Tyto modifikace dovolují značně oslabit podmínky pro globální konvergenci.

Věta 33 *Uvažujme modifikované metody DY, FR, CD dané předpisem*

$$s_1 = -g_1 \quad a \quad s_{i+1} = -\vartheta_i g_{i+1} + \beta_i s_i \quad pro \quad i \in N, \tag{65}$$

kde hodnoty β_i^{DY}, β_i^{FR}, β_i^{CD} *jsou určeny podle (43) a*

$$\vartheta_i^{DY} = \frac{y_i^T s_i}{y_i^T s_i} = 1, \quad \vartheta_i^{FR} = \frac{y_i^T s_i}{g_i^T g_i}, \quad \vartheta_i^{CD} = \frac{y_i^T s_i}{|g_i^T s_i|}. \tag{66}$$

Pak platí věta 22. Splňuje-li funkce $F : \mathcal{D} \to R$ *podmínky (F1) a (F3) a používáme-li při výběru délky kroku zobecněnou Wolfeho podmínku (S2a) a (S3a), kde* $0 < \varepsilon_1 < \varepsilon_2 < 1$ *a* $0 \le \varepsilon_3 < \infty$, *jsou tyto metody globálně konvergentní.*

Důkaz Provádíme-li přesný výběr délky kroku, platí $y_i^T s_i = -g_i^T s_i = g_i^T g_i$ (poznámka 49), neboli $\vartheta_i^{DY} = \vartheta_i^{FR} = \vartheta_i^{CD} = 1$, takže (65) přejde na (36) a platnost věty 22 zůstane zachována. Nyní dokážeme globální konvergenci.

(a) Jelikož $\vartheta_i^{DY} = 1$, metoda DY se použitím (65) nezmění, takže globální konvergence plyne z věty 23.

(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 (64). Modifikovaná metoda FR je tedy totožná s modifikovanou metodou CD a pro obě tyto metody je splněna rovnost (64).

(c) Uvažujme modifikovanou metodu CD. Jelikož je splněna rovnost (64), jsou směrové vektory s_i, $i \in N$, spádové a platí (14) s $\tilde{\varepsilon}_0 = 1$, což podle poznámky 24 implikuje nerovnost (15). Protože při výběru délky kroku používáme zobecněnou Wolfeho podmínku (S2a) a (S3a), platí

$$y_i^T s_i = g_{i+1}^T s_i - g_i^T s_i \le \varepsilon_3 |g_i^T s_i| - g_i^T s_i = (1 + \varepsilon_3)|g_i^T s_i|,$$

neboli $\vartheta_i \le 1 + \varepsilon_3$. Použijeme-li tento odhad spolu se vztahy (64)–(66), můžeme psát

$$\begin{aligned} \|s_{i+1}\|^2 &= \left(-\vartheta_i g_{i+1} + \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right)^T \left(-\vartheta_i g_{i+1} + \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} s_i\right) \\ &= \vartheta_i^2 \|g_{i+1}\|^2 - 2\vartheta_i \frac{\|g_{i+1}\|^2}{|g_i^T s_i|} g_{i+1}^T s_i + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2 \\ &\le (1 + \varepsilon_3)^2 \|g_{i+1}\|^2 + 2\varepsilon_2 (1 + \varepsilon_3) \|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{|g_i^T s_i|^2} \|s_i\|^2, \end{aligned}$$

neboli

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \le \frac{(1 + \varepsilon_3)(1 + 2\varepsilon_2 + \varepsilon_3)}{\|g_{i+1}\|^2} + \frac{\|s_i\|^2}{\|g_i\|^4}.$$

Předpokládejme, že neplatí

$$\liminf_{i \to \infty} \|g_i\| = 0.$$

Pak existuje konstanta $\underline{\varepsilon} > 0$ taková, že $\|g_i\| \ge \underline{\varepsilon}$ $\forall i \in N$, takže z předchozí nerovnosti plyne

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \le \frac{(1 + \varepsilon_3)(1 + 2\varepsilon_2 + \varepsilon_3)}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \le \frac{(1 + \varepsilon_3)(1 + 2\varepsilon_2 + \varepsilon_3)}{\underline{\varepsilon}^2} (i + 1)$$

(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 (15) uvedenou v poznámce 24. □

Poznámka 66 Z věty 33 plyne, že modifikace (65) dovoluje značně oslabit podmínky pro globální konvergenci metod FR a CD. Stačí vybíráme-li délku kroku pomocí zobecněné Wolfeho podmínky (S2a), 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 (65) lze také použít ke zlepšení konjugovanosti směrových vektorů v metodách PR a LS.

Věta 34 *Uvažujme modifikace metod 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 (42) a

$$\vartheta_i^{HS} = \frac{y_i^T s_i}{y_i^T s_i} = 1, \quad \vartheta_i^{PR} = \frac{y_i^T s_i}{g_i^T g_i}, \quad \vartheta_i^{LS} = \frac{y_i^T s_i}{|g_i^T s_i|}. \tag{67}$$

Pak platí věta 22 a navíc

$$y_i^T s_{i+1} = 0 \quad pro \quad i \in N. \tag{68}$$

Důkaz Tak jako v důkazu věty 33 platí $\vartheta_i^{HS} = \vartheta_i^{PR} = \vartheta_i^{LS} = 1$, používáme-li přesný výběr déky kroku, takže (65) přejde na (36) a platnost věty 22 zůstane zachována. Metoda HS, pro kterou platí (68), se nezmění. V případě metod PR a LS dostaneme

$$y_i^T s_{i+1} = -\frac{y_i^T s_i}{g_i^T g_i} y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{g_i^T g_i} y_i^T s_i = 0$$

a

$$y_i^T s_{i+1} = -\frac{y_i^T s_i}{|g_i^T s_i|} y_i^T g_{i+1} + \frac{y_i^T g_{i+1}}{|g_i^T s_i|} y_i^T s_i = 0.$$

□

Poznámka 67 Použitím vzorce (65) nelze zajistit spádovost směrových vektorů metod HS, PR, LS. To umožňuje vztah (63) nebo vzorec

$$s_1 = -g_1 \quad a \quad s_{i+1} = -g_{i+1} + \beta_i s_i - \gamma_i y_i \quad pro \quad i \in N, \tag{69}$$

kde hodnoty β_i^{HS}, β_i^{PR}, β_i^{LS} jsou určeny podle (42) a

$$\gamma_i^{HS} = \frac{g_{i+1}^T s_i}{y_i^T s_i}, \quad \gamma_i^{PR} = \frac{g_{i+1}^T s_i}{g_i^T g_i}, \quad \gamma_i^{LS} = \frac{g_{i+1}^T s_i}{|g_i^T s_i|}. \tag{70}$$

Vynásobením vzorce (69) skalárně vektorem g_{i+1} se můžeme snadno přesvědčit, že platí (64). Z hlediska praktické účinnosti je vzorec (69) méně vhodný než vztah (63).

Poznámka 68 Základní metody sdružených gradientů lze také kombinovat tak, že volíme

$$\beta_i = \frac{\lambda_i^1 g_{i+1}^T y_i + \lambda_i^2 g_{i+1}^T g_{i+1}}{\mu_i^1 y_i^T s_i + \mu_i^2 g_i^T g_i - \mu_i^3 g_i^T s_i} = \frac{g_{i+1}^T (g_{i+1} - \lambda_i^1 g_i)}{\mu_i^1 y_i^T s_i + \mu_i^2 g_i^T g_i - \mu_i^3 g_i^T s_i}, \tag{71}$$

kde λ_i^1, λ_i^2, μ_i^1, μ_i^2, μ_i^3 jsou nezáporná čísla taková, že $\lambda_i^1 + \lambda_i^2 = 1$ a $\mu_i^1 + \mu_i^2 + \mu_i^3 = 1$. Jednou z možností je volba $\lambda_i^1 = \min(1, \|g_{i+1}\|/\|g_i\|)$, která vede k modifikacím

$$\beta_i^{HSM} = \frac{g_{i+1}^T \tilde{y}_i}{y_i^T s_i}, \quad \beta_i^{PRM} = \frac{g_{i+1}^T \tilde{y}_i}{g_i^T g_i}, \quad \beta_i^{LSM} = \frac{g_{i+1}^T \tilde{y}_i}{|g_i^T s_i|}, \tag{72}$$

kde

$$\tilde{y}_i = g_{i+1} - \min\left(1, \frac{\|g_{i+1}\|}{\|g_i\|}\right) g_i. \tag{73}$$

Poznámka 69 Jak již bylo zmíněno, můžeme do vzorce (36) 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 uvedené v oddílu 4.1. V tomto případě platí

$$H_{i+1} = I + \left(\frac{y_i^T y_i}{y_i^T d_i} + 1\right) \frac{1}{y_i^T d_i} d_i d_i^T - \frac{1}{y_i^T d_i}(y_i d_i^T + d_i y_i^T),$$

kde $d_i = x_{i+1} - x_i = \alpha_i s_i$ a $y_i = g_{i+1} - g_i$. Pokud $s_i^T g_{i+1} = 0$, můžeme psát

$$-H_{i+1} g_{i+1} = -g_{i+1} - \left(\frac{y_i^T y_i}{y_i^T d_i} + 1\right) \frac{d_i^T g_{i+1}}{y_i^T d_i} d_i + \frac{d_i^T g_{i+1}}{y_i^T d_i} y_i + \frac{y_i^T g_{i+1}}{y_i^T d_i} d_i = -g_{i+1} + \frac{y_i^T g_{i+1}}{y_i^T d_i} d_i.$$

Matice H_{i+1} je pozitivně definitní, takže směrový vektor s_{i+1} je spádový i když $s_i^T g_{i+1} \neq 0$. Tato myšlenka tvoří podklad pro metody s proměnnou metrikou s omezenou pamětí a bude dále rozvíjena v oddílu 8.1.

Poznámka 70 Matice uvedená v předchozí posnámce splňuje kvazinewtonovskou podmínku $H_{i+1} y_i = d_i$, takže $-y_i^T s_{i+1} = y_i^T H_{i+1} g_{i+1} = d_i^T g_{i+1}$, zatímco pro metodu HS platí $-y_i^T s_{i+1} = 0$. 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 (36), 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

$$\beta_i^{DL} = \frac{y_i^T g_{i+1} - \rho_i d_i^T g_{i+1}}{y_i^T s_i} = \beta_i^{HS} - \rho_i \frac{d_i^T g_{i+1}}{y_i^T s_i} \tag{74}$$

(DL – Dai a Liao). Také se používá hodnota

$$\beta_i^{DL+} = \max(0, \beta_i^{HS}) - \rho_i \frac{d_i^T g_{i+1}}{y_i^T s_i}. \tag{75}$$

Parametr ρ_i lze volit různým způsobem, například tak jako v oddílu 4.4. Dai a Liao používají konstantní hodnotu $\rho_i = 0.1$. Položíme-li ve vzorci (74) $\rho_i = 2 y_i^T y_i / y_i^T d_i$, dostaneme

$$\beta_i^{HZ} = \frac{y_i^T g_{i+1}}{y_i^T s_i} - 2 \frac{y_i^T y_i}{y_i^T s_i} \frac{s_i^T g_{i+1}}{y_i^T s_i} \tag{76}$$

(HZ – Hager a Zhang).

Věta 35 *Nechť s_{i+1} je směrový vektor určený metodou HZ. Pak platí*

$$-g_{i+1}^T s_{i+1} \geq \frac{7}{8} g_{i+1}^T g_{i+1}. \tag{77}$$

Důkaz Použijeme-li (36) a (76), dostaneme

$$\begin{aligned} g_{i+1}^T s_{i+1} &= -g_{i+1}^T g_{i+1} + \left(\frac{y_i^T g_{i+1}}{y_i^T s_i} - 2\frac{y_i^T y_i}{y_i^T s_i}\frac{s_i^T g_{i+1}}{y_i^T s_i} \right) g_{i+1}^T s_i \\ &= \frac{y_i^T g_{i+1} y_i^T s_i s_i^T g_{i+1} - g_{i+1}^T g_{i+1}(y_i^T s_i)^2}{(y_i^T s_i)^2} - 2\frac{y_i^T y_i (s_i^T g_{i+1})^2}{(y_i^T s_i)^2} \end{aligned}$$

Dosadíme-li do nerovnosti

$$u^T v \le \|u\|\|v\| \le \frac{1}{2}(\|u\|^2 + \|v\|^2)$$

vektory

$$u = \frac{1}{2}(y_i^T s_i) g_{i+1}, \quad v = 2(s_i^T g_{i+1}) y_i,$$

můžeme psát

$$y_i^T g_{i+1}\, y_i^T s_i\, s_i^T g_{i+1} \le \frac{1}{2}\left(\frac{1}{4}\|g_{i+1}\|^2 (y_i^T s_i)^2 + 4\|y_i\|^2 (s_i^T g_{i+1})^2 \right)$$

a tedy

$$\begin{aligned} g_{i+1}^T s_{i+1}(y_i^T s_i)^2 &\le \frac{1}{8}\|g_{i+1}\|^2 (y_i^T s_i)^2 - \|g_{i+1}\|^2 (y_i^T s_i)^2 + 2\|y_i\|^2 (s_i^T g_{i+1})^2 - 2\|y_i\|^2 (s_i^T g_{i+1})^2 \\ &= -\frac{7}{8}\|g_{i+1}\|^2 (y_i^T s_i)^2, \end{aligned}$$

což bezprostředně dává (77). □

Poznámka 71 Někdy je výhodné metodu sdružených gradientů škálovat. Škálování je nejjednodušším předpodmíněním, kdy $H = \gamma_{i+1} I$. V tomto případě se místo (36) používá vzorec

$$-s_{i+1} = \gamma_{i+1}(g_{i+1} - \beta_i s_i),$$

kde

$$\gamma_{i+1} = \frac{y_i^T d_i}{y_i^T y_i} \tag{78}$$

a $d_i = x_{i+1} - x_i = \alpha_i s_i, y_i = g_{i+1} - g_i$. Hodnoty β_i^{HS}, β_i^{LS}, β_i^{DY}, β_i^{CD} se nezmění. Zbylé hodnoty je třeba upravit tak, že

$$\beta_i^{PR} = \frac{1}{\gamma_i}\frac{y_i^T g_{i+1}}{g_i^T g_i}, \quad \beta_i^{FR} = \frac{1}{\gamma_i}\frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}.$$

Parametr γ_{i+1}, je nutné udržovat v určitých mezích, například v intervalu $0.01 \le \gamma_{i+1} \le 100$. Pokud hodnota (78) v tomto intervalu neleží, položíme $\gamma_{i+1} = 1$.

Poznámka 72 Účinnost metod sdružených gradientů lze zvýšit vhodným restartováním. Restartování 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 $-g_{i+1}^T s_{i+1} \ge \varepsilon_0 \|g_{i+1}\|\|s_{i+1}\|$, kde $\varepsilon_0 > 0$ je nějaké malé číslo (například $\varepsilon_0 = 10^{-8}$). V poznámce 55 je uvedeno, že takto upravená metoda sdružených gradientů je globálně konvergentní, aniž by k restartům docházelo příliš často. Používáme-li metody (43), je výhodné testovat sdruženost směrových vektorů. V tomto případě se iterační proces přeruší, pokud neplatí

$$y_i^T s_{i+1} \le \eta_1 \|s_{i+1}\|\|y_i\|, \tag{79}$$

kde hodnota η_1 závisí na zvolené Wolfeho podmínce. Také je možné testovat ortogonalitu gradientů. V tomto případě se iterační proces přeruší, pokud neplatí

$$g_i^T g_{i+1} \le \eta_2 \|g_{i+1}\|\|g_i\|. \tag{80}$$

kde hodnota η_2 závisí na zvolené Wolfeho podmínce. Je-li počet proměnných dostatečně velký, vyplatí se iterační proces přerušovat vždy po n krocích, počítaných od posledního restartu (pak jsou splněny předpoklady věty 28).

Poznámka 73 Závěrem uvedeme několik poznámek k implementaci metod sdružených gradientů.

- Chceme-li implementovat metodu sdružených gradientů, musíme se rozhodnout, kterou základní metodu použijeme. Zde se omezíme na metody HS, PR, LS, DY, FR, CD. Metody DL a HZ v této poznámce uvažovat nebudeme.
- Zvolíme-li metody HS, PR, LS, můžeme použít základní hodnoty (42), nezáporné hodnoty (50), kombinované metody (51), nebo modifikované metody (72). Místo vztahu (36) můžeme použít vzorec (63), vzorec (65) s hodnotami (67), nebo vzorec (69) s hodnotami (70). Je vhodné provést restart, není-li směrový vektor dostatečně spádový.
- Zvolíme-li metody DY, FR, CD, můžeme použít základní hodnoty (43). Místo vztahu (36) můžeme použít vzorec (63), nebo vzorec (65) s hodnotami (66). Je vhodné provést restart, nejsou-li směrové vektory dostatečně sdružené (podmínka (79)).
- Pro metody sdružených gradienů je nejvhodnější používat silnou Wolfeho podmínku s parametry $\varepsilon_1 = 10^{-4}$ a $\varepsilon_2 = 10^{-1}$. V tomto případě je vhodné pokládat $\eta_1 = 0.05$ a $\eta_2 = 0.5$. Algoritmus 1 je třeba pozměnit tak, že v něm ponecháme podmínku (S3b) ale společně s podmínkou (S2a) vyhodnocujeme podmínku
$$s_i^T g_{i+1} \leq \varepsilon_2 |s_i^T g_i|,$$
která je částí podmínky (S3a).
- Metody sdružených gradienů jsou citlivé na počáteční volbu délky kroku. Obvykle je vhodné položit
$$\alpha_i^1 = \min\left(1, \frac{2(F_i - F_{i-1})}{s_i^T g_i}, \frac{2(\underline{F} - F_i)}{s_i^T g_i}\right),$$
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.

Algoritmus metody sdružených gradientů lze popsat zhruba takto:

Algoritmus 2 Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-1}$, $\eta_1 = 0.05$, $\eta_2 = 0.5$, $\underline{\varepsilon} > 0$.

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 73. Rozhodneme-li se pro škálování, určíme škálovací koeficient $\gamma_i > 0$ podle poznámky 71 (obvykle se škálování neprovádí, takže $\gamma_i = 1$). Pokud $\gamma_i \neq 0$ vynásobíme směrový vektor s_i číslem γ_i.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1 upraveného podle poznámky 73. 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.

Následující dvě tabulky ukazují srovnání jednotlivých metod sdružených gradientů při minimalizaci 22 testovacích funkcí s 1000 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a selhání S, jakož i celkový čas výpočtu). V první tabulce jsou uvedeny výsledky metod používajících silnou Wolfeho podmínku (S2a) a (S3a), kde $\varepsilon_1 = 10^{-4}$ a $\varepsilon_2 = 10^{-1}$. V podmínce (79) byla použita hodnota $\eta_1 = 0.05$. Druhá tabulka obsahuje výsledky metod používajících slabou Wolfeho podmínku (S2a) a (S3b), kde $\varepsilon_1 = 10^{-4}$ a $\varepsilon_2 = 0.9$. V podmínce (79) byla použita hodnota $\eta_1 = 0.2$.

	Metoda HS			Metoda PR			Metoda LS		
Realizace	NIT - NFV	S	čas	NIT - NFV	S	čas	NIT - NFV	S	čas
(42)	28861 - 42022	-	14.4	41491 - 65603	-	22.2	40696 - 61925	-	21.7
(51)	27259 - 39556	-	12.6	34991 - 51760	-	17.0	33269 - 48688	-	16.9
(63)	27692 - 39905	-	13.0	30326 - 44581	-	15.1	29772 - 43519	-	14.3
(65) + (67)	28861 - 42022	-	14.4	29321 - 44248	-	14.8	30360 - 44452	-	15.0
(69) + (70)	36389 - 57339	-	19.6	36132 - 55544	-	19.3	37120 - 57401	-	19.6
(72) + (73)	30985 - 43307	-	14.9	41891 - 65560	-	22.8	39537 - 60093	-	20.7
	Metoda DY			Metoda FR			Metoda CD		
Realizace	NIT - NFV	S	čas	NIT - NFV	S	čas	NIT - NFV	S	čas
(43)	58777 - 66087	1	16.0	61462 - 80726	1	20.1	93678 -116997	2	37.9
(63)	91617 -115841	2	35.5	60173 - 67924	1	16.0	59440 - 66822	1	15.9
(65) + (66)	58777 - 66087	1	16.0	58218 - 65532	1	15.9	59095 - 66288	1	16.0
(43) + (79)	30358 - 49816	-	12.7	31277 - 51394	-	13.9	35170 - 57068	-	16.1
(63) + (79)	36341 - 61332	-	12.7	29600 - 48584	-	12.4	30115 - 49309	-	12.7
(65) + (79)	30358 - 49816	-	12.7	31111 - 51291	-	13.0	29225 - 47775	-	12.3

	Metoda HS			Metoda PR			Metoda LS		
Realizace	NIT - NFV	S	čas	NIT - NFV	S	čas	NIT - NFV	S	čas
(42)	72582 - 88844	-	31.4	68670 - 90081	1	23.3	79986 -104883	2	27.1
(51)	71904 - 87072	-	27.2	109825 -137358	4	48.0	96854 -119264	2	39.2
(63)	66217 - 80905	-	25.6	74880 - 91404	1	26.6	73783 - 90037	1	25.3
(65) + (67)	72582 - 88844	-	31.3	24342 - 53003	-	14.0	72343 - 88302	-	26.4
	Metoda DY			Metoda FR			Metoda CD		
Realizace	NIT - NFV	S	čas	NIT - NFV	S	čas	NIT - NFV	S	čas
(43)	62785 - 63457	1	15.1	110743 -133252	1	43.9	160486 -177113	4	60.7
(63)	155859 -181578	5	78.7	61272 - 62014	1	14.8	61381 - 62185	1	14.8
(65) + (66)	62785 - 63457	1	15.1	62213 - 62900	1	14.8	61618 - 62321	1	15.0
(43) + (79)	42755 - 46596	-	12.9	63714 - 76588	1	23.1	99616 -113680	2	43.2
(63) + (79)	140668 -169084	-	55.1	42740 - 46553	-	12.5	43380 - 47363	-	13.5
(65) + (79)	42755 - 46596	-	12.9	42156 - 45914	-	12.6	42287 - 46093	-	13.1

Z údajů uvedených v těchto tabulkách lze vyvodit několik závěrů:

- Při realizaci metod sdružených gradientů, zejména metod HS, PR, LS a jejich modifikací, je výhodné používat silnou Wolfeho podmínku s $\varepsilon_2 = 10^{-1}$ (tato hodnota byla získána experimentálně).
- V případě, že používáme silnou Wolfeho podmínku, zlepšují kombinace (51) nebo modifikace (63) a (65) + (67) účinnost metod HS, PR, LS. Metoda HS dává nejlepší výsledky.
- Modifikace (69) + (70) nebo (72) + (73) mírně zlepšují účinnost metod PR, LS, ale zhoršují účinnost metody HS.
- V případě, že používáme silnou Wolfeho podmínku, dávají metody DY, FR, CD horší výsledky než metody HS, PR, LS. Vlastnosti metod DY, FR, CD se výrazně zlepší, přerušujeme-li je vždy tehdy, není-li splněna podmínka (79). Volba hodnoty η_1 v (79) závisí na použité Wolfeho podmínce (vhodnou hodnotu je třeba určit experimentálně).
- Účinnost metod FR a CD výrazně zlepšují modifikace (63) nebo (65) + (66) (které jsou v tomto případě ekvivalentní). U těchto modifikací nezávisí na volbě Wolfeho podmínky, což potvrzuje význam věty 33. Doplníme-li uvedené modifikace o test sdruženosti (79), jsou výsledné metody rovnocenné nejlepším modifikacím metody HS.

3.6 Předpodmíněná metoda sdružených gradientů pro řešení soustav lineárních rovnic

Podle vět 22 a 30 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

$$Q(s) = g^T s + \frac{1}{2} s^T B s, \tag{81}$$

která se používá k určení směrového vektoru v metodách spádových směrů (i v metodách s lokálně omezeným krokem popsaných v páté kapitole). V poznámce 50 jsme ukázali, že metodu sdružených gradientů lze předpodmínit tak, že se místo kvadratické funkce $Q(s)$ minimalizuje kvadratická funkce

$$\tilde{Q}(\tilde{s}) = \tilde{g}^T \tilde{s} + \frac{1}{2} \tilde{s}^T \tilde{B} \tilde{s}, \tag{82}$$

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 50 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 30 zaručuje rychlou konvergenci metody.

Definice 22 *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

$$\begin{aligned} q_i &= B p_i, & \alpha_i &= g_i^T C^{-1} g_i, / p_i^T q_i, \\ s_{i+1} &= s_i + \alpha_i p_i, & g_{i+1} &= g_i + \alpha_i q_i, \\ \beta_i = g_{i+1}^T C^{-1} g_{i+1} / g_i^T C^{-1} g_i, & & p_{i+1} &= -C^{-1} g_{i+1} + \beta_i p_i \end{aligned}$$

pro $1 \le i \le n$, nazveme předpodmíněnou metodou sdružených gradientů (s předpodmiňovačem C) pro řešení soustavy lineárních rovnic $Bs = -g$.

Poznámka 74 Z definice 22 plyne, že $g_i = Bs_i + g$, kde

$$s_i = \sum_{j=1}^{i-1} \alpha_j p_j, \tag{83}$$

Poznámka 75 Hodnota

$$\alpha_i = \frac{g_i^T C^{-1} g_i}{p_i^T B p_i} \tag{84}$$

realizuje přesný výběr délky kroku, neboť z definice 22 plyne, že

$$p_i^T g_{i+1} = p_i^T g_i + \alpha_i p_i^T q_i = -g_i^T C^{-1} g_i + (g_i^T C^{-1} g_i / p_i^T B p_i) p_i^T B p_i = 0.$$

Poznámka 76 Některé vlastnosti nepředpodmíněné metody sdružených gradientů jsou ukázány v důkazu věty 22 (vztahy (38)–(40)). 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 $\tilde{B} = C^{-1/2} B C^{-1/2}$. Platí

$$p_j^T g_i = 0, \tag{85}$$

$$g_j^T C^{-1} g_i = 0, \tag{86}$$

$$p_j^T B p_i = 0 \tag{87}$$

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 77 Budeme 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 36 *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*

$$Q(s_{i+1}) < Q(s_i(\alpha)) < Q(s_i), \tag{88}$$

$$g^T s_{i+1} < g^T s_i(\alpha) < g^T s_i, \tag{89}$$

$$\|s_{i+1}\| > \|s_i(\alpha)\| > \|s_i\|, \tag{90}$$

$$\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} > \frac{g^T s_i(\alpha)}{\|g\|\|s_i(\alpha)\|} > \frac{g^T s_i}{\|g\|\|s_i\|}. \tag{91}$$

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 poznámky 74 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 (85). 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 $0 \leq \alpha < \alpha_i$ a nulová pro $\alpha = \alpha_i$, takže funkce $Q(s_i(\alpha))$ klesá pro $0 \leq \alpha < \alpha_i$ a nabývá svého minima pro $\alpha = \alpha_i$. Pro $\alpha = \alpha_i$ 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 (85)–(86) 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 $1 \leq j < i$ 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 $\alpha \geq 0$.

(c) Použijeme-li vztah (83), 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

$$\begin{aligned} -s_i(\alpha) &= -\sum_{j=1}^{i-1} \alpha_j p_j - \alpha p_i = \sum_{j=1}^{i-1} \alpha_j \|g_j\|^2 \left(\sum_{k=1}^{j} \frac{g_k}{\|g_k\|^2} \right) + \alpha \|g_i\|^2 \left(\sum_{k=1}^{i} \frac{g_k}{\|g_k\|^2} \right) \\ &= \alpha_1 \|g_1\|^2 \left(\frac{g_1}{\|g_1\|^2} \right) + \alpha_2 \|g_2\|^2 \left(\frac{g_1}{\|g_1\|^2} + \frac{g_2}{\|g_2\|^2} \right) + \ldots + \alpha \|g_i\|^2 \left(\frac{g_1}{\|g_1\|^2} + \frac{g_2}{\|g_2\|^2} + \ldots + \frac{g_i}{\|g_i\|^2} \right) \\ &= \sum_{j=1}^{i} \left(\sum_{k=j}^{i-1} \alpha_k \|g_k\|^2 + \alpha \|g_i\|^2 \right) \frac{g_j}{\|g_j\|^2}. \end{aligned}$$

Použijeme-li (86) s $C = I$, dostaneme

$$s_i^T(\alpha) s_i(\alpha) = \sum_{j=1}^{i} \left(\sum_{k=1}^{i-1} \alpha_k \|g_k\|^2 + \alpha \|g_i\|^2 \right)^2 \frac{1}{\|g_j\|^2}$$

a

$$-g^T s_i(\alpha) = \sum_{k=j}^{i-1} \alpha_k \|g_k\|^2 + \alpha \|g_i\|^2,$$

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 (91). □

Důsledek 2 *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*

$$Q(s_{i+1}) < Q(s_i(\alpha)) < Q(s_i),$$
$$g^T s_{i+1} < g^T s_i(\alpha) < g^T s_i,$$
$$\|s_{i+1}\|_C > \|s_i(\alpha)\|_C > \|s_i\|_C, \tag{92}$$
$$\frac{g^T s_{i+1}}{\|g\|_D \|s_{i+1}\|_C} > \frac{g^T s_i(\alpha)}{\|g\|_D \|s_i(\alpha)\|_C} > \frac{g^T s_i}{\|g\|_D \|s_i\|_C}, \tag{93}$$

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 76. □

Věta 37 *Jsou-li splněny předpoklady věty 36, platí*

$$-Q(s_{i+1}) \ge \frac{1}{2}\frac{\|g\|^2}{\|B\|}, \quad -g^T s_{i+1} \ge \frac{\|g\|^2}{\|B\|}. \tag{94}$$

Je-li navíc matice B pozitivně definitní, platí

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \ge \frac{1}{\sqrt{\kappa(B)}}. \tag{95}$$

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

$$-Q(s_2) = \frac{(g^T g)^2}{g^T B g} - \frac{1}{2}\frac{(g^T g)^2 g^T B g}{(g^T B g)^2} = \frac{1}{2}\frac{(g^T g)^2}{g^T B g} \ge \frac{1}{2}\frac{\|g\|^2}{\|B\|},$$

a

$$-g^T s_2 = \frac{(g^T g)^2}{g^T B g} \ge \frac{\|g\|^2}{\|B\|},$$

takže podle (88)–(89) dostaneme (94).
(b) Pokud B je pozitivně definitní, platí $p_j^T B p_j > 0$, kdykoliv $\|g_j\| > 0$, neboť z (85) plyne

$$p_j^T p_j = (-g_j + \beta_{j-1} p_{j-1})^T (-g_j + \beta_{j-1} p_{j-1}) = g_j^T g_j + \beta_{j-1}^2 p_{j-1}^T p_{j-1} \ge g_j^T g_j,$$

neboli $\|p_j\| \ge \|g_j\|$. Podle věty 22 existuje index $m \le n$ takový, že $\|g_j\| > 0$ pro $1 \le j \le m$ a $g_{m+1} = 0$. Podle (91) můžeme pro $i \le m$ psát

$$\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \le \frac{g^T s_{m+1}}{\|g\|\|s_{m+1}\|}.$$

Jelikož $g_{m+1} = g + B s_{m+1} = 0$, je vektor s_{m+1} řešením soustavy $g + Bs = 0$, což podle věty 8 dává

$$-\frac{g^T s_{m+1}}{\|g\|\|s_{m+1}\|} \ge \frac{1}{\sqrt{\kappa(B)}}.$$

Po dosazení do předchozí nerovnosti dostaneme (95). □

Důsledek 3 *Jsou-li splněny předpoklady důsledku 2, platí*

$$-Q(s_{i+1}) \geq \frac{1}{2}\frac{\|g\|^2}{\kappa(C)\|B\|}, \quad -g^T s_{i+1} \geq \frac{\|g\|^2}{\kappa(C)\|B\|}. \tag{96}$$

Je-li navíc matice B pozitivně definitní, platí

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{\kappa(C)\sqrt{\kappa(B)}}. \tag{97}$$

Důkaz (a) Podobně jako v důkazu věty 37 dostaneme

$$-Q(s_2) = -Q(\tilde{s}_2) = \frac{1}{2}\frac{(\tilde{g}^T\tilde{g})^2}{\tilde{g}^T\tilde{B}\tilde{g}} \geq \frac{1}{2}\frac{\|\tilde{g}\|^2}{\|\tilde{B}\|} \geq \frac{1}{2}\frac{g^T C^{-1} g}{\|C^{-1}\|\|B\|} \geq \frac{1}{2}\frac{\|g\|^2}{\kappa(C)\|B\|},$$

a

$$-g^T s_2 = \frac{(\tilde{g}^T\tilde{g})^2}{\tilde{g}^T\tilde{B}\tilde{g}} \geq \frac{\|g\|^2}{\kappa(C)\|B\|},$$

což spolu s (88)–(89) dává (96).
(b) Jelikož

$$\frac{1}{\kappa(C)}\|g\|^2\|s_{i+1}\|^2 \leq g^T C^{-1} g s_{i+1}^T C s_{i+1} \leq \kappa(C)\|g\|^2\|s_{i+1}\|^2$$

pro $1 \leq i \leq m$, můžeme podle (93) psát

$$\frac{(g^T s_{i+1})^2}{\|g\|^2\|s_{i+1}\|^2} \geq \frac{1}{\kappa(C)}\frac{(g^T s_{i+1})^2}{g^T C^{-1} g s_{i+1}^T C s_{i+1}} \geq \frac{1}{\kappa(C)}\frac{(g^T s_{m+1})^2}{g^T C^{-1} g s_{m+1}^T C s_{m+1}} \geq \frac{1}{\kappa^2(C)}\frac{(g^T s_{m+1})^2}{\|g\|^2\|s_{m+1}\|^2}$$

a jelikož vektor s_{m+1} je řešením soustavy $g + Bs = 0$, můžeme použít stejnou nerovnost jako v důkazu věty 37. □

Poznámka 78 Předpokládáme-li, že matice C je vybrána tak, že $\kappa(\tilde{B}) \leq \kappa(B)$, můžeme druhou nerovnost v důsledku 3 nahradit nerovností

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{\sqrt{\kappa(B)\kappa(C)}},$$

neboť podle věty 37 platí

$$\frac{(g^T s_{i+1})^2}{\|g\|^2\|s_{i+1}\|^2} \geq \frac{1}{\kappa(C)}\frac{(g^T s_{i+1})^2}{g^T C^{-1} g s_{i+1}^T C s_{i+1}} = \frac{1}{\kappa(C)}\frac{(\tilde{g}^T\tilde{s}_{i+1})^2}{\tilde{g}^T\tilde{g}\tilde{s}_{i+1}^T\tilde{s}_{i+1}} \geq \frac{1}{\kappa(C)\kappa(\tilde{B})}$$

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 38 *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*

$$Q(s_i(\alpha)) < Q(s_i), \tag{98}$$

$$g^T s_i(\alpha) < g^T s_i, \tag{99}$$

$$\|s_i(\alpha)\| > \|s_i\|, \tag{100}$$

$$\frac{g^T s_i(\alpha)}{\|g\|\|s_i(\alpha)\|} > \frac{g^T s_i}{\|g\|\|s_i\|}. \tag{101}$$

Důkaz Podobně jako v části (a) důkazu věty 36 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) a (c) důkazu věty 36, neboť se v těchto částech nepoužívá výraz $p_i^T B p_i$. □

Věta 38 má velký význam pro vyšetřování nepřesných metod s lokálně omezeným krokem (oddíl 6.3). Používáme-li metodu sdružených gradientů pro výpočet směrového vektoru v metodách spádových směrů, mohou nastat problémy. Jednak může být $p_i^T B p_i \approx 0$, což vede k selhání metody ($\alpha_i \approx \infty$ a $\|s_{i+1}\| \approx \infty$), nebo platí $p_i^T B p_i < 0$, takže stacionární bod získaný metodou sdružených gradientů není minimem funkce $Q(s)$. Proto je třeba výpočet ukončit pokud neplatí $p_i^T B p_i \geq \underline{c}\, p_i^T p_i$, kde $\underline{c}$ je zvolená dolní mez. Jelikož tato podmínka nemusí být splněna pro všechny indexy $1 \leq i \leq m$, nemůžeme použít nerovnost (97). Platí však tato věta

Věta 39 *Aplikujeme-li předpodmíněnou metodu sdružených gradientů s pozitivně definitní maticí C na kvadratickou funkci $Q(s)$ a platí-li $p_j^T B p_j \geq \underline{c}\, p_j^T p_j$ pro $1 \leq j \leq i$, pak*

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{\underline{c}}{n\kappa(C)\|B\|}.$$

Důkaz Použijeme-li vztahy (85) a (87), dostaneme

$$g_j^T C^{-1} g_j = g_j^T(-p_j + \beta_{j-1} p_{j-1}) = -g_j^T p_j = -\left(g + \sum_{k=1}^{j-1} \alpha_k B p_k\right)^T p_j = -g^T p_j$$

pro $1 \leq j \leq i$, takže

$$\begin{aligned} -g^T s_{i+1} &= -\sum_{j=1}^{i} \alpha_j g^T p_j = \sum_{j=1}^{i} \alpha_j g_j^T C^{-1} g_j \geq \alpha_1 g_1^T C^{-1} g_1 \\ &= \frac{(g_1^T C^{-1} g_1)^2}{p_1^T B p_1} = \frac{p_1^T C p_1}{p_1^T p_1} \frac{p_1^T p_1}{p_1^T B p_1} g^T C^{-1} g \geq \frac{\|g\|^2}{\kappa(C)\|B\|}. \end{aligned}$$

Dále platí

$$s_{i+1} = \sum_{j=1}^{i} \alpha_j p_j = -\sum_{j=1}^{i} \frac{g^T p_j}{p_j^T B p_j} p_j = -\sum_{j=1}^{i} \frac{p_j p_j^T}{p_j^T B p_j} g,$$

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 obě nerovnosti, 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 79 Je-li matice B pozitivně definitní, můžeme položit $\underline{c}/\|B\| = 1/\kappa(B)$, takže dostaneme

$$-\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} \geq \frac{1}{n\kappa(B)\kappa(C)}.$$

Tento odhad je horší než odhad (97) uvedený v důsledku 3.

Algoritmus předpodmíněné metody sdružených gradientů pro výpočet směrových vektorů v metodách spádových směrů lze popsat zhruba takto:

Algoritmus 3 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 $\tau < \underline{c}\|p\|$, 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 80 Platí-li v každém iteračním kroku $\|B\| \leq \overline{B}$ a $\kappa(C) \leq \overline{\kappa}$, kde $\overline{B}$ a $\overline{\kappa}$ jsou konstanty společné všem iteračním krokům, pak směrový vektor s vypočtený algoritmem 3, splňuje podle věty 39 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 3 lze tedy použít k realizaci globálně konvergentní modifikované Newtonovy metody spádových směrů.

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.

4.1 Základní vlastnosti metod s proměnnou metrikou

Definice 23 *Řekneme, že metoda spádových směrů (definice 15) je metodou s proměnnou metrikou, jestliže*

$$s_i = -H_i g_i, \tag{102}$$

kde H_i, $i \in N$, jsou symetrické pozitivně definitní matice konstruované podle rekurentního vztahu

$$H_{i+1} = \gamma_i (H_i + U_i M_i U_i^T), \tag{103}$$

kde $U_i \in R^{n\times 2}$, $M_i \in R^{2\times 2}$ a $\gamma_i > 0$, a vyhovující podmínce

$$H_{i+1} y_i = \rho_i d_i, \tag{104}$$

kde $y_i = g_{i+1} - g_i$, $d_i = x_{i+1} - x_i = \alpha_i s_i$ a $\rho_i > 0$.

Poznámka 81 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 (104) se nazývá (zobecněnou) kvazinewtonovskou podmínkou (předpokládáme, že $d_i \neq 0$ a $y_i \neq 0$, neboť v opačném případě nemá podmínka (104) smysl). Původní metody s proměnnou metrikou byly navrženy s hodnotami $\rho_i = 1$ a $\gamma_i = 1$ (bez škálování a korekce). Jelikož efektivní škálování a vhodná korekce zlepšují účinnost metod s proměnnou metrikou, budeme vyšetřovat obecný případ, kdy $\rho_i > 0$ a $\gamma_i > 0$.

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 40 *(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*

$$Q(x) = \frac{1}{2}(x - x^*)^T G (x - x^*).$$

Nechť $g_i \neq 0$, $1 \leq i \leq n$. Pak $g_{n+1} = 0$ a $x_{n+1} = x^*$.

Důkaz Předpokládejme, že $g_i \neq 0$, $\leq i \leq n$. Dokážeme indukcí, že $s_i \neq 0$, $\alpha_i \neq 0$, $1 \leq i \leq n$, a že pro $1 \leq j < i \leq n+1$ platí

$$H_i y_j = \lambda_i^j d_j, \tag{105}$$

$$s_j^T g_i = 0, \tag{106}$$

$$s_j^T G s_i = 0, \tag{107}$$

$$s_j^T y_i = y_j^T s_i = 0, \tag{108}$$

kde $\lambda_i^j > 0$, $1 \leq j < i \leq n+1$, (viz (109)). Rovnosti (107) a (108) jsou ekvivalentní, neboť pro kvadratickou funkci $Q(x)$ platí $y_i = g_{i+1} - g_i = G(x_{i+1} - x_i) = G d_i = \alpha_i G s_i$ a $\alpha_i \neq 0$ podle indukčního předpokladu. Z (102) a (107) plyne, že s_i, $1 \leq i \leq n$, jsou nenulové a vzájemně sdružené (G-ortogonální), tudíž lineárně

nezávislé, takže podle (106) nutně $g_{n+1} = 0$. Pro $i = 1$ platí $s_1^T g_1 = -g_1^T H_1 g_1 < 0$ (H_1 je symetrická pozitivně definití) takže $s_1 \neq 0$ a $\alpha_1 \neq 0$ a dále není co dokazovat. Indukční krok:

(a) Nechť $i \leq n$. Z (108) plyne $d_i^T y_j = \alpha_i s_i^T y_j = 0$ a (105) navíc dává $y_i^T H_i y_j = \lambda_i^j y_i^T d_j = 0$, takže $U_i^T y_j = 0$, $1 \leq j < i$. Podle (103) a (105) tedy platí

$$H_{i+1} y_j = \gamma_i (H_i y_j + U_i^T M_i U_i^T y_j) = \gamma_i H_i y_j = \gamma_i \lambda_i^j d_j \stackrel{\Delta}{=} \lambda_{i+1}^j d_j,$$

$1 \leq j < i$, a použijeme-li (104) dostaneme $H_{i+1} y_i = \rho_i d_i \stackrel{\Delta}{=} \lambda_{i+1}^i d_i$. Platí tedy $H_{i+1} y_j = \lambda_{i+1}^j d_i$, $1 \leq j \leq i$.

(b) Nechť $i \leq n$. Z (106) a (108) plyne $s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = 0$, $1 \leq j < i$. Z přesného výběru délky kroku dostaneme $s_i^T g_{i+1} = 0$, takže celkem $s_j^T g_{i+1} = 0$, $1 \leq j \leq i$.

(c) Podle (102) 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 $\alpha_{i+1} \neq 0$. Použijeme-li (102), (a), (b) dostaneme

$$y_j^T s_{i+1} = -y_j^T H_{i+1} g_{i+1} = -\frac{\lambda_{i+1}^j}{\alpha_j} d_j^T g_{i+1} = -\lambda_{i+1}^j s_j^T g_{i+1} = 0,$$

$1 \leq j \leq i$. □

Poznámka 82 Může se stát, že $g_m = 0$ pro nějaký index $m < n$. V tom případě sice nejsou splněny předpoklady věty 40, ale získáme minimum kvadratické funkce $Q(x)$ již v m-tém iteračním kroku.

Poznámka 83 Z části (a) důkazu věty 40 plyne, že pro $1 \leq j < i \leq n+1$ platí

$$\lambda_i^j = \left(\prod_{k=j}^{i-1} \gamma_k \right) \frac{\rho_j}{\gamma_j}. \tag{109}$$

Důsledek 4 *Nechť jsou splněny předpoklady věty 40. Pak pro $1 \leq i \leq n$ tvoří vektory s_i a s_{i+1} bázi v $\mathcal{L}(U_i)$ a pro $1 \leq i < j \leq n+1$ platí*

$$H_i g_j = \left(\prod_{k=1}^{i-1} \gamma_k \right) H_1 g_j. \tag{110}$$

Důkaz (a) Nechť 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

$$v^T s_{i+1} = -v^T H_{i+1} g_{i+1} = -\gamma_i v^T (H_i g_i + H_i y_i) = \gamma_i v^T (s_i - H_i y_i) = 0,$$

takže vektory s_i a s_{i+1} leží v $\mathcal{L}(U_i)$, a jelikož podle (107) jsou tyto vektory lineárně nezávislé, tvoří bázi v $\mathcal{L}(U_i)$.

(b) Podle (106) platí $s_k^T g_j = 0$ a $s_{k+1}^T g_j = 0$ $\forall 1 \leq k < i < j$, což podle (a) dává $U_k^T g_j = 0$ $\forall 1 \leq k < i < j$. Několikanásobným použitím (103) pak dostaneme

$$H_i g_j = \gamma_{i-1} H_{i-1} g_j = \ldots = \left(\prod_{k=1}^{i-1} \gamma_k \right) H_1 g_j.$$

□

Ukážeme nyní, že jsou-li splněny předpoklady věty 40, generují všechny metody s proměnnou metrikou z Broydenovy třídy stejnou posloupnost bodů x_i, $i \in N$.

Věta 41 *Nechť jsou splněny předpoklady věty 40. Pak všechny metody s proměnnou metrikou z Broydenovy třídy generují stejnou posloupnost bodů x_i, $i \in N$.*

Důkaz Protože 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 paralelní. 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 (103) a (104)). 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 4 leží vektor s_{i+1} v $\mathcal{L}(U_i)$ a podle (108) 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 (105) a (108), dostaneme $y_j^T H_{i+1} y_{i+1} = \alpha_j \lambda_{i+1}^j s_j^T y_{i+1} = 0$ 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

$$y_j^T \left(\bar{H}_{i+1} - \lambda_i H_{i+1} \right) y_{i+1} = 0 \qquad \forall 1 \leq j < i+1 \tag{111}$$

pro libovolné číslo $\lambda_i > 0$. Nechť $i+1 < j \leq n$. Použijeme-li (110), 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 $\mathcal{L}(\bar{U}_j) = \mathcal{L}(U_j)$ $\forall 1 \leq j \leq i$, platí také

$$\bar{H}_{i+1} y_j = \left(\prod_{k=1}^{i} \bar{\gamma}_k \right) H_1 y_j \triangleq \bar{\omega}_i H_1 y_j,$$

což spolu s předchozí rovností dává

$$y_j^T \left(\bar{H}_{i+1} - \lambda_i H_{i+1} \right) y_{i+1} = 0 \qquad \forall i+1 < j \leq n \tag{112}$$

pro $\lambda_i = \omega_i / \bar{\omega}_i$. Vektory y_j, $1 \leq j \leq n$ jsou lineárně nezávislé a podle (108) 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 (111) a (112), 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 5 *Nechť jsou splněny předpoklady věty 40. 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 41 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 (118)). 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 (110) 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 50). Protože 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 □

Velmi výhodné je volit parametry γ_i a ρ_i tak, že $\gamma_i = 1$ a $\rho_i = 1$ $\forall i \in N$.

Věta 42 *(Aproximace Hessovy matice). Nechť jsou splněny předpoklady věty 40 s $\gamma_i = 1$ a $\rho_i = 1$ $\forall i \in N$. Pak platí $H_{n+1} = G^{-1}$.*

Důkaz Jestliže $\gamma_i = 1$ a $\rho_i = 1$ $\forall i \in N$, platí podle (109) $\lambda_i^j = 1$, $1 \le j < i \le n+1$. 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 (107) lineárně nezávislé, musí platit $H_{n+1} = G^{-1}$. □

V předchozích větách jsme využívali toho, že minimalizovaná funkce je kvadratická. Překvapivě se dá dokázat, 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 43 *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 (102) 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 4 se ukáže, že vektor s_{i+1} leží v $\mathcal{L}(U_i)$. Navíc podle (103) a (104) 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 41, 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 (103) a (104), dostaneme

$$H_{i+1} v = H_{i+1} w + \lambda H_{i+1} y_i = \gamma_i H_i w + \lambda \rho_i d_i = \bar{\gamma}_i \bar{H}_i w + \lambda \bar{\rho}_i \bar{d}_i = \bar{H}_{i+1} w + \lambda \bar{H}_{i+1} \bar{y}_i = \bar{H}_{i+1} v, \qquad (113)$$

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 v 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 aktualizace (103). Pro zjednodušení budeme index i vynechávat a index $i+1$ nahradíme symbolem $+$. Nejprve uvedeme několik pomocných tvrzení týkajících se aktualizací.

Lemma 10 *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) Matice $I + UV^T$ má stejná nejednotková vlastní čísla jako matice $I + V^TU$.

(c) Platí $\det(I + UV^T) = \det(I + V^TU)$.

(d) 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) Zřejmě $(I + UV^T)x = \lambda x$ právě tehdy, jestliže $UV^Tx = (\lambda - 1)x$, a $(I + V^TU)y = \lambda y$ právě tehdy, jestliže $V^TUy = (\lambda - 1)y$. Tvrzení (b) tedy plyne z (a).
(c) Determinant matice je roven součinu jejích vlastních čísel. Tvrzení (c) tedy plyne z (b).
(d) Je-li matice $I + UV^T$ regulární, je podle (b) 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 10 má několik důležitých důsledků.

Důsledek 6 *(Woodbury) Nechť jsou splněny předpoklady lemmatu 10 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 (c) a (d) z lemmatu 10 (matice U se nahradí maticí $H^{-1}U$). □

Poznámka 84 (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 7 *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) *Matice* $I + UMU^T$ *má stejná nejednotková vlastní čísla jako matice* $I + MU^TU$ *(nebo jako matice* $I + U^TUM$*).*

(c) *Platí* $\det(I + UMU^T) = \det(I + MU^TU) = \det(I + U^TUM)$. *Je-li matice* M *regulární, platí* $\det(I + UMU^T) = \det M \ \det(M^{-1} + U^TU)$.

(d) *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 10 použít UM místo V (nebo UM místo U a U místo V). Tvrzení (c) a (d) jsou jednodušší verzí důsledku 8. □

Důsledek 8 *Nechť jsou splněny předpoklady důsledku 7 a nechť* $H \in R^{n\times n}$ *je regulární symetrická matice. Pak*

(a) *Platí* $\det(H + UMU^T) = \det H \ \det(I + MU^TH^{-1}U) = \det H \ \det(I + U^TH^{-1}UM)$. *Je-li matice* M *regulární, platí* $\det(H + UMU^T) = \det H \ \det M \ \det(M^{-1} + U^TH^{-1}U)$.

(b) *Jsou-li matice* M *a* $H + UMU^T$ *regulární, platí*

$$(H + UMU^T)^{-1} = H^{-1} - H^{-1}U(M^{-1} + U^TH^{-1}U)^{-1}U^TH^{-1}.$$

Důkaz Označme $V = H^{-1}UM$. Jelikož $H + UMU^T = (I + UMU^TH^{-1})H = (I + UV^T)H$, můžene podle lemmatu 10 psát

$$\begin{aligned}
\det(H + UMU^T) &= \det H \ \det(I + UV^T) = \det H \ \det(I + V^TU) \\
&= \det H \ \det(I + MU^TH^{-1}U) \\
(H + UMU^T)^{-1} &= H^{-1}(I + UV^T)^{-1} = H^{-1}(I - U(I + V^TU)^{-1}V^T) \\
&= H^{-1} - H^{-1}U(I + MU^TH^{-1}U)^{-1}MU^TH^{-1}.
\end{aligned}$$

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 44 *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*

$$M = \begin{bmatrix} \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right), & -\frac{\eta}{b} \\ -\frac{\eta}{b}, & \frac{\eta - 1}{a} \end{bmatrix},$$

kde η *je volný parametr a kde*

$$a = y^THy, \qquad b = y^Td, \qquad c = d^TH^{-1}d.$$

Důkaz Označme m_1, m_2, m_3 prvky matice M. Pak podle (103) a (104) platí

$$\begin{aligned}\frac{1}{\gamma}H_+y &= Hy + [d, Hy]\begin{bmatrix} m_1, & m_2 \\ m_2, & m_3 \end{bmatrix}\begin{bmatrix} b \\ a \end{bmatrix} \\ &= Hy + (m_1b + m_2a)d + (m_2b + m_3a)Hy = \frac{\rho}{\gamma}d,\end{aligned}$$

takže nutně

$$\begin{aligned}m_1b + m_2a &= \rho/\gamma, \\ m_2b + m_3a &= -1.\end{aligned}$$

Jeden parametr je nadbytečný. Zvolíme $m_2 = -\eta/b$ a zbylé prvky m_1, m_3 určíme řešením uvedených rovnic, takže

$$m_1 = \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right), \qquad m_2 = -\frac{\eta}{b}, \qquad m_3 = \frac{\eta - 1}{a}. \tag{114}$$

Tím dostaneme matici M uvedenou ve větě 44. □

Poznámka 85 Při vyšetřování metod s proměnnou metrikou budeme často používat označení

$$\mu = \frac{1}{ab}\left(\eta\frac{a}{b} + (1 - \eta)\frac{\rho}{\gamma}\right). \tag{115}$$

Přímým výpočtem se snadno přesvědčíme, že $\mu = -\det M$, kde M je matice vystupující ve větě 44.

Poznámka 86 Číslo $c = d^TH^{-1}d$ se v matici M ani ve vzorci (115) 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 $c = -\alpha d^Tg$ a není třeba řešit soustavu rovnic s maticí H.

Poznámka 87 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

$$b = y^Td = \alpha(g_+ - g)^Ts \geq \alpha(\varepsilon_2 - 1)g^Ts > 0.$$

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 88 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 (116) lze zapsat ve tvaru

$$\frac{1}{\gamma}H_+ = H + \left(\frac{\rho}{\gamma} - \frac{a}{b}\right)dd^T.$$

Zvolíme-li $\rho/\gamma = a/b$, lze kvazinewtonovskou podmínku $H_+y = \rho d$ splnit prostým vynásobením matice H číslem $\gamma = \rho\, b/a$.

Poznámka 89 Vztah $H_+ = \gamma(H + UMU^T)$ můžeme roznásobit. Pak platí

$$\frac{1}{\gamma}H_+ = H + \frac{\rho}{\gamma b}dd^T - \frac{1}{a}Hy(Hy)^T + \frac{\eta}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \tag{116}$$

(Broydenova třída). Nejznámější členy Broydenovy třídy dostaneme, položíme-li $\eta = \eta^{DFP} = 0$ (metoda Davidona, Fletchera a Powella), takže

$$\frac{1}{\gamma}H_+^{DFP} = H + \frac{\rho}{\gamma b}dd^T - \frac{1}{a}Hy(Hy)^T, \tag{117}$$

nebo $\eta = \eta^{BFGS} = 1$ (metoda Broydena, Fletchera, Goldfarba a Shanno), takže

$$\frac{1}{\gamma}H_+^{BFGS} = H + \left(\frac{a}{b} + \frac{\rho}{\gamma}\right)\frac{1}{b}dd^T - \frac{1}{b}\left(Hyd^T + d(Hy)^T\right), \tag{118}$$

nebo $\eta = \eta^{R1} = (\rho/\gamma)/(\rho/\gamma - a/b)$ (metoda hodnosti 1), takže

$$\frac{1}{\gamma}H_+^{R1} = H + \frac{1}{(\rho/\gamma)b - a}\left(\frac{\rho}{\gamma}d - Hy\right)\left(\frac{\rho}{\gamma}d - Hy\right)^T, \tag{119}$$

nebo $\eta = \eta^{H} = (\rho/\gamma)/(\rho/\gamma + a/b)$ (Hoshinova metoda), takže

$$\frac{1}{\gamma}H_+^{H} = H + \frac{2\rho}{\gamma b}dd^T - \frac{1}{(\rho/\gamma)b + a}\left(\frac{\rho}{\gamma}d + Hy\right)\left(\frac{\rho}{\gamma}d + Hy\right)^T. \tag{120}$$

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 podminku pro pozitivní definitnost matice H_+ (zdůvodnění je uvedeno v poznámce 93), takže může dojít ke ztrátě globální konvergence vlivem porušení podmínky spádovosti (S1a).

Poznámka 90 Pro konstrukci metod s omezenou pamětí je užitečný pseudosoučinový tvar

$$\frac{1}{\gamma}H_+ = \left(I - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)y^T\right)H\left(I - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)y^T\right)^T + \frac{\rho}{\gamma b}dd^T \tag{121}$$

který lze použít pouze tehdy, když $\eta \geq 0$. Tento vzorec se velmi zjednoduší pro metodu BFGS, kdy $\eta = 1$. Platí také

$$\frac{1}{\gamma}H_+ = \left(I - \frac{1}{b}dy^T\right)\left(H + \frac{\eta - 1}{a}Hy(Hy)^T\right)\left(I - \frac{1}{b}yd^T\right) + \frac{\rho}{\gamma b}dd^T. \tag{122}$$

Vzorec (116) lze zapsat ve tvaru

$$\begin{aligned}
\frac{1}{\gamma}H_+ &= \frac{1}{\gamma}H_+^{DFP} + \frac{\eta}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \\
&= \frac{1}{\gamma}H_+^{BFGS} + \frac{\eta - 1}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T \\
&= \left(I - \frac{1}{b}dy^T\right)H\left(I - \frac{1}{b}yd^T\right) + \frac{\rho}{\gamma b}dd^T + \frac{\eta - 1}{a}\left(\frac{a}{b}d - Hy\right)\left(\frac{a}{b}d - Hy\right)^T,
\end{aligned} \tag{123}$$

kde první část posledního řádku je pseudosoučinový tvar pro metodu BFGS a druhá část plyne ze vzorce (116). Platí také

$$\begin{aligned}
\frac{1}{\gamma}H_+ &= H - \frac{\mu}{m_3}dd^T + m_3\left(\frac{m_2}{m_3}d + Hy\right)\left(\frac{m_2}{m_3}d + Hy\right)^T \\
&= H + \frac{\mu a}{1-\eta}dd^T - \frac{1-\eta}{a}\left(\frac{\eta a}{(1-\eta)b}d + Hy\right)\left(\frac{\eta a}{(1-\eta)b}d + Hy\right)^T,
\end{aligned} \tag{124}$$

nebo

$$\frac{1}{\gamma}H_+ = H + m_1\left(d + \frac{m_2}{m_1}Hy\right)\left(d + \frac{m_2}{m_1}Hy\right)^T - \frac{\mu}{m_1}Hy(Hy)^T$$

$$= H + \frac{1}{(\rho/\gamma)b + \eta a}\left(\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right)d - \eta Hy\right)\left(\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right)d - \eta Hy\right)^T - \frac{\mu b^2}{(\rho/\gamma)b + \eta a}Hy(Hy)^T. \quad (125)$$

kde m_1, m_2, m_3 jsou čísla určená vztahy (114). 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 (116) vyžaduje zhruba $3n^2$ aritmetických operací. Pomocí vzorce (125) lze aktualizaci metody BFGS vyjádřit ve tvaru

$$\frac{1}{\gamma}H_+^{BFGS} = H + \frac{1}{(\rho/\gamma)b + a}\left(\left(\frac{a}{b} + \frac{\rho}{\gamma}\right)d - Hy\right)\left(\left(\frac{a}{b} + \frac{\rho}{\gamma}\right)d - Hy\right)^T - \frac{1}{(\rho/\gamma)b + a}Hy(Hy)^T, \quad (126)$$

Teoretický význam má též vzorec

$$\frac{1}{\gamma}H_+^{BFGS} = H + \frac{1}{b}\left(d\left(\frac{\rho}{\gamma}d - Hy\right)^T + \left(\frac{\rho}{\gamma}d - Hy\right)d^T\right) - \left(\frac{\rho}{\gamma} - \frac{a}{b}\right)\frac{1}{b}dd^T, \quad (127)$$

který lze získat variačním odvozením (poznámka 116). Aktualizaci metody DFP lze vyjádřit v součinovém tvaru

$$\frac{1}{\gamma}H_+^{DFP} = \left(I - \frac{1}{a}\left(Hy \pm \sqrt{\frac{\rho a}{\gamma b}}d\right)y^T\right)H\left(I - \frac{1}{a}\left(Hy \pm \sqrt{\frac{\rho a}{\gamma b}}d\right)y^T\right)^T. \quad (128)$$

O správnosti všech těchto vztahů se můžeme přesvědčit jejich roznásobením a porovnáním odpovídajících si členů.

Lemma 11 *Nechť H je symetrická pozitivně definití matice, $a > 0$, $b > 0$, $c > 0$ a nechť H_+ je matice získaná pomocí aktualizace (116), 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

$$\sigma = \frac{1}{b^2}(\eta(ac - b^2) + b^2) + \frac{\rho}{\gamma}\frac{c}{b} = \frac{\rho}{\gamma}\frac{c}{b} + \left(1 - \frac{\eta}{\eta^*}\right), \quad (129)$$

$$\delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac - b^2) + b^2) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right) \quad (130)$$

a kde

$$\eta^* = -\frac{b^2}{ac - b^2} < 0$$

je kritická hodnota parametru η (pokud $ac - b^2 = 0$, můžeme položit $1/\eta^* = 0$ a $\eta^* = -\infty$).

Důkaz Podle (103) platí

$$\frac{1}{\gamma}H^{-1/2}H_+H^{-1/2} = I + H^{-1/2}UMU^TH^{-1/2}. \quad (131)$$

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 7 vlastními čísly matice $I + MU^TH^{-1}U$, takže pro ně musí platit $\det((1-\lambda)I + MU^TH^{-1}U) = 0$. Použijeme-li pro M vztah uvedený ve větě 44 a pro $U^TH^{-1}U$ vztah

$$U^T H^{-1} U = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix},$$

můžeme psát

$$\begin{aligned} \det((1-\lambda)I + MU^T H^{-1} U) &= \det\left(\begin{bmatrix} 1-\lambda, & 0 \\ 0, & 1-\lambda \end{bmatrix} + M \begin{bmatrix} c, & b \\ b, & a \end{bmatrix} \right) \\ &= \det \begin{bmatrix} \eta \dfrac{ac-b^2}{b^2} + \dfrac{\rho c}{\gamma b} + 1 - \lambda, & \dfrac{\rho}{\gamma} \\ -\eta \dfrac{ac-b^2}{ab} - \dfrac{b}{a}, & -\lambda \end{bmatrix} = 0, \end{aligned}$$

což po úpravě dává $\lambda^2 - \sigma\lambda + \delta = 0$ s koeficienty uvedenými v lemmatu 11. □

Poznámka 91 Poznamenejme, že δ se jako součin vlastních čísel matice $I + MU^T H^{-1} U$ podle důsledku 7 rovná determinantu matice $(1/\gamma)H^{-1/2} H_+ H^{-1/2} = I + H^{-1/2} U M U^T H^{-1/2}$. Platí tedy $\det((1/\gamma)H_+) = \delta \det H$, což lze zapsat ve tvaru.

$$\det\left(\frac{1}{\gamma} H_+\right) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right) \det H \tag{132}$$

(pokud $ac - b^2 = 0$, můžeme položit $1/\eta^* = 0$).

Věta 45 *Nechť jsou splněny předpoklady lemmatu 11. 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 11 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

$$\sigma = \frac{\gamma}{\rho}\frac{a}{b}\delta + \frac{\rho}{\gamma}\frac{c}{b}.$$

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 92 Ve větě 45 předpokládáme, že $b > 0$. Pokud $b = 0$, není matice H_+ definována. Pokud $b < 0$ a $\delta > 0$, plyne z důkazu věty 45, že $\sigma < 0$, takže matice H_+ není pozitivně definitní. Podmínka $b > 0$ je tedy pro pozitivní definitnost nutná.

Poznámka 93 Z věty 45 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 minimalizaci s nelineárními omezeními, je velmi důležitý inverzní tvar rekurentního vztahu (116).

Věta 46 *(Aktualizace matice $B = H^{-1}$). Nechť jsou splněny předpoklady lemmatu 11. Nechť $B = H^{-1}$ a $B_+ = H_+^{-1}$. Pak platí*

$$\gamma B_+ = B + \frac{\gamma}{\rho b} y y^T - \frac{1}{c} Bd(Bd)^T + \frac{\beta}{c}\left(\frac{c}{b}y - Bd\right)\left(\frac{c}{b}y - Bd\right)^T, \tag{133}$$

kde

$$\beta\eta(ac - b^2) + (\beta + \eta)b^2 = b^2. \tag{134}$$

Důkaz Inverzí vztahu $(1/\gamma)H_+ = H + UMU^T$ podle důsledku 8 dostaneme

$$\gamma B_+ = B - BU(M^{-1} + U^T BU)^{-1}U^T B \triangleq B + BUKU^T B,$$

kde $K \in R^{2\times 2}$. Jelikož podle (104) 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

$$K = \begin{bmatrix} \frac{\beta - 1}{c}, & -\frac{\beta}{b} \\ -\frac{\beta}{b}, & \frac{1}{b}\left(\beta\frac{c}{b} + \frac{\gamma}{\rho}\right) \end{bmatrix},$$

což po dasazení do $\gamma B_+ = B + BUKU^T B$ dává (133). Vztah (134), svazující β s η lze získat například z rovnosti

$$K = -(M^{-1} + U^T BU)^{-1}.$$

Jednodušší způsob je uveden v poznámce 95. □

Poznámka 94 (Dualita) Vztah (133) dostaneme ze vztahu (116) 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

$$\gamma B_+^{DFP} = B + \left(\frac{c}{b} + \frac{\gamma}{\rho}\right)\frac{1}{b}yy^T - \frac{1}{b}\left(Bdy^T + y(Bd)^T\right). \tag{135}$$

Metodu BFGS dostaneme pro $\beta = \beta^{BFGS} = 0$, takže

$$\gamma B_+^{BFGS} = B + \frac{\gamma}{\rho b}yy^T - \frac{1}{c}Bd(Bd)^T. \tag{136}$$

Metoda hodnosti 1 je samoduální, dostaneme ji pro $\beta = \beta^{R1} = (\gamma/\rho)/(\gamma/\rho - c/b)$, takže

$$\gamma B_+^{R1} = B + \frac{1}{(\gamma/\rho)b - c}\left(\frac{\gamma}{\rho}y - Bd\right)\left(\frac{\gamma}{\rho}y - Bd\right)^T. \tag{137}$$

Hoshinova metoda je také samoduální, dostaneme ji pro $\beta = \beta^H = (\gamma/\rho)/(\gamma/\rho + c/b)$, takže

$$\gamma B_+^H = B + \frac{2\gamma}{\rho b}yy^T - \frac{1}{(\gamma/\rho)b + c}\left(\frac{\gamma}{\rho}y + Bd\right)\left(\frac{\gamma}{\rho}y + Bd\right)^T. \tag{138}$$

Poznámka 95 Z duality lze snadno určit vztah mezi β a η. Platí totiž

$$\det(\gamma B_+) = \frac{\gamma}{\rho}\frac{b}{c}\left(1 - \frac{\beta}{\beta^*}\right)\det B, \tag{139}$$

což spolu s výrazem (132) pro $\det H_+$ a s identitami $\det B \det H = 1$, $\det B_+ \det H_+ = 1$ dává

$$\frac{b^2}{ac}\left(1-\frac{\beta}{\beta^*}\right)\left(1-\frac{\eta}{\eta^*}\right)=1. \tag{140}$$

Po dosazení, roznásobení a úpravě tohoto vztahu dostaneme rovnost (134). Z vyjádření (139) vyplývá, že matice B_+ je pozitivně definitní právě tehdy, jestliže $\beta > \beta^*$, kde

$$\beta^* = \eta^* = -\frac{b^2}{ac-b^2} < 0. \tag{141}$$

Poznámka 96 Ze vztahů (134) a (140)–(141) 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 $\beta = \beta^*$. Podle (140)–(141) pro $\eta \leq \overline{\eta}$, $\overline{\eta} \geq 1$, platí

$$1-\frac{\beta}{\beta^*} = \frac{ac}{b^2}\left(1-\frac{\eta}{\eta^*}\right)^{-1} = \frac{ac}{\eta(ac-b^2)+b^2} \geq \frac{ac}{\overline{\eta}(ac-b^2)+b^2} = \frac{ac}{\overline{\eta}ac+b^2(1-\overline{\eta})} \geq \frac{1}{\overline{\eta}}. \tag{142}$$

Poznámka 97 Z úvahy použité v poznámce 95 plyne, že při přechodu od vztahu (116) ke vztahu (133) provádíme záměnu $\delta \to 1/\delta$. Z důkazu věty 46 víme, že $-K^{-1} = M^{-1} + U^T B U = M^{-1}(I + MU^T BU)$, což podle lemmatu 11 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 (116) ke vztahu (133) provádíme záměnu $\mu \to \mu/\delta$. Z duality plyne, že

$$\frac{1}{\delta} = \frac{\gamma}{\rho}\frac{1}{bc}(\beta(ac-b^2)+b^2) = \frac{\gamma}{\rho}\frac{b}{c}\left(1-\frac{\beta}{\beta^*}\right), \tag{143}$$

$$\frac{\mu}{\delta} = \frac{1}{bc}\left(\beta\frac{c}{b}+(1-\beta)\frac{\gamma}{\rho}\right). \tag{144}$$

Jelikož vlastní čísla matice $\gamma B^{-1/2} B_+ B^{-1/2}$ jsou převrácenými hodnotami vlastních čísel matice (131), jsou to buď jednotky nebo řešení kvadratické rovnice

$$\lambda^2 - \frac{\sigma}{\delta}\lambda + \frac{1}{\delta} = 0.$$

Z duality plyne, že

$$\frac{\sigma}{\delta} = \frac{1}{b^2}(\eta(ac-b^2)+b^2) + \frac{\gamma}{\rho}\frac{a}{b} = \frac{\gamma}{\rho}\frac{a}{b} + \left(1-\frac{\beta}{\beta^*}\right). \tag{145}$$

Poznámka 98 Jelikož $Bs = -g$ (poznámka 86), lze ve vztahu (133) nahradit vektor Bd vektorem $-\alpha g$, takže odpadne maticové násobení.

Poznámka 99 V následující tabulce jsou uvedeny hodnoty parametrů nejznámějších metod s proměnnou metrikou z Broydenovu třídy.

	DFP	BFGS	R1	H
η	0	1	$\frac{\rho b}{\rho b - \gamma a}$	$\frac{\rho b}{\rho b + \gamma a}$
δ	$\frac{\rho b}{\gamma a}$	$\frac{\rho c}{\gamma b}$	$\frac{\rho}{\gamma}\frac{\rho c - \gamma b}{\rho b - \gamma a}$	$\frac{\rho}{\gamma}\frac{\rho c + \gamma b}{\rho b + \gamma a}$
μ	$\frac{\rho}{\gamma a b}$	$\frac{1}{b^2}$	0	$\frac{2}{b^2}\frac{\rho b}{\rho b + \gamma a}$
β	1	0	$\frac{\gamma b}{\gamma b - \rho c}$	$\frac{\gamma b}{\gamma b + \rho c}$
$\frac{1}{\delta}$	$\frac{\gamma a}{\rho b}$	$\frac{\gamma a}{\rho c}$	$\frac{\gamma}{\rho}\frac{\gamma a - \rho b}{\gamma b - \rho c}$	$\frac{\gamma}{\rho}\frac{\gamma a + \rho b}{\gamma b + \rho c}$
$\frac{\mu}{\delta}$	$\frac{1}{b^2}$	$\frac{\gamma}{\rho b c}$	0	$\frac{2}{b^2}\frac{\gamma b}{\gamma b + \rho c}$

4.2 Součinový tvar metod s proměnnou metrikou

Nyní budeme vyšetřovat součinový tvar metod s proměnnou metrikou. Budeme 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. Matice S_+ se určuje pomocí aktualizace

$$\frac{1}{\sqrt{\gamma}}S_+ = S + p\tilde{q}^T, \tag{146}$$

kde $p \in R^n$ a $\tilde{q} \in R^m$. Součinový tvar metod s proměnnou metrikou se používá zejména tehdy, když $m < n$ a to buď při minimalizaci na lineární varietě rovnoběžné s podprostorem $\mathcal{L}(S)$ dimenze $m < n$, nebo v případě metod s omezenou pamětí, kdy uchováváme pouze $m < n$ vektorů, tvořících sloupce matice S. Používáme-li metody spádových směrů, platí $s = -Hg = -SS^Tg$, takže $d = \alpha s \in \mathcal{L}(S)$. Budeme tedy 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 100 Pokud $H = SS^T$, počítáme směrový vektor $s = -Hg$ podle vzorců

$$\tilde{s} = -S^Tg, \qquad s = S\tilde{s}.$$

Pak $\tilde{d} = \alpha\tilde{s}$ a $\tilde{y} = S^Ty$.

Poznámka 101 Pokud $p \in \mathcal{L}(S)$ (což předpokládáme), lze vzorec (146) zapsat dvojím způsobem, buď

$$\frac{1}{\sqrt{\gamma}}S_+ = S(I + \tilde{p}\tilde{q}^T), \tag{147}$$

nebo

$$\frac{1}{\sqrt{\gamma}}S_+ = (I + pq^T)S. \tag{148}$$

Pokud ale $p \notin \mathcal{L}(S)$ (jako u posunutých metod s proměnnou metrikou popsaných v oddílu 8.3) 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 24 *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 47 *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

$$MM^\dagger M = UV^TV(V^TV)^{-1}(U^TU)^{-1}U^TUV^T = UV^T = M.$$

$$M^\dagger MM^\dagger = V(V^TV)^{-1}(U^TU)^{-1}U^TUV^TV(V^TV)^{-1}(U^TU)^{-1}U^T = V(V^TV)^{-1}(U^TU)^{-1}U^T = M^\dagger.$$

(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 102 Má-li matice $S \in R^{n\times m}$, $m \leq n$, plnou hodnost, lze podle definice 24 snadno ověřit, že

$$S^\dagger = (S^TS)^{-1}S^T, \qquad S^\dagger S = I \tag{149}$$

$$(SS^T)^\dagger = (S^\dagger)^TS^\dagger = S(S^TS)^{-2}S^T, \qquad (S^TS)^{-1} = S^\dagger(S^\dagger)^T. \tag{150}$$

Poznámka 103 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 $S \in R^{n\times m}$, $m \leq n$, plnou hodnost, lze podle definice 24 snadno ověřit, že

$$(SS^T)^{1/2} = S(S^TS)^{-1/2}S^T, \qquad ((SS^T)^\dagger)^{1/2} = S(S^TS)^{-3/2}S^T. \tag{151}$$

Věta 48 *Nechť matice $S \in R^{n\times m}$, $m \leq n$, má plnou hodnost a nechť $U \in R^{n\times k}$, $V \in R^{m\times k}$, $k \leq m$, jsou matice takové, že $S + UV^T$ má plnou hodnost. Pak platí*

$$(S + UV^T)^\dagger = S^\dagger - S^\dagger U(I + V^TS^\dagger U)^{-1}V^TS^\dagger. \tag{152}$$

Důkaz (a) Nejprve ukážeme, že matice $I + V^T S^\dagger U$ je regulární. Předpokládejme naopak, že pro nějaký vektor $x \neq 0$ platí $(I+V^T S^\dagger U)x = 0$. Pak nutně $y = S^\dagger U x \neq 0$. Musí tedy platit $(S^\dagger U + S^\dagger U V^T S^\dagger U)x = S^\dagger(S + UV^T)y = 0$, kde $y \neq 0$, což je ve sporu s předpokladem, že matice $S^\dagger$ 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 (149) je tato matice určena vztahem $(S + UV^T)^\dagger(S + UV^T) = I$. Protože

$$\begin{aligned}(S^\dagger - S^\dagger U(I + V^T S^\dagger U)^{-1} V^T S^\dagger)(S + UV^T) &= I + S^\dagger U(I + V^T S^\dagger U)(I + V^T S^\dagger U)^{-1} V^T \\ &\quad - S^\dagger U(I + V^T S^\dagger U)^{-1} V^T - S^\dagger U V^T S^\dagger U(I + V^T S^\dagger U)^{-1} V^T = I,\end{aligned}$$

platí $(S + UV^T)^\dagger = S^\dagger - S^\dagger U(I + V^T S^\dagger U)^{-1} V^T S^\dagger$. □

Věta 49 *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*

$$(H + UMU^T)^\dagger = B - BU(M^{-1} + U^T BU)^{-1} U^T B, \qquad B = H^\dagger. \tag{153}$$

Důkaz Jelikož $U = HV$, můžeme vzorec (153) zapsat ve tvaru

$$(H + HVMV^T H)^\dagger = B - BHV(M^{-1} + V^T HV)^{-1} V^T HB.$$

Platí

$$\begin{aligned}&\big(B - BHV(M^{-1} + V^T HV)^{-1} V^T HB\big)\big(H + HVMV^T H\big) \\ &\quad = BH - BHV(M^{-1} + V^T HV)^{-1} V^T HBH + BHVMV^T H \\ &\qquad - BHV(M^{-1} + V^T HV)^{-1} V^T HVMV^T H = BH,\end{aligned}$$

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 104 Podmínka $U = [d, Hy] = HV$ je splněna, pokud $s = -Hg$. Pak $d = -\alpha Hg$, takže $V = [-\alpha g, y]$.

Poznámka 105 Použijeme-li pseudoinverzní matice $S^\dagger$ a $B = H^\dagger$, můžeme psát

$$\tilde{y} = S^T y = S^\dagger H y, \qquad \tilde{q} = S^T q = S^\dagger H q, \tag{154}$$

$$\tilde{d} = S^\dagger d = S^T B d, \qquad \tilde{p} = S^\dagger p = S^T B p, \tag{155}$$

$$Hy = S\tilde{y}, \qquad Hq = S\tilde{q}, \tag{156}$$

$$Bd = (S^\dagger)^T \tilde{d}, \qquad Bp = (S^\dagger)^T \tilde{p}, \tag{157}$$

neboť podle (149) a (150) platí $S^\dagger H = S^T$ a $S^T B = S^\dagger$. Při odvozování součinového tvaru metod s proměnnou metrikou budeme používat čísla

$$a = y^T SS^T y = \tilde{y}^T \tilde{y}, \qquad b = y^T d = \tilde{y}^T \tilde{d}, \qquad c = d^T (SS^T)^\dagger d = \tilde{d}^T \tilde{d}. \tag{158}$$

Nejprve je třeba zobecnit lemma 11.

Lemma 12 *Nechť $H = SS^T$, kde matice $S \in R^{n\times m}$, $m \leq n$, 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 $I+MU^TH^\dagger U$. 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 102 a 103, 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 7 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 11 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 (146), aby byla splněna kvazinewtonovská podmínka

$$S_+S_+^Ty = \rho d. \tag{159}$$

Lemma 13 *Uvažujme aktualizaci (146), s nenulovým vektorem $\tilde{q}$ zvoleným tak, že*

$$D^2 \triangleq (\tilde{q}^T\tilde{y})^2 + (\frac{\rho}{\gamma}b - a)\tilde{q}^T\tilde{q} > 0, \tag{160}$$

kde $\tilde{y} = S^Ty$ (pokud $(\rho/\gamma)b > a$, lze volit $\tilde{q}$ libovolně). Pak kvazinewtonovská podmínka (159) je splněna právě tehdy, platí-li

$$p = \frac{(\rho/\gamma)d - S(\tilde{y}+\tau\tilde{q})}{\tilde{q}^T(\tilde{y}+\tau\tilde{q})} = \frac{(\rho/\gamma)d - S(\tilde{y}+\tau\tilde{q})}{D}.$$

Číslo $\tau = p^Ty$ se vypočte z rovnosti $\tilde{q}^T\tilde{y} + \tau\tilde{q}^T\tilde{q} = D$ (pokud $(\rho/\gamma)b = a$, lze volit $\tau = 0$).

Důkaz Použitím vztahu (146) 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 $\tilde{y} = S^T y$. Označíme-li $\tau = p^T y$, 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 $\tau = p^T y$, 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

$$D^2 = (\tilde{q}^T\tilde{y})^2 + (\frac{\rho}{\gamma}b - a)\tilde{q}^T\tilde{q}. \tag{161}$$

Výraz (161) 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 $a = \tilde{y}^T\tilde{y} = \tilde{q}^T\tilde{y} = \tilde{q}^T\tilde{q} > 0$. □

Nyní se budeme zabývat součinovým tvarem metod z Broydenovy třídy.

Lemma 14 *Nechť $U = S\tilde{U} = S[\tilde{d}, \tilde{y}]$. Uvažujme aktualizaci (146) (splňující kvazinewtonovskou podmínku (159)), kde $\tilde{q} = \tilde{U}\hat{q}$ a kde vektor $\hat{q} \in R^2$ je zvolen tak, aby byla splněna nerovvnost (160). Pak 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 13 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 (146), 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 12 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 (147) a (151)). Tento součin je podle lemmatu 10 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

$$\delta = \det(I + \tilde{p}\tilde{q}^T)\det(I + \tilde{q}\tilde{p}^T) = (1 + \tilde{q}^T\tilde{p})^2 \geq 0.$$

□

Poznámka 106 Podle lemmatu 14 existuje součinový tvar pouze pro ty metody z Broydenovy třídy, pro které $\delta \geq 0$ a $\mu \geq 0$. Ve větě 51 a důsledku 9 ukážeme, že tyto nutné podmínky jsou i podmínkami postačujícími (poznámka 108).

Nyní ukážeme, jak lze volit vektor $\tilde{q} = \tilde{U}\hat{q}$, abychom dostali jednotlivé metody z Broydenovy třídy. Jak vyplývá z důkazu lemmatu 13, je vektor p určen vektorem $\tilde{q}$ (existují obvykle dvě řešení). Navíc výsledná aktualizace nezávisí na normě vektoru $\tilde{q}$, neboť z (146) plyne, že vynásobíme-li vektor $\tilde{q}$ nějakým číslem, stačí tímto číslem vydělit vektor p. Proto budeme hledat vektor $\tilde{q}$ ve tvaru $\tilde{q} = \tilde{y} - \vartheta\tilde{d}$.

Věta 50 *Nechť jsou splněny předpoklady lemmatu 14 a nechť $\tilde{q} = \tilde{y} - \vartheta\tilde{d}$. Pak aktualizace (146) je ekvivalentní aktualizaci (116), pokud*

$$\frac{\rho}{\gamma}\frac{(b-\vartheta c)^2}{D^2} = \frac{1}{ab}(\eta(ac-b^2)+b^2), \tag{162}$$

kde

$$D^2 = \frac{\rho}{\gamma}b(a-2\vartheta b+\vartheta^2 c) - \vartheta^2(ac-b^2). \tag{163}$$

Jestliže $\eta = 0$, pak buď $\vartheta = 0$ nebo

$$\frac{1}{\vartheta} = \frac{1}{2}\left(\frac{\gamma}{\rho}+\frac{c}{b}\right).$$

V ostatních případech platí

$$\frac{1}{\vartheta} = -\frac{m_3}{m_2} \pm \sqrt{\left(\frac{\gamma}{\rho}+\frac{m_3}{m_2}\right)\left(\frac{c}{b}+\frac{m_3}{m_2}\right)},$$

kde m_1, m_2, m_3 jsou čísla určená vztahy (114), což lze zapsat ve tvaru

$$\frac{1}{\vartheta} = \frac{b}{\eta}\left(\frac{\eta-1}{a} \pm \frac{\gamma}{\rho}\sqrt{\delta\mu}\right).$$

Důkaz V důkazu lemmatu 14 jsme ukázali, že $\delta = (1+\tilde{q}^T\tilde{p})^2$. Výraz vystupující na pravé straně této rovnosti je podle lemmatu 13 roven číslu

$$1+\tilde{q}^T\tilde{p} = 1+q^Tp = \frac{\rho}{\gamma}\frac{q^Td}{D} = \frac{\rho}{\gamma}\frac{\tilde{q}^T\tilde{d}}{D} = \frac{\rho}{\gamma}\frac{b-\vartheta c}{D}.$$

Použijeme-li tuto rovnost a vztah (130), dostaneme (162). Jelikož $\tilde{q}^T\tilde{y} = a-\vartheta b$ a $\tilde{q}^T\tilde{q} = a-2\vartheta b+\vartheta^2 c$, dostaneme po dosazení těchto vztahů do (161) a po úpravě výraz (163). Vynásobíme-li rovnost (162) číslem aD^2 dostaneme

$$\frac{\rho}{\gamma}a(b-\vartheta c)^2 - bD^2 = \frac{\eta}{b}D^2(ac-b^2).$$

Dosadíme-li (163) do levé strany této rovnosti, sdružíme-li odpovídající si členy a vydělíme-li vzniklou rovnici číslem $ac-b^2$, můžeme psát

$$\vartheta^2\left(\frac{\rho}{\gamma}c+b\right) - 2\vartheta\frac{\rho}{\gamma}b = \frac{\eta}{b}D^2.$$

Jestliže $\eta = 0$, je buď $\vartheta = 0$ nebo $1/\vartheta = (\gamma/\rho + c/b)/2$. V opačném případě, dosazením za D^2 a dalšími úpravami spočívajícími v tom, že na obou stranách vytkneme výraz v hranaté závorce, převedeme všechny členy na levou stranu a výslednou rovnici vydělíme číslem ηa, dostaneme

$$\frac{(1-\eta)b}{\eta a}\left[\vartheta^2\left(\frac{\rho c}{\gamma b}+1\right)-2\vartheta\frac{\rho}{\gamma}\right]+\vartheta^2\frac{c}{b}-\frac{\rho}{\gamma}=0,$$

což po vydělení číslem $(\rho/\gamma)\vartheta^2$ s použitím (114) dává

$$\frac{1}{\vartheta^2}+\frac{2}{\vartheta}\frac{m_3}{m_2}-\left(\frac{m_3}{m_2}\left(\frac{\gamma}{\rho}+\frac{c}{b}\right)+\frac{\gamma c}{\rho b}\right)=0.$$

Tato kvadratická rovnice má řešení

$$\frac{1}{\vartheta}=-\frac{m_3}{m_2}\pm\sqrt{\left(\frac{m_3}{m_2}\right)^2+\frac{m_3}{m_2}\left(\frac{\gamma}{\rho}+\frac{c}{b}\right)+\frac{\gamma c}{\rho b}}=-\frac{m_3}{m_2}\pm\sqrt{\left(\frac{\gamma}{\rho}+\frac{m_3}{m_2}\right)\left(\frac{c}{b}+\frac{m_3}{m_2}\right)}.$$

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 107 Věta 50 udává způsob, jak lze k dané metodě s proměnnou metrikou (charakterizované parametrem η) nalézt součinový tvar (146). K dané hodnotě η najdeme podle věty 50 hodnotu ϑ určující vektor $\hat{q}$ a číslo D^2 (existují obvykle dvě řešení). Pak podle lemmatu 13 určíme vektor p (existují opět dvě řešení).

(a) Pro metodu DFP platí $\eta=0$, takže lze volit $\vartheta=0$.

(b) Pro metodu BFGS platí $\eta=1$, takže $m_3=0$, což dává $\vartheta=\pm\sqrt{\rho b/\gamma c}$.

(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 $\vartheta=\rho/\gamma$. 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 93).

Použití věty 50 není příliš vhodné pro explicitní vyjádření součinového tvaru. Jinou možnost udává následující věta, 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 51 *Nechť $a>0$, $b>0$, $c>0$ a $ac-b^2>0$. Uvažujme aktualizaci (116), kde $H=SS^T$, $\rho>0$, $\gamma>0$. Nechť $\delta\ge0$, $\mu\ge0$ a buď $\delta>0$ nebo $\mu>0$. Nechť $d=S\tilde{d}$ a $\tilde{y}=S^Ty$. Pak platí $H_+=S_+^TS_+$, kde*

$$\frac{1}{\sqrt{\gamma}}S_+=S+U\hat{p}\hat{q}^T\tilde{U}^T,\tag{164}$$

přičemž

$$\hat{p}\hat{q}^T=\frac{1}{\lambda}\begin{bmatrix}\frac{\rho}{\gamma}+a\sqrt{\mu}\\ -1-\frac{b}{\sqrt{\mu}}\end{bmatrix}\begin{bmatrix}\sqrt{\delta}-b\sqrt{\mu}\\ -\frac{\gamma}{\rho}\sqrt{\delta}+c\sqrt{\mu}\end{bmatrix}^T$$

a

$$\lambda=\left(\frac{\rho}{\gamma}c-b\right)+\left(b-\frac{\gamma}{\rho}a\right)\sqrt{\delta}+\left(ac-b^2\right)\sqrt{\mu}.$$

Důkaz (a) Z důkazu lemmatu 14 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 (164), 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 14 víme, že

$$\det(I + \tilde{U}\hat{p}\hat{q}^T\tilde{U}^T) = \sqrt{\delta},$$

takže podle lemmatu 10 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 $\lambda\neq 0$.
(c) Nechť $\delta=0$ a $\mu>0$. Jelikož podle předpokladu platí $ac-b^2>0$, můžeme vhodnou volbou znaménka $\sqrt{\mu}$ docílit toho, že $\lambda\neq 0$. Jelikož podle poznámky 85 je $\delta(\eta)$ lineární funkcí parametru η (se směrnicí $ac-b^2>0$), platí $\delta(\eta+\varepsilon)>0$ pro libovolné číslo $\varepsilon>0$. Pro $\delta(\eta+\varepsilon)>0$ můžeme použít postup uvedený v (b) a protože všechny veličiny v rozkladu (164) závisí spojitě na ε, lze použít limitní přechod a tvrzení platí i pro $\delta=\delta(\eta)=0$. □

Obě předchozí věty obsahují poměrně komplikované výrazy. Tyto výrazy se velmi zjednoduší pro základní metody (117), (118), (119).

Důsledek 9 Pro metodu DFP platí $\eta=0$, čili $\delta=\rho b/(\gamma a)$ a $\mu=\rho/(\gamma ab)$, takže

$$\frac{1}{\sqrt{\gamma}}S_+^{DFP}=S-\frac{1}{a}\left(SS^Ty\pm\sqrt{\frac{\rho a}{\gamma b}}d\right)\tilde{y}^T. \tag{165}$$

Pro metodu BFGS platí $\eta=1$, čili $\delta=\rho c/(\gamma b)$ a $\mu=1/b^2$, takže

$$\frac{1}{\sqrt{\gamma}}S_+^{BFGS}=S-\frac{1}{b}d\left(\tilde{y}\pm\sqrt{\frac{\rho b}{\gamma c}}\tilde{d}\right)^T. \tag{166}$$

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

$$\frac{1}{\sqrt{\gamma}}S_+^{R1}=S+\frac{\sqrt{\delta}-1}{(\rho/\gamma)^2c-2(\rho/\gamma)b+a}\left(\frac{\rho}{\gamma}d-SS^Ty\right)\left(\frac{\rho}{\gamma}\tilde{d}-\tilde{y}\right)^T. \tag{167}$$

V těchto vzorcích je $d=S\tilde{d}$ a $SS^Ty=S\tilde{y}$. Čitatel i jmenovatel v posledním vzorci je vždy nenulový (i když $\delta=0$).

Důkaz K odvození těchto vztahů můžeme použít buď větu 50 nebo větu 51. Použití věty 50 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 51. Přímé dosazení do (164) 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

$$\lambda=\left(\frac{\rho}{\gamma}c-b\right)+\frac{\gamma a}{\rho b}\left(\frac{\rho}{\gamma}c-b\right)\sqrt{\delta}=\left(\frac{\rho}{\gamma}c-b\right)(\sqrt{\delta}+1)/\sqrt{\delta},$$

$$\hat{p} = \begin{bmatrix} \frac{a}{b}(\frac{\rho b}{\gamma a} + \sqrt{\delta}) \\ -1 - \sqrt{\delta} \end{bmatrix} = \begin{bmatrix} \frac{a}{b}\sqrt{\delta}(\sqrt{\delta}+1) \\ -(\sqrt{\delta}+1) \end{bmatrix}, \quad \hat{q} = \frac{1}{a}\begin{bmatrix} 0 \\ \frac{\gamma a}{\rho b}(\frac{\rho}{\gamma}c - b)\sqrt{\delta} \end{bmatrix} = \frac{1}{a}\begin{bmatrix} 0 \\ (\frac{\rho}{\gamma}c - b)/\sqrt{\delta} \end{bmatrix},$$

takže po vykrácení

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{1}{a}\begin{bmatrix} \frac{a}{b}\sqrt{\delta} \\ -1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}^T = -\frac{1}{a}\begin{bmatrix} \pm\sqrt{\frac{\rho a}{\gamma b}} \\ 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}^T.$$

(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

$$\begin{aligned} \lambda &= \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} - \frac{1}{b}(ac - b^2) = \frac{c}{b}\left(\frac{\rho}{\gamma}b - a\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} \\ &= \left(\frac{\rho}{\gamma}b - a\right)\frac{\gamma}{\rho}\left(\frac{\rho c}{\gamma b} + \sqrt{\delta}\right) = \left(\frac{\rho}{\gamma}b - a\right)\frac{\gamma}{\rho}\sqrt{\delta}(\sqrt{\delta}+1), \end{aligned}$$

$$\hat{p} = \frac{1}{b}\begin{bmatrix} \frac{\rho}{\gamma}b - a \\ 0 \end{bmatrix}, \quad \hat{q} = \begin{bmatrix} \sqrt{\delta}+1 \\ -\frac{\gamma}{\rho}(\sqrt{\delta} + \frac{\rho c}{\gamma b}) \end{bmatrix} = \begin{bmatrix} \sqrt{\delta}+1 \\ -\frac{\gamma}{\rho}\sqrt{\delta}(\sqrt{\delta}+1) \end{bmatrix},$$

takže po vykrácení

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}} \\ -1 \end{bmatrix}^T = -\frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \pm\sqrt{\frac{\rho b}{\gamma c}} \\ 1 \end{bmatrix}^T.$$

(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

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c - b}{\frac{\rho}{\gamma}b - a} + \sqrt{\delta}\right) = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta}(\sqrt{\delta}+1),$$

$$\hat{p} = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}, \quad \hat{q} = \sqrt{\delta}\begin{bmatrix} 1 \\ -\frac{\gamma}{\rho} \end{bmatrix} = \frac{\gamma}{\rho}\sqrt{\delta}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix},$$

a po vykrácení

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T}{\left(\frac{\rho}{\gamma}b - a\right)(\sqrt{\delta}+1)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T(\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c - b}{\frac{\rho}{\gamma}b - a} - 1\right)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T(\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}\right)^2 c - 2\left(\frac{\rho}{\gamma}\right)b + a}.$$

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 108 Ve větě 51 jsme předpokládali, že buď $\delta > 0$ nebo $\mu > 0$, neboť v opačném případě není matice $\hat{p}\hat{q}^T$ vystupující v (164) definovaná. I v tomto případě je však možné vyjádřit aktualizaci (116) v součinovém tvaru. Hodnota $\mu = 0$ odpovídá metodě hodnosti 1, pro kterou (po vykrácení výrazem $\sqrt{\delta}$ umožněným spojitou závislostí δ na η) platí (167). Pokud $ac - b^2 > 0$, nemůže být jmenovatel ani čitatel v (167) nulový.

Poznámka 109 Vzorec (167) lze upravit tak, aby se v něm neodečítala blízká čísla. Dosadíme-li do (167) 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 (133). 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 110 Má-li matice $S \in R^{n\times m}$, $m \geq n$, plnou hodnost, lze použitím definice 24 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

$$B = H^{-1} = (SS^T)^{-1} = (S^\dagger)^T S^\dagger = A^T A.$$

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 52 *Nechť $a > 0$, $b > 0$, $c > 0$ a $ac - b^2 > 0$. Uvažujme aktualizaci (133), 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*

$$\sqrt{\gamma} A_+ = A - \frac{1}{\sqrt{\delta}} \tilde{U}\hat{p}\hat{q}^T (BU)^T, \tag{168}$$

přičemž $\hat{p}$ a $\hat{q}$ jsou vektory vystupující ve větě 51.

Důkaz Podle (148) platí

$$\frac{1}{\gamma} S_+ S_+^T = (I + pq^T)SS^T(I + qp^T),$$

což s použitím lemmatu 10 dává

$$\gamma A_+^T A_+ = (I + qp^T)^{-1} A^T A (I + pq^T)^{-1} = \left(I - \frac{1}{\sqrt{\delta}} qp^T\right) A^T A \left(I - \frac{1}{\sqrt{\delta}} pq^T\right).$$

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

$$Ap = AU\hat{p} = A\left[d,\, (A^T A)^{-1} y\right]\hat{p} = \left[\tilde{d},\, \tilde{y}\right]\hat{p} = \tilde{U}\hat{p}$$

a

$$q = A^T \tilde{q} = A^T \tilde{U}\hat{q} = A^T \left[\tilde{d},\, \tilde{y}\right]\hat{q} = \left[A^T Ad,\, y\right]\hat{q} = BU\hat{q}.$$

□

Poznámka 111 Pro metodu DFP platí

$$\sqrt{\gamma} A_+^{DFP} = A - \frac{1}{b}\left(\tilde{d} \pm \sqrt{\frac{\gamma b}{\rho a}}\tilde{y}\right) y^T. \tag{169}$$

Pro metodu BFGS platí

$$\sqrt{\gamma} A_+^{BFGS} = A - \frac{1}{c}\tilde{d}\left(A^T Ad \pm \sqrt{\frac{\gamma c}{\rho b}} y\right)^T. \tag{170}$$

Pro metodu hodnosti 1 platí

$$\sqrt{\gamma} A_+^{R1} = A + \frac{1/\sqrt{\delta} - 1}{(\gamma/\rho)^2 a - 2(\gamma/\rho)b + c}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho} y - A^T Ad\right)^T. \tag{171}$$

kde $1/\delta = (\gamma/\rho)((\gamma/\rho)a - b)/((\gamma/\rho)b - c)$. Vzorec (171) lze upravit na tvar

$$\sqrt{\gamma} A_+^{R1} = A + \frac{1}{\frac{\gamma}{\rho} b - c \pm \sqrt{\frac{\gamma}{\rho}\left(\frac{\gamma}{\rho} b - c\right)\left(\frac{\gamma}{\rho} a - b\right)}}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho} y - A^T Ad\right)^T.$$

Ve všech těchto vzorcích je $y = A^T \tilde{y}$ a $A^T Ad = A^T \tilde{d}$ (neboť $\tilde{d} = Ad$). Poznamenejme, že pro minimalizaci součtu čtverců má praktický význam pouze metoda BFGS, která používá vektor $\tilde{d} = Ad$. Ostatní metody potřebují navíc vektor $\tilde{y} = A(A^T A)^{-1} y$, takže je nutné invertovat matici $A^T A$.

Součinový tvar $H = SS^T$, kde $S \in R^{n\times m}$ a $m \leq n$, lze modifikovat tak, že se místo matice S používá matice Z, jejíž sloupce tvoří ortonormální bázi v $\mathcal{L}(S)$.

Věta 53 *Nechť $H = SS^T$, kde matice $S \in R^{n\times m}$, $m \leq n$, má plnou hodnost, a $Z \in R^{n\times m}$ je matice jejíž sloupce tvoří bázi v $\mathcal{L}(S)$. Pak platí $H = Z(Z^T BZ)^{-1} Z^T$, 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

$$SS^T = ZMM^T Z^T.$$

Použitím definice 24 se snadno ověří, že $S^\dagger = M^{-1} Z^\dagger$ (neboť $Z^\dagger Z = I$ podle (149)). Protože podle (150) platí $(SS^T)^\dagger = (S^\dagger)^T S^\dagger$, můžeme psát

$$Z^T (SS^T)^\dagger Z = Z^T (S^\dagger)^T S^\dagger Z = Z^T (M^{-1} Z^\dagger)^T M^{-1} Z^\dagger Z = (MM^T)^{-1}, \tag{172}$$

takže

$$Z(Z^TBZ)^{-1}Z^T = Z(Z^T(SS^T)^\dagger Z)^{-1}Z^T = ZMM^TZ = SS^T = H$$

. □

Poznámka 112 V předchozí větě nejsou kladeny žádné požadavky na výběr matice Z, takže lze položit $Z = S$. V tomto případě podle (150) platí $S^TBS = I$ (sloupce matice S jsou B-ortogonální), takže $S(S^TBS)^{-1}S^T = SS^T$.

Poznámka 113 V dalším textu budeme používat redukované matice $\tilde{B} = Z^TBZ$ a $\tilde{H} = \tilde{B}^{-1}$. Směrový vektor $s = -Hg$ se vypočte podle vzorců

$$\tilde{g} = Z^Tg, \qquad \tilde{s} = -\tilde{H}\tilde{g}, \qquad s = Z\tilde{s}$$

(vektor $\tilde{s}$ lze také získat řešením soustavy rovnic $\tilde{B}\tilde{s} = -\tilde{g}$). Pak $\tilde{d} = \alpha\tilde{s}$ a $\tilde{y} = Z^Ty$. 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 15 *Matice $\tilde{B}$ a $\tilde{H}$ jsou pozitivně definitní.*

Důkaz Podle (172) 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 $Z^TZ = I$.

Lemma 16 *Jestliže $Z^TZ = I$, platí $\tilde{H} = Z^THZ$, kde $H = SS^T = Z(Z^TBZ)^{-1}Z^T$.*

Důkaz Podle (172) platí $\tilde{B} = (MM^T)^{-1}$, takže $\tilde{H} = \tilde{B}^{-1} = MM^T$. Pokud $ZZ^T = I$, můžeme psát

$$Z^THZ = Z^TSS^TZ = Z^T(ZMM^TZ^T)Z = MM^T,$$

odkud plyne dokazované tvrzení. □

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 54 *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 $Z\tilde{U} = [Z\tilde{d}, Z\tilde{H}Z^Ty] = [d, Hy] = U$. Platí tedy $(1/\gamma)H_+ = H + UMU$. □

Poznámka 114 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

$$\gamma\tilde{B}_+ = \tilde{B} + \frac{\gamma}{\rho b}\tilde{y}\tilde{y}^T - \frac{1}{c}\tilde{B}\tilde{d}(\tilde{B}\tilde{d})^T + \frac{\beta}{c}\left(\frac{c}{b}\tilde{y} - \tilde{B}\tilde{d}\right)\left(\frac{c}{b}\tilde{y} - \tilde{B}\tilde{d}\right)^T.$$

Poznámka 115 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^{HR} hodnotu odpovídající metodě hodnosti 1. Pokud $0 < \rho/\gamma \leq b/c$, jsou rozložitelné ty metody pro něž $\eta \geq \eta^{HR}$, kde $\eta^* < \eta^{HR} < 0$. Pokud $a/b < \rho/\gamma$, jsou rozložitelné ty metody pro něž $\eta^* \leq \eta \leq \eta^{HR}$, kde $\eta^{HR} > 1$.

(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.

4.3 Variační odvození metod s proměnnou metrikou

Velmi zajímavý způsob jak lze získat metody s proměnnou metrikou spočívá v použití minimalizačního principu.

Lemma 17 *Nechť $\psi(X) : R^{n\times n} \to R$ je symetrická funkce matice X (takže $\psi(X^T) = \psi(X)$). Nechť X^* je symetrická matice, která minimalizuje $\psi(X)$ na množině symetrických matic řádu n splňujících podmínku $Xp = q$. Pak existuje vektor Lagrangeových multiplikátorů u takový, že*

$$\frac{\partial\psi(X^*)}{\partial X} = up^T + pu^T.$$

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

$$\begin{aligned} \frac{\partial L(X,u,v)}{\partial x_{kl}} &= \frac{\partial\psi(X)}{\partial x_{kl}} - 2u_kp_l + v_{kl} - v_{lk} = 0, \\ \frac{\partial L(X,u,v)}{\partial x_{lk}} &= \frac{\partial\psi(X)}{\partial x_{lk}} - 2u_lp_k + v_{lk} - v_{kl} = 0, \end{aligned}$$

kde k, l je libovolná dvojice indexů. Sečteme-li obě rovnosti a použijeme-li symetrii funkce $\psi(X)$, dostaneme

$$2\frac{\partial\psi(X)}{\partial x_{kl}} - 2u_kp_l - 2u_lp_k = 0,$$

což maticově zapsáno dává tvrzení lematu. □

Pro variační odvození metod s proměnou metrikou se nejčastěji používá Frobeniova norma matice. V tomto případě má funkce $\psi(x)$ tvar $\psi(x) = (1/2)\|X\|_F^2 = (1/2)\mathrm{Tr}X^TX$ (tato funkce je symetrická a konvexní).

Lemma 18 *Symetrická matice X^* má minimální Frobeniovu normu na množině symetrických matic řádu n splňujících podmínku $Xp = q$ právě tehdy, když*

$$X^* = \frac{1}{p^T p}(pq^T + qp^T) - \frac{q^T p}{(p^T p)^2} pp^T. \tag{173}$$

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 17 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 55 *Nechť W je symetrická pozitivně definitní 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*

$$\left(\frac{1}{\gamma}\tilde{H} - H\right) y = \frac{\rho}{\gamma} d - Hy \triangleq w$$

právě tehdy, když

$$\frac{1}{\gamma}H_+ = H + \frac{Wyw^T + w(Wy)^T}{y^T Wy} - \frac{w^T y}{y^T Wy}\frac{Wy(Wy)^T}{y^T Wy}. \tag{174}$$

Důkaz Položme $X = W^{-1/2}((1/\gamma)\tilde{H} - H)W^{-1/2}$. 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 18, podle kterého

$$\begin{aligned} X^* &= \frac{1}{p^T p}(pq^T + qp^T) - \frac{q^T p}{(p^T p)^2} pp^T \\ &= \frac{1}{y^T Wy}(W^{1/2}yw^T W^{-1/2} + W^{-1/2}wy^T W^{1/2}) - \frac{w^T y}{(y^T Wy)^2} W^{1/2}yy^T W^{1/2}. \end{aligned}$$

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 116 Zvolíme-li matici W tak, aby platilo $Wy = d$, přejde vzorec (174) na vztah (127). 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 (174) patří do Broydenovy třídy právě tehdy, je-li vektor Wy lineární kombinací vektorů d a Hy. Protože aktualizace (174) nezávisí na normě vektoru Wy, budeme předpokládat, že $Wy = d - \vartheta Hy$.

Věta 56 *Nechť jsou splněny předpoklady věty 55 a nechť $Wy = d - \vartheta Hy$. Pak aktualizace (174) je ekvivalentní aktualizaci (116), pokud*

$$\eta = \frac{b(b - \vartheta^2(\rho/\gamma)a)}{(b - \vartheta a)^2}. \tag{175}$$

Jestliže $\eta = 1$, pak buď $\vartheta = 0$, nebo

$$\frac{1}{\vartheta} = \frac{1}{2}\left(\frac{\rho}{\gamma} + \frac{a}{b}\right).$$

V ostatních případech platí

$$\frac{1}{\vartheta} = \frac{a}{\eta - 1}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

Důkaz Jestliže $Wy = d - \vartheta Hy$, platí $y^T w = (\rho/\gamma)b - a$ a $y^T Wy = b - \vartheta a$. Dosadíme-li tyto vztahy do (174) a porovnáme-li záporně vzaté koeficienty u smíšených členů (v aktualizaci (116) je tento záporně vzatý koeficient roven η/b), dostaneme

$$\frac{b - \vartheta^2(\rho/\gamma)a}{(b - \vartheta a)^2} = \frac{\eta}{b},$$

odkud plyne (175). Jestliže $\eta = 1$, dostaneme použitím (175) rovnici $b^2 - 2\vartheta ab + \vartheta^2 a^2 = b^2 - \vartheta^2(\rho/\gamma)ab$, neboli

$$\vartheta^2\left(\frac{\rho}{\gamma}b + a\right) - 2\vartheta b = 0,$$

takže buď $\vartheta = 0$, nebo $1/\vartheta = (\rho/\gamma + a/b)/2$. V opačném případě dostaneme $\eta(b^2 - 2\vartheta ab + \vartheta^2 a^2) = b^2 - \vartheta^2(\rho/\gamma)ab$, což po vydělení číslem $\vartheta^2 ab^2$ dává

$$\frac{\eta - 1}{a}\frac{1}{\vartheta^2} - 2\frac{\eta}{b}\frac{1}{\vartheta} + \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right) = m_3\frac{1}{\vartheta^2} + 2\,m_2\frac{1}{\vartheta} + m_1 = 0,$$

kde m_1, m_2, m_3 jsou čísla určená vztahy (114). Tato kvadratická rovnice má řešení

$$\frac{1}{\vartheta} = \frac{-m_2 \pm \sqrt{m_2^2 - m_1 m_3}}{m_3} = \frac{-m_2 \pm \sqrt{\mu}}{m_3} = \frac{a}{\eta - 1}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

□

Poznámka 117 Věta 56 udává způsob, jak lze k dané metodě s proměnnou metrikou (charakterizované parametrem η) nalézt aktualizaci tvaru (174). K dané hodnotě η najdeme podle věty 56 hodnotu ϑ určující vektor Wy (existují obvykle dvě řešení).

(a) Pro metodu DFP platí $\eta = 0$ a $\mu = \rho/(\gamma ab)$, takže $\vartheta = \sqrt{\gamma b/(\rho a)}$.

(b) Pro metodu BFGS platí $\eta = 1$, takže lze volit $\vartheta = 0$.

(c) Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$ a $\mu = 0$, takže $\vartheta = \gamma/\rho$.

Poznámka 118 Analogický postup lze použít pro aktualizaci matice B. Nechť V je symetrická pozitivně definití matice. Pak symetrická matice B_+ minimalizuje Frobeniovu normu $\|V^{-1/2}(\gamma\tilde{B} - B)V^{-1/2}\|_F$ na množině symetrických matic $\tilde{B}$ řádu n splňujících kvazinewtonovskou podmínku

$$\left(\gamma\tilde{B} - B\right) d = \frac{\gamma}{\rho} y - Bd \triangleq v$$

právě tehdy, když

$$\gamma B_+ = \left(B + \frac{Vdv^T + v(Vd)^T}{d^T V d} - \frac{v^T d}{d^T V d}\frac{Vd(Vd)^T}{d^T V d}\right). \tag{176}$$

Poznámka 119 Zvolíme-li matici V tak, aby platilo $Vd = y$, přejde vzorec (176) na vztah duální k (127). Metoda DFP tedy minimalizuje Frobeniovu normu $\|V^{-1/2}(\gamma B_+ - BH)V^{-1/2}\|_F$, pokud $Vd = y$.

Poznámka 120 Zvolíme-li matici V tak že $Vd = y - \vartheta Bd$, je aktualizace (176) ekvivalentní aktualizaci (133), pokud

$$\beta = \frac{b(b - \vartheta^2(\gamma/\rho)c)}{(b - \vartheta c)^2}.$$

Jestliže $\beta = 1$, pak buď $\vartheta = 0$, nebo

$$\frac{1}{\vartheta} = \frac{1}{2}\left(\frac{\gamma}{\rho} + \frac{c}{b}\right).$$

V ostatních případech platí

$$\frac{1}{\vartheta} = \frac{c}{\beta - 1}\left(\frac{\beta}{b} \pm \sqrt{\frac{\mu}{\delta}}\right)$$

(číslo μ/δ je určeno vzorcem (144)).

(a) Pro metodu DFP platí $\beta = 0$, takže lze volit $\vartheta = 0$.

(b) Pro metodu BFGS platí $\beta = 1$ a $\mu/\delta = \gamma/(\rho bc)$, takže $\vartheta = \sqrt{\rho b/(\gamma c)}$.

(c) Pro metodu hodnosti 1 platí $\beta = (\gamma/\rho)/(\gamma/\rho - c/b)$ a $\mu/\delta = 0$, takže $\vartheta = \rho/\gamma$.

Poznamenejme, že aktualizaci (176) používají strukturované metody s proměnnou metrikou pro minimalizaci součtu čtverců (vzorec (296)).

Poznámka 121 Zvolíme-li $V = I$, dostaneme metodu, která nepatří do Broydenovy třídy a která se nazývá Powellovou symetrizací Broydenovy metody. Platí

$$B_+^{PSB} = \frac{1}{\gamma}\left(B + \frac{dv^T + vd^T}{d^T d} - \frac{v^T d}{d^T d}\frac{dd^T}{d^T d}\right).$$

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 149).

Kromě Frobeniovy normy lze použít i jiná minimalizační kriteria. Velmi se osvědčila funkce $\psi(X) = \operatorname{Tr} X - \ln\det X$. Nechť λ_i, $1 \le i \le n$, jsou vlastní čísla 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 nebudou ani příliš málá ani příliš veliká. Je důležité, že použití této funkce vede na aktualizaci s nejvýše dvěma korekčními členy.

Lemma 19 *Nechť* $\psi(X) = \operatorname{Tr} X - \ln\det X$. *Pak platí*

$$\frac{\partial\psi(X)}{\partial X} = I - X^{-1}.$$

Důkaz To, že $\partial \operatorname{Tr}X/\partial X = I$, je zřejmé. Dokážeme, že $\partial \ln\det X/\partial X = X^{-1}$. Použijeme toho, že každou symetrickou matici X lze zapsat ve tvaru $X = Q^T\Lambda Q$, kde Q je ortogonální matice (takže $Q^TQ = QQ^T = I$ a $\det Q = 1$) a Λ je diagonální matice obsahující vlastní čísla matice X. Použijeme-li toto vyjádření a větu o násobení determinantů, můžeme psát

$$\ln\det(X) = \ln\det(Q^T\Lambda Q) = \ln\det(\Lambda) = \ln\prod_{i=1}^{n}\lambda_i = \sum_{i=1}^{n}\ln\lambda_i.$$

Platí tedy

$$\frac{\partial\ln\det(X)}{\partial x_{kl}} = \sum_{i=1}^{n}\frac{\partial\ln\det(X)}{\partial\lambda_i}\frac{\partial\lambda_i}{\partial x_{kl}} = \sum_{i=1}^{n}\lambda_i^{-1}\frac{\partial\lambda_i}{\partial x_{kl}}.$$

Z druhé strany $\Lambda = QXQ^T$, takže

$$\lambda_i = e_i^T\Lambda e_i = e_i^TQXQ^Te_i = \sum_{k=1}^{n}\sum_{l=1}^{n}Q_{ik}x_{kl}Q_{il},$$

takže $\partial\lambda_i/\partial x_{kl} = Q_{ik}Q_{il}$ a

$$\frac{\partial\ln\det(X)}{\partial x_{kl}} = \sum_{i=1}^{n}\lambda_i^{-1}\frac{\partial\lambda_i}{\partial x_{kl}} = Q_{ik}\lambda_i^{-1}Q_{il},$$

Jelikož $X = Q^T\Lambda Q$, platí $X^{-1} = Q^T\Lambda^{-1}Q$, takže

$$(X^{-1})_{kl} = e_k^TX^{-1}e_l = \sum_{i=1}^{n}Q_{ik}\lambda_i^{-1}Q_{il}.$$

Porovnáme-li obě vyjádření, vidíme, že platí $\partial\ln\det X/\partial X = X^{-1}$. □

Lemma 20 *Symetrická matice* X^* *minimalizuje funkci* $\psi(X) = \operatorname{Tr}X - \ln\det X$ *na množině symetrických matic řádu* n *splňujících podmínku* $Xp = q$ *právě tehdy, když*

$$(X^*)^{-1} = I - \frac{1}{q^Tp}(pq^T + qp^T) + \frac{1}{p^Tq}pp^T + \frac{q^Tq}{(p^Tq)^2}pp^T. \tag{177}$$

Důkaz Jelikož podle lemmatu 19 lze psát $\partial\psi(X)/\partial X = I - X^{-1}$, musí podle lemmatu 17 platit $(X^*)^{-1} = I - up^T - pu^T$. Z podmínky $(X^*)^{-1}q = p$ dostaneme $p = q - up^Tq - pu^Tq$, neboli

$$u = \frac{1}{p^Tq}(q - p - u^Tqp),$$

takže

$$u^Tq = \frac{1}{p^Tq}(q^Tq - p^Tq - u^Tqp^Tq),$$

neboli $2u^Tq = q^Tq/p^Tq - 1$, což dává

$$u = \frac{1}{p^Tq}\left(q - p - \frac{1}{2}\left(\frac{q^Tq}{p^Tq} - 1\right)p\right) = \frac{1}{p^Tq}\left(q - \frac{1}{2}\left(\frac{q^Tq}{p^Tq} + 1\right)p\right).$$

Dosadíme-li tento vektor do vztahu $(X^*)^{-1} = I - up^T - +pu^T$ a uvážíme-li konvexitu funkce $\psi(X)$, dostaneme tvrzení lemmatu. □

Věta 57 *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 $a = y^THy$, $b = y^Td$, $c = d^TBd$.

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 20, podle kterého

$$\begin{aligned}
(X^*)^{-1} &= I - \frac{1}{q^Tp}(pq^T + qp^T) + \frac{1}{p^Tq}pp^T + \frac{q^Tq}{(p^Tq)^2}pp^T \\
&= I - \frac{\gamma}{\rho b}\left(\frac{\rho}{\gamma}H^{1/2}yd^TH^{-1/2} + \frac{\rho}{\gamma}H^{-1/2}dy^TH^{1/2}\right) \\
&\quad + \frac{\gamma}{\rho b}H^{1/2}yy^TH^{1/2} + \left(\frac{\gamma}{\rho b}\right)^2\frac{\rho^2 c}{\gamma^2}H^{1/2}yy^TH^{1/2} \\
&= I - \frac{1}{b}(H^{1/2}yd^TH^{-1/2} + H^{-1/2}dy^TH^{1/2}) + \frac{1}{b}\left(\frac{\gamma}{\rho} + \frac{c}{b}\right)H^{1/2}yy^TH^{1/2}.
\end{aligned}$$

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 122 Podle věty 57 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 123 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 $a = y^THy$, $b = y^Td$, $c = d^TBd$.

Poznámka 124 Podle poznámlky 123 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 (164) a které obsahuje i aktualizace hodnosti 2. Abychom mohli použít variační princip, zapíšeme kvazinewtonovskou podmínku ve tvaru

$$\frac{1}{\sqrt{\gamma}}S_+^Ty = \tilde{z}, \qquad \frac{1}{\sqrt{\gamma}}S_+\tilde{z} = \frac{\rho}{\gamma}d, \qquad \tilde{z}^T\tilde{z} = \frac{\rho}{\gamma}b, \tag{178}$$

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 58 *Nechť T je symetrická pozitivně definití matice. Pak Frobeniova norma $\|T^{-1/2}(S_+/\sqrt{\gamma} - S)\|_F$ je minimální na množině všech matic splňujících kvazinewtonovskou podmínku (178) právě tehdy, platí-li*

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{Ty}{y^TTy}\tilde{y}^T + \left(\frac{\rho}{\gamma}d - z + \frac{y^Tz}{y^TTy}Ty\right)\frac{\tilde{z}^T}{\tilde{z}^T\tilde{z}}, \tag{179}$$

kde $\tilde{y} = S^Ty$ a $z = S\tilde{z}$.

Důkaz Označme $X = S_+/\sqrt{\gamma}$. Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2}\left\|T^{-1/2}(X-S)\right\|_F^2 + \tilde{u}^T\left(X^Ty - \tilde{z}\right) + v^T\left(X\tilde{z} - \frac{\rho}{\gamma}d\right) \\ &= \sum_{i=1}^m\left[\frac{1}{2}(x_i-s_i)^T T^{-1}(x_i-s_i) + \tilde{u}_i y^T x_i + \tilde{z}_i v^T x_i\right] - \tilde{u}^T\tilde{z} - \frac{\rho}{\gamma}v^T d, \end{aligned}$$

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 (178) dostaneme

$$X^Ty = S^Ty - y^TTy\tilde{u} - v^TTy\tilde{z} = \tilde{z} \quad \Rightarrow \quad \tilde{u} = \frac{1}{y^TTy}\left(S^Ty - (1+v^TTy)\tilde{z}\right),$$

což po dosazení do předchozí rovnosti dává

$$X - S = -\frac{Ty}{y^TTy}\tilde{y}^T + w\tilde{z}^T,$$

kde $w \in R^n$ je zatím neznámý vektor (jednoznačně určený vektorem v). Užitím druhé podmínky z (178) dostaneme

$$X\tilde{z} = S\tilde{z} - \frac{y^TS\tilde{z}}{y^TTy}Ty + \tilde{z}^T\tilde{z}w = \frac{\rho}{\gamma}d \quad \Rightarrow \quad w = \frac{1}{\tilde{z}^T\tilde{z}}\left(\frac{\rho}{\gamma}d - z + \frac{y^Tz}{y^TTy}Ty\right),$$

což po dosazení do předchozí rovnosti (s využitím vztahu $X = S_+/\sqrt{\gamma}$) dává (179). Postačitelnost plyne z konvexity Frobeniovy normy. □

Poznámka 125 Zvolíme-li matici T tak, aby platilo $Ty = (\rho/\gamma)d - z$, výraz (179) se velmi zjednoduší. Po dosazení a úpravě dostaneme

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{(\rho/\gamma)d - z}{(\rho/\gamma)b - y^Tz}(\tilde{y} - \tilde{z})^T \tag{180}$$

Položíme-li $\tilde{z} = \pm\sqrt{\rho b/(\gamma c)}\,\tilde{d}$, dostaneme metodu BFGS (vzorec (166)).

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 (179) zapsat ve tvaru

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{Hv}{y^THv}y^TS + \left(\frac{\rho}{\gamma}d - z + \frac{y^Tz}{y^THv}Hv\right)\frac{\tilde{z}^T}{\tilde{z}^T\tilde{z}}. \tag{181}$$

Věta 59 *Nechť H_+ je symetrická matice určená podle (116), kde $H = SS^T$, $d = -\alpha Hg$, $a > 0$, $b > 0$, $c > 0$ a $\eta \ge 0$ (takže $\delta > 0$). Nechť $B = H^\dagger = S(S^TS)^{-1}S^T$ a S_+ je matice určená podle (179) nebo (181), kde*

$$Ty = Hv = \frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy, \qquad \tilde{z} = \frac{\rho}{\gamma}\frac{b}{\sqrt{\delta}}S^TBTy = \frac{\rho}{\gamma}\frac{b}{\sqrt{\delta}}S^TBHv$$

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 (130)). 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 (178)). Jelikož $\tilde{z}^T\tilde{z} = \vartheta^2 v^T Hv$ (neboť podle definice 24 platí $HBHBH = H$), je tato hodnota dána výrazem

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{b}{v^T Hv}.$$

Speciální volbu $\tilde{z} = \vartheta S^T BHv$ používáme proto, že se tím velmi zjednoduší aktualizace (181), neboť v tomto případě platí $z = S\tilde{z} = \vartheta Hv$, takže

$$\frac{y^T z}{y^T Hv}Hv - z = \vartheta\frac{y^T Hv}{y^T Hv}Hv - \vartheta Hv = 0.$$

(b) Jelikož norma vektoru Hv neovlivní tvar aktualizace (181), budeme předpokládat, že $y^T Hv = 1$. Položíme-li

$$Hv = \frac{\alpha_1}{b}d + \frac{\alpha_2}{a}Hy$$

(což lze, neboť $d = -\alpha g$), 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 výrazu odvozeného v (a), dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{b}{v^T Hv} = \frac{\rho}{\gamma}\frac{ab^3}{\alpha_1^2(ac-b^2)+b^2}$$

(c) Nyní využijeme toho, že vektory $\tilde{z}$ a Hv, uvedené v (a) a (b), umožňují zapsat aktualizaci (181) ve velmi jednoduchém tvaru

$$\frac{1}{\sqrt{\gamma}}S_+ = S - Hvy^T S + \frac{\vartheta}{b}dv^T HBS. \tag{182}$$

Položíme-li $H = SS^T$ a $H_+ = S_+S_+^T$, dostaneme z předchozího vztahu (po vynásobení)

$$\frac{1}{\gamma}H_+ = H - (Hvy^T H + Hyv^T H) + \frac{\vartheta}{b}(dv^T H + Hvd^T) + aHvv^T H - \frac{\vartheta}{b}(dv^T H + Hvd^T) + \frac{\vartheta^2}{b^2}v^T Hvdd^T$$

$$= H - (Hvy^T H + Hyv^T H) + aHvv^T H + \frac{\vartheta^2}{b^2}v^T Hvdd^T.$$

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 44. K určení parametru η stačí porovnat koeficienty u $Hyy^T H$ v obou vyjádřeních. Podle věty 44 se tento koeficient rovná $(\eta-1)/a$ a dosazením vektoru $Hv = (\alpha_1/b)d + (\alpha_2/a)Hy$ do předchozího vztahu 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 $\alpha_1^2 = (1-\alpha_2)^2 = \alpha_2^2 - 2\alpha_2 + 1 = \eta$. Dosadíme-li $\alpha_1^2 = \eta$ do výrazu odvozeného v (b) a použijeme-li číslo δ definované v poznámce 85, dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{ab^3}{\eta(ac-b^2)+b^2} = \left(\frac{\rho}{\gamma}\right)^2\frac{b^2}{\delta}.$$

□

Důsledek 10 *Nechť jsou splněny předpoklady věty 59 a nechť*

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)\tilde{y} + \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}}d\left(\frac{\sqrt{\eta}}{b}\tilde{d} + \frac{1-\sqrt{\eta}}{a}\tilde{y}\right)^T$$

$$= S - \left(\left(\frac{\sqrt{\eta}}{b} - \vartheta\frac{1-\sqrt{\eta}}{a}\right)d - \frac{1-\sqrt{\eta}}{a}Hy\right)\tilde{y} + \vartheta\frac{\sqrt{\eta}}{b}d\tilde{d}^T, \tag{183}$$

kde $\sqrt{\eta}$ *a* $\sqrt{\delta}$ *jsou libovolné hodnoty (kladné i záporné) takové, že* $(\sqrt{\eta})^2 = \eta$ *a* $(\sqrt{\delta})^2 = \delta$ *a kde* $\vartheta = (\rho/\gamma)/\sqrt{\delta}$. *Pak platí* $H_+ = S_+S_+^T$.

Důkaz Dokazovaný vztah dostaneme prostým dosazením vektoru Hv a čísla ϑ, uvedených ve větě 59, do vzorce (182) a použitím vztahů (154)–(155). □

Poznámka 126 Věta 59 používá jiné předpoklady než věta 51, nerovnost $\mu \geq 0$ je nahražena nerovností $\eta \geq 0$. Vztah (183) lze tedy použít pro každou perfektní metodu z Broydenovy třídy. Na druhé straně matice $(1/\sqrt{\gamma})S_+ - S$ v (183) má obecně hodnost 2, takže (183) vyžaduje více numerických operací než (164). Pro $\eta = 0$ (DFP) nebo $\eta = 1$ (BFGS) dávají oba vztahy stejné výsledeky, dosazení do (183) je však nesrovnatelně jednodušší. Je zajímavé porovnat (183) s pseudosoučinovým tvarem (121) uvedeným v poznámce 90.

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 34.

Věta 60 *Nechť* $\tilde{G}$ *je matice taková, že* $\tilde{G}d = y$, *tedy například*

$$\tilde{G} = \int_0^1 G(x + \lambda d)\mathrm{d}\lambda. \tag{184}$$

Označme $R = \tilde{G}^{1/2}H\tilde{G}^{1/2}$ *a* $R'_+ = \tilde{G}^{1/2}H_+\tilde{G}^{1/2}$. *Pak jestliže* $0 \leq \eta \leq 1$ *a* $b/c \leq \rho/\gamma \leq a/b$, *platí* $\kappa(R'_+) \leq \kappa(R)$.

Důkaz Označme $z = \tilde{G}^{1/2}d$, takže $y = \tilde{G}^{1/2}z$. Použijeme-li (116), můžeme psát

$$\frac{1}{\gamma}R'_+ = R + \frac{\rho}{\gamma b}zz^T - \frac{1}{a}Rz(Rz)^T + \frac{\eta}{a}\left(\frac{a}{b}z - Rz\right)\left(\frac{a}{b}z - Rz\right)^T,$$

kde $a = z^TRz$ a $b = z^Tz$. Transformací kvazinewtonovské podmínky dostaneme $R^+z = \rho z$, takže matice R^+ má vlastní číslo ρ příslušné vlastnímu vektoru z. Vlastní vektory $v \in R^n$ příslušné ostatním vlastním číslům $\lambda \neq \rho$ můžeme volit tak, aby $v^Tz = 0$ a $v^Tv = 1$. Potom

$$\frac{1}{\gamma}v^TR'_+v = v^TRv + \frac{\eta - 1}{a}(v^TRz)^2 \leq v^TRv$$

(neboť $\eta - 1 \leq 0$) a $\lambda = v^TR'_+v \leq \gamma v^TRv \leq \gamma\|R\|$. Můžeme tedy psát $\|R'_+\| \leq \max(\rho, \gamma\|R\|)$. Protože $a = z^TRz$ a $b = z^Tz$, platí $a/b = z^TRz/z^Tz \leq \|R\|$, takže pro $\rho/\gamma \leq a/b$ dostaneme $\rho \leq \gamma\|R\|$. Platí tedy $\|R'_+\| \leq \gamma\|R\|$. Nyní můžeme použít dualitu (poznámka 94) a provést stejnou úvahu pro matici $(R'_+)^{-1}$ (v tomto případě se používá nerovnost $\gamma/\rho \geq c/b$). Dostaneme tak $\|(R'_+)^{-1}\| \leq (1/\gamma)\|R^{-1}\|$. Spojením obou nerovností dostaneme dokazované tvrzení □

Poznámka 127 Podle věty 60 je vhodné volit podíl ρ/γ tak, aby platilo $b/c \leq \rho/\gamma \leq a/b$. V tomto případě platí $\mu \geq 0$ pro libovolnou hodnotu parametru η (poznámka 115). Metoda hodnosti 1 však vyžaduje, aby $0 < \rho/\gamma < b/c$ nebo $a/b < \rho/\gamma$ (poznámka 93), 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 130.

K volbě parametru η lze použít různé minimalizační principy. Nejvíce se ujal princip spočívající v minimalizaci čísla podmíněnosti matice $(1/\gamma)H^{-1/2}H_+H^{-1/2}$.

Lemma 21 *Nechť jsou splněny předpoklady lemmatu 11 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

$$\eta = \eta^{OC} = \frac{bc(\rho/\gamma - b/c)}{ac - b^2} > \eta^*. \tag{185}$$

Důkaz (a) Podle lemmatu 11 jsou čísla $\underline{\lambda}(\eta)$ a $\overline{\lambda}(\eta)$ vlastními čísly matice $(1/\gamma)H^{-1/2}H_+H^{-1/2} \triangleq T+\eta uu^T$ (použili jsme vztah (116)). Nechť $\eta_2 > \eta_1$. Je-li $v_1 \in R^n$, $\|v_1\| = 1$, vlastním vektorem matice $T + \eta_1 uu^T$ příslušným vlastnímu číslu $\overline{\lambda}(\eta_1)$, platí $\overline{\lambda}(\eta_1) = v_1^T(T + \eta_1 uu^T)v_1 < v_1^T(T + \eta_2 uu^T)v_1 \leq \overline{\lambda}(\eta_2)$. Je-li $v_2 \in R^n$, $\|v_2\| = 1$, vlastním vektorem matice $T + \eta_2 uu^T$ příslušným vlastnímu číslu $\underline{\lambda}(\eta_2)$, platí $\underline{\lambda}(\eta_2) = v_2^T(T + \eta_2 uu^T)v_2 > v_2^T(T + \eta_1 uu^T)v_2 \geq \overline{\lambda}(\eta_1)$.
(b) Podle (103) 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 tak jako v důkazu věty 45 platí

$$\sigma^2 - 4\delta = \left(\frac{\gamma a}{\rho b}\delta + \frac{\rho c}{\gamma b}\right)^2 - 4\delta = \left(\frac{\gamma a}{\rho b}\delta - \frac{\rho c}{\gamma b}\right)^2 + 4\frac{ac - b^2}{b^2}\delta > 0$$

(předpokládáme, že $\delta > 0$ a $ac - b^2 > 0$). Vyjádříme-li kořeny rovnice $\lambda^2 - \sigma\lambda + \delta = 0$ v explicitním tvaru a použijeme-li substituci

$$\omega = \frac{\sigma}{2\sqrt{\delta}},$$

dostaneme po rozšíření zlomku

$$\frac{\overline{\lambda}}{\underline{\lambda}} = \left(\omega + \sqrt{\omega^2 - 1}\right)^2.$$

Derivujeme-li tento podíl podle parametru η, dostaneme

$$\left(\frac{\overline{\lambda}}{\underline{\lambda}}\right)' = \frac{2\omega'}{\sqrt{\omega^2 - 1}}\left(\omega + \sqrt{\omega^2 - 1}\right)^2$$

(jmenovatel je stejně jako diskriminant rovnice $\lambda^2 - \sigma\lambda + \delta = 0$ kladný). Tento výraz je nulový právě tehdy, jestliže

$$\omega' = \frac{2\sigma'\delta - \sigma\delta'}{4\delta\sqrt{\delta}} = 0,$$

neboli $2\sigma'\delta - \sigma\delta' = 0$ (neboť $\delta > 0$). Použijeme-li výrazy uvedené v lemmatu 11, 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 61 *Nechť jsou splněny předpoklady lemmatu 21. 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, jestliže $\eta = \max(\eta^{R1}, \eta^{OC})$.

(b) Pokud $b/c \le \rho/\gamma \le a/b$, je $\kappa(\eta)$ minimální právě tehdy, jestliže $\eta = \eta^{OC}$.

(c) Pokud $a/b < \rho/\gamma$, je $\kappa(\eta)$ minimální právě tehdy, jestliže $\eta = \min(\eta^{R1}, \eta^{OC})$.

Důkaz Podle Lemmatu 11 platí

$$\kappa(\eta) = \frac{\max(1, \overline{\lambda}(\eta))}{\min(1, \underline{\lambda}(\eta))}. \tag{186}$$

Jestliže $\mu \ge 0$, podle (b) lemmatu 21 platí $0 < \underline{\lambda}(\eta) \le 1 \le \overline{\lambda}(\eta)$, takže podle (c) lemmatu 21 je $\kappa(\eta)$ minimální, pokud $\eta = \eta^{OC}$. Jestliže $\mu < 0$, lze podle (a) lemmatu 21 oba kořeny rovnice $\lambda^2 + \sigma\lambda + \delta = 0$ současně zvětšit nebo zmenšit změnou parametru η. Podíl (186) 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 93 a poznámky 115. □

Poznámka 128 Hodnota (185) je samoduální. Dosadíme-li ji do (134), dostaneme

$$\beta = \beta^{OC} = \frac{ab(\gamma/\rho - b/a)}{ac-b^2} > \beta^*. \tag{187}$$

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, jestliže $\beta = \max(\beta^{R1}, \beta^{OC})$.

(b) Pokud $b/a \le \gamma/\rho \le c/b$, je $\kappa(\beta)$ minimální právě tehdy, jestliže $\beta = \beta^{OC}$.

(c) Pokud $c/b < \gamma/\rho$, je $\kappa(\beta)$ minimální právě tehdy, jestliže $\beta = \min(\beta^{R1}, \beta^{BM})$.

Poznámka 129 Položíme-li $\delta = 1$, dostaneme použitím (130) hodnotu

$$\eta = \frac{ab(\gamma/\rho - b/a)}{ac-b^2} > \eta^*,$$

která je shodná s výrazem vystupujícím ve vzorci (185). Tato hodnota je samoduální. Dosadíme-li ji do (134), dostaneme

$$\beta = \frac{bc(\rho/\gamma - b/c)}{ac-b^2} > \beta^*,$$

což je zase výraz vystupující ve vzorci (187).

Poznámka 130 Větu 61 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 61 je však určení vhodného podílu ρ/γ pro danou hodnotu parametru η. V tomto případě z $\eta = \eta^{OC}$ plyne

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{b}{c}\left(1 - \frac{\eta}{\eta^*}\right)$$

(stejně jako v části (c) důkazu lemmatu 21 se lze přesvědčit, že tato hodnota minimalizuje podíl $\overline{\lambda}/\underline{\lambda}$ pro zadanou hodnotu parametru η).

(a) Pro metodu DFP je $\eta = 0$, takže je vhodné volit $\rho/\gamma = b/c$.

(b) Pro metodu BFGS je $\eta = 1$, takže je vhodné volit $\rho/\gamma = a/b$.

(c) Pro Hoshinovu metodu je $\eta = (\rho/\gamma)/(\rho/\gamma + a/b)$, takže je vhodné volit

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{a}{b}\frac{\rho/\gamma + b/c}{\rho/\gamma + a/b},$$

neboli $\rho/\gamma = \sqrt{a/c}$

(d) Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$. Této hodnotě odpovídá podle věty 61 podíl

$$\frac{\rho}{\gamma} = \frac{\eta(ac - b^2) + b^2}{bc} = \frac{a}{b}\frac{\rho/\gamma - b/c}{\rho/\gamma - a/b},$$

neboli

$$\rho/\gamma = \frac{a}{b}\left(1 \pm \sqrt{\frac{ac - b^2}{ac}}\right).$$

Menší z těchto hodnot je nevhodná, větší dává metodu, která patří mezi perfektní metody s proměnnou metrikou ale není omezená (platí $\eta > 1$).

(e) Zvolíme-li $\eta = 2$ dostaneme metodu která patří mezi perfektní metody s proměnnou metrikou ale není omezená. Hodnotě $\eta = 2$ odpovídá podle věty 61 podíl

$$\frac{\rho}{\gamma} = \frac{2(ac - b^2) + b^2}{bc} = \frac{a}{b}\left(1 + \frac{ac - b^2}{ac}\right) > \frac{a}{b}$$

(předpokládáme že $ac - b^2 > 0$). Hodnotě $\eta = 2$ odpovídá hodnota $\beta = -b^2/(2ac - b^2)$.

Pokud $\eta > 1$, platí podle věty 61 $\rho/\gamma > a/b$. Použití této hodnoty však není vhodné, neboť v tomto případě nejsou splněny předpoklady věty 60. 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 131 Větu 61 lze 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 podíl ρ/γ a definuje (pro neoptimální hodnotu podílu 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 metodu OS (Oren, Spedicato)

$$\eta = \frac{b}{b + \sqrt{ac}}.$$

Tato metoda 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

$$\eta = 1 + \frac{1}{\sqrt{1 - b^2/(ac)}}.$$

Tato metoda patří mezi perfektní metody s proměnnou metrikou ale není omezená (platí $\eta > 1$).

Existují další minimalizační postupy, kterými lze získat hodnotu parametru η. Jeden z nich je založen na použití nerovnosti mezi geometrickým a aritmetickým průměrem

$$\frac{1}{n}\mathrm{Tr}(X) = \frac{1}{n}\sum_{i=1}^{n}\lambda_i \geq \left(\prod_{i=1}^{n}\lambda_i\right)^{1/n} = (\det(X))^{1/n}$$

(lemma 2), která platí pro libovolnou symetrickou pozitivně definitní matici X s vlastními čísly $\lambda_i > 0$, $1 \leq i \leq n$. Protože rovnost nastává pouze tehdy, mají-li všechna vlastní čísla stejnou hodnotu, je účelné minimalizovat funkci

$$\frac{\mathrm{Tr}(\gamma B^{-1/2}B_+B^{-1/2})}{n(\det(\gamma B^{-1/2}B_+B^{-1/2}))^{1/n}} = \frac{n - 2 + \sigma/\delta}{n\,(1/\delta)^{1/n}}, \tag{188}$$

kde σ a δ jsou čísla určená podle vzorců (129) a (130).

Věta 62 *Hodnota β minimalizuje funkci (188) 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 (143) a (145), 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

$$1 - \frac{\beta}{\beta^*} = \frac{n - 2 + (\gamma/\rho)(a/b)}{n - 1} = 1 + \frac{(\gamma/\rho)(a/b) - 1}{n - 1}, \tag{189}$$

odkud plyne tvrzení věty. □

Poznámka 132 Hodnota β uvedená ve větě 62 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 (140) a (189) (je však vhodné nahradit tuto hodnotu nulou, vyjde-li záporná).

Odvodíme ještě jednu hodnotu patametru 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 124). 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.

Lemma 22 *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 (133), 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 $\tilde{B} = G^{-1/2}BG^{-1/2}$, $\tilde{d} = G^{1/2}d$, $\tilde{y} = G^{-1/2}y$ a $\tilde{v} = G^{-1/2}v$, což spolu s (139) 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.

Lemma 23 *Zvolíme-li $G^{-1} = \lambda H$, dostaneme*

$$\beta = \frac{b^2}{ac - b^2}\left(\frac{1}{\lambda} - 1\right), \tag{190}$$

$$\eta = \frac{\lambda ac - b^2}{ac - b^2} \tag{191}$$

Důkaz Zvolíme-li $G^{-1} = \lambda H$, můžeme psát

$$\frac{v^T G^{-1}v}{c} = \frac{\lambda}{c}((c/b)y - Bd)^T H((c/b)y - Bd) = \frac{\lambda}{c}\left(\frac{c^2}{b^2}a - 2\frac{c}{b}b + c\right) = \lambda\frac{ac - b^2}{b^2}.$$

Použijeme-li lemma 22 a (141) dostaneme $1 - \beta/\beta^* = 1/\lambda$, což spolu s (140) dává (190) a (191). □

Poznámka 133 Dobré výsledky dostaneme, položíme-li $\lambda = \sqrt{c/a}$. Poněkud horší výsledky dávají volby $\lambda = b/a$ a $\lambda = c/b$. Hodnota (191) je definována pokud $ac - b^2 > 0$, v opačném případě pokládáme $\eta = 1$. Pokud je hodnota (191) 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 $B_+d = (1/\rho)y$ se vypočte hodnota $\rho = d^T y/d^T B_+ d$. Tato hodnota 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

$$A = \frac{F_+ - F}{d^T g}, \qquad B = \frac{d^T g_+}{d^T g}.$$

Poznámka 134 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 $d^T B_+ d$, dostaneme

$$d^T B_+ d = 2(F - F_+ + d^T g_+),$$

takže

$$\rho = \frac{d^T y}{2(F - F_+ + d^T g_+)} = \frac{B-1}{2(B-A)}. \tag{192}$$

Poznámka 135 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

$$\begin{aligned} 6(F - F_+ + d^T g_+) &= 3 d^T B_+ d - d^T (T_+ d) d, \\ 2(d^T g - d^T g_+) &= -2 d^T B_+ d + d^T (T_+ d) d, \end{aligned}$$

což po sečtení dává

$$d^T B_+ d = 6(F - F_+) + 4 d^T g_+ + 2 d^T g,$$

takže

$$\rho = \frac{d^T y}{6(F - F_+) + 4 d^T g_+ + 2 d^T g} = \frac{B-1}{4B - 6A + 2} = \frac{B-1}{4(B-1) - 6(A-1)}. \tag{193}$$

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

$$\rho = \frac{\varphi'(\alpha) - \varphi'(0)}{\alpha \varphi''(\alpha)}.$$

Zvolíme-li vhodný tvar funkce $\varphi(\alpha)$ a spočteme-li $\varphi'(\alpha)$, $\varphi''(\alpha)$, můžeme podle předchozího vzorce určit odpovídající hodnotu parametru ρ.

Poznámka 136 Jednou z možností 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 $\varphi'(0)$, $\varphi'(\alpha)$. Po dosazení těchto hodnot do příslušných výrazů a po formálních úpravavách dostaneme vzorec (193), stejný jako v případě použití čtyř členů zpětného Taylorova rozvoje.

Poznámka 137 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 63 *Uvažujme homogenní model $\varphi(\alpha) = a\alpha^r + b\alpha + c$ s $r > 1$. Pak platí*

$$\rho = \frac{A-1}{B-A}, \tag{194}$$

kde

$$\begin{aligned} A &= \frac{\varphi(\alpha)-\varphi(0)}{\alpha\varphi'(0)} = \frac{F_+ - F}{d^T g}, \\ B &= \frac{\varphi'(\alpha)}{\varphi'(0)} = \frac{d^T g_+}{d^T g}. \end{aligned}$$

Pokud pro hodnotu (194) platí $\rho > 0$, odpovídá tato hodnota homogennímu modelu s $r > 1$.

Důkaz Zřejmě

$$\begin{aligned} \varphi'(\alpha) &= ar\alpha^{r-1} + b, \\ \varphi''(\alpha) &= ar(r-1)\alpha^{r-2}, \end{aligned}$$

takže pro $r > 1$ platí

$$\begin{aligned} B - 1 &= \frac{\varphi'(\alpha)}{\varphi'(0)} - 1 = \frac{ar\alpha^{r-1} + b}{b} - 1 = \frac{ar\alpha^{r-1}}{b}, \\ A - 1 &= \frac{\varphi(\alpha)-\varphi(0)}{\alpha\varphi'(0)} - 1 = \frac{a\alpha^r + b\alpha}{\alpha b} - 1 = \frac{a\alpha^{r-1}}{b}, \end{aligned}$$

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 výrazu pro ρ (poznámka 137) 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 $\rho > 0$ implikuje $r > 1$. Použijeme-li (S3b), dostaneme $\varphi'(\alpha) \geq \varepsilon_2\varphi'(0)$, kde $0 < \varepsilon_2 < 1$ a $\varphi'(0) < 0$, takže $B \leq \varepsilon_2 < 1$ a $B - 1 < 0$. Pokud

$$\frac{1}{r} = \frac{A-1}{B-1} \geq 1,$$

pak $A - 1 \leq B - 1$, takže $B - A \geq 0$ a $\rho \leq 0$. □

Parametr γ slouží ke škálování matice H, neboť aktualizace (116) s $\gamma \neq 1$ je ekvivalentní aktualizaci (116) s $\gamma = 1$ aplikované na matici γH. Důvod pro škálování poskytuje věta 60 a následující věta.

Věta 64 *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 138 Zvolíme-li $T = G^{-1/2}$, platí $\tilde{G} = T^TGT = I$. 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 $T\tilde{H}_1T^T$) 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 $\gamma Hy \approx \rho d$, 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 139 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 145). 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. 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. 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. 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 $F_1 = F(x+\alpha_1s)$, $g_1 = g(x + \alpha_1s)$, $\lambda_1 = s^Tg_1/s^Tg$, 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).

4.5 Globální konvergence

Nyní se budeme zabývat globální konvergencí metod s proměnnou metrikou. Důkaz globální konvergence vyžaduje silnější předpoklady než tomu bylo v případě metod sdružených gradientů. Potřebujeme aby minimalizovaná funkce byla stejnoměrně konvexní (podmínka (F4)) a navíc se globální konvergence dá dokázat pouze pro perfektní metody z Broydenovy třídy takové, že $(1-\lambda)\beta_i^* \leq \beta_i \leq 1-\lambda$, kde $0 < \lambda < 1$.

Lemma 24 *Uvažujme metodu s proměnnou metrikou z Broydenovy třídy takovou, že $\underline{\gamma} \leq \gamma_i \leq \overline{\gamma}$, $\underline{\rho} \leq \rho_i \leq \overline{\rho}$ a $(1-\lambda)\beta_i^* \leq \beta_i \leq 1-\lambda$, kde $0 < \lambda < 1$. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (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 (133) můžeme po roznásobení zapsat takto

$$B_{i+1} = \frac{1}{\gamma_i}\left(B_i + \frac{\gamma_i}{\rho_i}\frac{y_i y_i^T}{y_i^T d_i} + \beta_i \frac{d_i^T B_i d_i}{y_i^T d_i}\frac{y_i y_i^T}{y_i^T d_i} - \frac{\beta_i}{y_i^T d_i}(B_i d_i y_i^T + y_i (B_i d_i)^T) + \frac{\beta_i - 1}{d_i^T B_i d_i} B_i d_i (B_i d_i)^T\right).$$

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

$$\begin{aligned} Tr\, B_{i+1} &= \frac{1}{\gamma_i}\left(Tr\, B_i + \frac{\gamma_i}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i)\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i}\right) \\ &\leq \frac{1}{\underline{\gamma}}\left(Tr\, B_i + \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T d_i}{y_i^T d_i}\|B_i\| + 2\frac{\|y_i\|\|d_i\|}{y_i^T d_i}\|B_i\|\right) + \frac{1}{\underline{\rho}}\frac{y_i^T y_i}{y_i^T d_i} \end{aligned}$$

(neboť $\beta_i \leq 1$). Protože $y_i = \tilde{G}_i d_i$ (věta 60), kde matice $\tilde{G}_i$ vyhovuje nerovnostem v podmínkách (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 nerovností $\|B_i\| \leq Tr\, B_i$ do vztahu pro $Tr\, B_{i+1}$, dostaneme

$$\begin{aligned} Tr\, B_{i+1} &\leq Tr\, B_{i+1} + 1 \leq \frac{1}{\underline{\gamma}}\left(1 + \frac{\overline{G}}{\underline{G}} + 2\sqrt{\frac{\overline{G}}{\underline{G}}}\right) Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + 1 \\ &\leq \overline{K}(Tr\, B_i + 1) \leq \overline{K}^i (Tr\, B_1 + 1) \leq \overline{C}^i, \end{aligned} \tag{195}$$

kde $\overline{K} = \max\left(\left(1 + \overline{G}/\underline{G} + 2\sqrt{\overline{G}/\underline{G}}\right)/\underline{\gamma},\ \overline{G}/\underline{\rho} + 1\right)$ a $\overline{C} = \overline{K}(Tr\, B_1 + 1)$.

(b) Použijeme-li vztah pro $\det B_{i+1}$ (poznámka 95), můžeme psát

$$\frac{\det B_{i+1}}{\det B_i} = \left(\frac{1}{\gamma_i}\right)^n \frac{\gamma_i}{\rho_i}\frac{b_i}{c_i}\left(1 - \frac{\beta_i}{\beta_i^*}\right) \geq \left(\frac{1}{\overline{\gamma}}\right)^n \frac{\underline{\gamma}}{\overline{\rho}}\frac{b_i}{c_i}\lambda = \underline{K}\frac{b_i}{c_i}, \tag{196}$$

kde $\underline{K} = (1/\overline{\gamma})^n(\underline{\gamma}/\overline{\rho})\lambda$. Platí tedy

$$\frac{\det B_{i+1}}{\det B_1} \geq \underline{K}^i \prod_{j=1}^{i} \frac{b_j}{c_j}$$

a protože $\det H_{i+1} = 1/\det B_{i+1}$, můžeme psát

$$\det H_{i+1} \leq \frac{\det H_1}{\underline{K}^i} \prod_{j=1}^{i} \frac{c_j}{b_j} \leq \frac{1}{K^i} \prod_{j=1}^{i} \frac{c_j}{b_j},$$

kde $K = \underline{K}/(\det H_1 + 1)$. Podle (a) platí

$$\det B_{i+1} \leq \left(\frac{Tr\, B_{i+1}}{n}\right)^n \leq (Tr\, B_{i+1})^n \leq \overline{C}^{in} \triangleq C^i$$

(používáme nerovnost mezi geometrickým a aritmetickým průměrem (8)), 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

$$\prod_{j=1}^{i} \frac{c_j}{b_j} \geq \left(\frac{K}{C}\right)^i \triangleq \underline{C}^i.$$

Použijeme-li znovu nerovnost mezi geometrickým a aritmetickým průměrem, dostaneme

$$\sum_{j=1}^{i} \frac{c_j}{b_j} \geq i\left(\prod_{j=1}^{i} \frac{c_j}{b_j}\right)^{1/i} \geq i\underline{C}.$$

□

Věta 65 *(Globální konvergence) Nechť jsou splněny předpoklady lemmatu 24, přičemž $\gamma_i \geq 1\ \forall i \in N$. Pak*

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Důkaz Použijeme opět základní vztah pro $Tr\, B_{i+1}$ uvedený na začátku důkazu lemmatu 24. Protože $y_i = \tilde{G}_i d_i$ a $B_i d_i = -\alpha_i g_i$, můžeme psát

$$\begin{aligned}
\frac{|y_i^T B_i d_i|}{y_i^T d_i} &= \frac{|y_i^T B_i d_i|}{d_i^T B_i d_i}\frac{c_i}{b_i} \leq \frac{\|\tilde{G} d_i\|\|\alpha_i g_i\|}{-\alpha_i d_i^T g_i}\frac{c_i}{b_i} \leq \frac{\overline{G}}{\cos\theta_i}\frac{c_i}{b_i}, \\
\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} &= \frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i}\frac{y_i^T d_i}{d_i^T B_i d_i}\frac{c_i}{b_i} \geq \frac{\alpha_i^2\|g_i\|^2\underline{G}\|d_i\|^2}{\alpha_i^2(d_i^T g_i)^2}\frac{c_i}{b_i} = \frac{\underline{G}}{\cos^2\theta_i}\frac{c_i}{b_i},
\end{aligned}$$

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

$$\begin{aligned} Tr\, B_{i+1} &\leq Tr\, B_i + \frac{1}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i)\frac{1}{\overline{\gamma}}\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \\ &\leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \left(\overline{G} + 2\frac{\overline{G}}{\cos\theta_i} - \frac{\lambda \underline{G}}{\overline{\gamma}\cos^2\theta_i}\right)\frac{c_i}{b_i} = Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \xi_i \frac{c_i}{b_i}, \end{aligned}$$

kde $\xi_i = \overline{G} + 2\overline{G}/\cos\theta_i - \lambda\underline{G}/(\overline{\gamma}\cos^2\theta_i)$. Přepokládejme nyní, že $\liminf_{i\to\infty}\|g_i\| > 0$. Pak podle věty 9 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 předchozí nerovnosti a podle (b) lemmatu 24 platí

$$Tr\, B_{i+1} \leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \xi_i \frac{c_i}{b_i} \leq Tr\, B_1 + i\frac{\overline{G}}{\underline{\rho}} - 2\frac{\overline{G}}{\underline{\rho}\underline{C}}\sum_{j=1}^{i}\frac{c_j}{b_j} \leq Tr\, B_1 - i\frac{\overline{G}}{\underline{\rho}}.$$

Zvolíme-li index i tak aby platilo $i > Tr\, B_1 \underline{\rho}/\overline{G}$, dostaneme $Tr\, B_{i+1} < 0$, což je spor, neboť stopa symetrické pozitivně definití matice je kladná. □

Poznámka 140 Věta 65 teoreticky zdůvodňuje špatné konvergenční vlastnosti metody DFP (s nepřesným výběrem délky kroku). Metoda DFP odpovídá volbě $\beta_i = 1$ $\forall i \in N$, takže není splněn předpoklad $\beta_i \leq 1 - \lambda < 1$ $\forall i \in N$. Ze vztahu pro $Tr\, B_{i+1}$ vymizí poslední člen a nelze použít princip důkazu.

Poznámka 141 Věta 65 je nejobecnějším tvrzením o globální konvergenci (nemodifikovaných a nerestartovaných) metod s proměnnou metrikou. Tato věta vyžaduje aby byla splněna podmínka (F5) (existence konstanty $\underline{G} > 0$), takže ji lze použít pouze pro konvexní funkce. Z důkazu věty 65 je zřejmé, že tento požadavek slouží pouze k tomu, aby platilo

$$y_i^T d_i \geq \underline{G}\|d_i\|^2. \tag{197}$$

Jednou z možností jak tento požadavek odstranit, je zvolit malé číslo $\underline{\tau} > 0$ a matici B_i aktualizovat pouze tehdy, je-li splněna podmínka (197) s $\underline{G} = \underline{\tau}$. Jiná úprava spočívá v náhradě vektoru $y_i = g_{i+1} - g_i$ 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 (S2a) a (S3a) s $\varepsilon_3 = \infty$, 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

$$\frac{\tilde{y}_i^T \tilde{y}_i}{\tilde{y}_i^T d_i} = \frac{\tilde{y}_i^T \tilde{y}_i}{\tilde{y}_i^T (\tilde{G}_i + \tau_i I)^{-1}\tilde{y}_i} \leq \overline{G} + \underline{\tau}.$$

Jsou tedy splněny všechny nerovnosti, které se používají v důkazu věty 65. Místo konstanty $\underline{\tau}$ můžeme použít proměnnou hodnotu $\underline{\tau}_i = \overline{\tau}\min(1, \|g_i\|)$, kde $\overline{\tau} > 0$. Protože důkaz věty 65 provádíme sporem a předpokládáme, že $\|g_i\| > \underline{\varepsilon}$, platí v tomto případě $\underline{\tau}_i \geq \overline{\tau}\min(1, \underline{\varepsilon}) \stackrel{\Delta}{=} \underline{\tau}$.

4.6 Asymptotická rychlost konvergence

Nyní se budeme zabývat rychlostí konvergence metod s proměnnou metrikou. K tomuto účelu není vhodné používat matice H_i a B_i, neboť nastávají potíže s komutativitou. Lze však použít matice R_i zavedené ve větě 60 nebo transformaci proměnných studovanou ve větě 64. Protože metody s proměnnou metrikou jsou

invariantní vůči této transformaci, zachovává se při ní R-lineární i Q-superlineární rychlost konvergence. Matici T budeme volit tak, že $T = (G^*)^{-1/2}$, takže po transformaci platí $G^* = I$. Budeme opět používat označení

$$\tilde{G}_i = \int_0^1 G(x_i + \lambda d_i)\mathrm{d}\lambda.$$

Věta 66 *(Lineární konvergence) Nechť jsou splněny předpoklady lemmatu 24, přičemž $\gamma_i \geq 1\ \forall i \in N$. Nechť $x_i \to x^*$ a $G^* = I$. Pak platí*

$$\sum_{i=1}^{\infty} \|e_i\| < \infty.$$

Důkaz Jelikož $x_i \to x^*$ (a tedy $d_i \to 0$), platí $\tilde{G}_i \to G^* = I$, takže pro libovolné číslo $0 < \varepsilon < 1$ existuje index $k \in N$ takový, že $\|\tilde{G}_i - I\| \leq \varepsilon\ \forall i \geq k$. Pak pro $i \geq k$ můžeme psát

$$\begin{aligned}
\frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\frac{y_i^T B_i d_i}{y_i^T d_i} &= \left(\frac{d_i^T (I + (\tilde{G}_i - I))^2 d_i}{d_i^T (I + (\tilde{G}_i - I)) d_i} - 2\right)\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\frac{d_i^T (\tilde{G}_i - I) B_i d_i}{y_i^T d_i} \\
&\leq (\|\tilde{G}_i - I\| - 1)\frac{d_i^T B_i d_i}{y_i^T d_i} + 2\|\tilde{G}_i - I\|\frac{\|d_i\|\|B_i d_i\|}{d_i^T B_i d_i}\frac{c_i}{b_i} \leq \frac{2\varepsilon}{\cos\theta_i}\frac{c_i}{b_i}
\end{aligned}$$

(první člen je záporný, neboť $\|\tilde{G}_i - I\| \leq \varepsilon < 1$). Zvolme číslo $0 < \varepsilon < 1$ tak, aby platilo $\varepsilon \leq \lambda\underline{G}/(4\overline{\gamma})$ a předpokládejme bez újmy na obecnosti, že $k = 1$ (v opačném případě můžeme provést přečíslování indexů). Pak po dosazení do vztahu pro $Tr\, B_{i+1}$ uvedeného v důkazu věty 65 dostaneme

$$\begin{aligned}
Tr\, B_{i+1} &\leq Tr\, B_i + \frac{1}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - 2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1 - \beta_i)\frac{1}{\overline{\gamma}}\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} \\
&\leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \left(2\varepsilon\cos\theta_i - \frac{\lambda\underline{G}}{\overline{\gamma}}\right)\frac{1}{\cos^2\theta_i}\frac{c_i}{b_i} \leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} - \frac{\lambda\underline{G}}{2\overline{\gamma}}\frac{1}{\cos^2\theta_i}\frac{c_i}{b_i} \\
&\leq Tr(B_1) + i\frac{\overline{G}}{\underline{\rho}} - \frac{\lambda\underline{G}}{2\overline{\gamma}}\sum_{j=1}^{i}\frac{1}{\cos^2\theta_j^2}\frac{c_j}{b_j},
\end{aligned}$$

takže

$$\sum_{j=1}^{i}\frac{1}{\cos^2\theta_j^2}\frac{c_j}{b_j} \leq \frac{2\overline{\gamma}}{\lambda\underline{G}}\left(Tr(B_1) + i\frac{\overline{G}}{\underline{\rho}}\right) \leq Li,$$

kde $L = 2\overline{\gamma}(\overline{G}/\underline{\rho} + Tr\, B_1 + 1)/(\lambda\underline{G})$. Použijeme-li nerovnost mezi geometrickým a aritmetickým průměrem (8), dostaneme

$$\prod_{j=1}^{i}\frac{1}{\cos^2\theta_j^2}\frac{c_j}{b_j} \leq \left(\frac{1}{i}\sum_{j=1}^{i}\frac{1}{\cos^2\theta_j^2}\frac{c_j}{b_j}\right)^i = L^i$$

a podle (b) lemmatu 24 platí

$$\prod_{j=1}^{i}\cos^2\theta_j^2 \geq \frac{1}{L^i}\prod_{j=1}^{i}\frac{c_j}{b_j} \geq \frac{\underline{K}^i}{L^i} \triangleq \underline{c}^i.$$

Použijeme-li ještě jednou nerovnost mezi geometrickým a aritmetickým průměrem, můžeme psát

$$\sum_{j=1}^{i} \cos^2 \theta_j^2 \geq i \left(\prod_{j=1}^{i} \cos^2 \theta_j^2 \right)^{1/i} = i\underline{c},$$

takže dokazované tvrzení plyne z věty 14 a poznámky 32. □

V dalších úvahách budeme používat princip omezeného znehodnocení zformulovaný v následujícím lemmatu.

Lemma 25 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost, která konverguje R-lineárně k bodu $x^* \in R^n$ 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 $\kappa_{i+1} \leq \kappa_i(1 + C\|e_i\|)$. Pak existuje konstanta $\overline{C} > 0$ taková, že $\kappa_i \leq \kappa_1 \exp(\overline{C})$ $\forall i \in N$.*

Důkaz Podle předpokladu platí

$$\kappa_{i+1} \leq \kappa_1 \prod_{j=1}^{i}(1 + C\|e_j\|) \leq \kappa_1 \left(\frac{1}{i} \sum_{j=1}^{i}(1 + C\|e_j\|) \right)^i = \kappa_1 \left(1 + \frac{C}{i} \sum_{j=1}^{i} \|e_j\| \right)^i$$

(používáme nerovnost mezi geometrickým a aritmetickým průměrem (8)). Jelikož z R-lineární konvergence plyne existence konstanty $\overline{C}$ takové, že

$$\sum_{j=1}^{i} \|e_j\| \leq \sum_{j=1}^{\infty} \|e_j\| = \frac{\overline{C}}{C}$$

(poznámka 32), můžeme psát

$$\kappa_{i+1} \leq \kappa_1 \left(1 + \frac{\overline{C}}{i} \right)^i.$$

$\forall i \in N$. 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 $\kappa_i \leq \kappa_1 \exp(\overline{C})$ $\forall i \in N$. □

Poznámka 142 Podle lemmatu 25 je posloupnost $\kappa_i \in R^n$, $i \in N$, shora omezená, pokud $x_i \to x^*$ R-lineárně a $\kappa_{i+1} = \kappa_i(1+O(\|e_i\|))$ $\forall i \in N$. Platí to i tehdy pokud $\kappa_{i+1} = \kappa_i(1+O(\|e_i\|))+O(\|e_i\|)$ $\forall i \in N$, neboť v tomto případě $\kappa_{i+1} + 1 = (\kappa_i + 1)(1 + O(\|e_i\|))$ a podle lemmatu 25 existuje konstanta $\overline{C}$ taková, že $\kappa_i < \kappa_i + 1 \leq (\kappa_1 + 1)\exp(\overline{C})$ $\forall i \in N$.

Nyní dokážeme větu o superlineární konvergenci metod s proměnnou metrikou. Jelikož superlineární konvergence vyžaduje aby v jistém smyslu platilo $B_i \to G^*$ (věta 16), budeme požadovat splnění předpokladů věty 42 ($\rho_i = 1$, $\gamma_i = 1$, $\forall i \in N$). Dále budeme používat označení

$$R_i = \tilde{G}_i^{1/2} H_i \tilde{G}_i^{1/2}, \quad R'_{i+1} = \tilde{G}_i^{1/2} H_{i+1} \tilde{G}_i^{1/2},$$

kde $\tilde{G}_i$ je matice definovaná ve větě 60 (takže $\tilde{G}_i d_i = y_i$). Poznamenejme, že $R_{i+1} = \tilde{G}_{i+1}^{1/2} H_{i+1} \tilde{G}_{i+1}^{1/2} \neq R'_{i+1}$.

Poznámka 143 Ke kvantitativnímu vyšetřování maticových rekurentních vztahů lze z výhodou použít Frobeniovu normu $\|A\|_F = \sqrt{Tr(A^T A)}$. Z této definice plyne, že $\|A\| \leq \|A\|_F \leq \sqrt{n}\|A\|$. Využijeme toho, že pro libovolné matice A, B platí $\|A\|_F^2 = Tr(A^T A)$, $\|B\|_F^2 = Tr(B^T B)$, takže $\|A + B\|^2 = Tr((A + B)^T(A + B)) = \|A\|_F^2 + \|B\|_F^2 + 2Tr(A^T B)$. Dále platí

$$\|uv^T\|_F^2 = Tr(vu^T uv^T) = u^T u Tr(vv^T) = u^T uv^T v.$$

Je-li matice A symetrická, je $\|A\|_F^2$ součtem druhých mocnin jejích vlastních čísel. Z toho plyne, že symetrické matice, které mají stejná vlastní čísla, mají stejnou Frobeniovu normu.

Lemma 26 *Uvažujme aktualizaci*

$$R'_+ = R + \frac{zz^T}{z^Tz} - \frac{Rz(Rz)^T}{z^TRz} + \frac{\eta}{z^TRz}\left(\frac{z^TRz}{z^Tz}z - Rz\right)\left(\frac{z^TRz}{z^Tz}z - Rz\right)^T.$$

Pak platí

$$\begin{aligned}\|R'_+ - I\|_F^2 \;=\; \|R-I\|_F^2 \;&-\; (1-\eta)\left(\left(1-\frac{z^TR^2z}{z^TRz}\right)^2 + 2\left(\frac{z^TR^3z}{z^TRz} - \left(\frac{z^TR^2z}{z^TRz}\right)^2\right)\right)\\ &-\; \eta\left(\left(1-\frac{z^TRz}{z^Tz}\right)^2 + 2\eta\left(\frac{z^TR^2z}{z^Tz} - \left(\frac{z^TRz}{z^Tz}\right)^2\right)\right)\\ &-\; \eta(1-\eta)\left(\left(\frac{z^TR^2z}{z^TRz}\right)^2 - \left(\frac{z^TRz}{z^Tz}\right)^2\right).\end{aligned}$$

Důkaz Aplikujeme-li pravidla uvedená v poznámce 143 na vztah

$$R'_+ - I = R - I + \frac{zz^T}{z^Tz} + \eta\frac{z^TRz}{z^Tz}\frac{zz^T}{z^Tz} - \eta\frac{zz^TR}{z^Tz} - \eta\frac{Rzz^T}{z^Tz} + (\eta-1)\frac{Rzz^TR}{z^TRz},$$

získaný úpravou aktualizace z lemmatu 26, dostaneme

$$\begin{aligned}\|R'_+ - I\|_F^2 = \|R-I\|_F^2 \;&+\; 1 + \eta^2\left(\frac{z^TRz}{z^Tz}\right)^2 + 2\eta^2\frac{z^TR^2z}{z^Tz} + (\eta-1)^2\left(\frac{z^TR^2z}{z^TRz}\right)^2\\ &+\; 2\frac{z^TRz}{z^Tz} + 2\eta\left(\frac{z^TRz}{z^Tz}\right)^2 - 4\eta\frac{z^TR^2z}{z^Tz} + 2(\eta-1)\frac{z^TR^3z}{z^TRz}\\ &-\; 2 - 2\eta\frac{z^TRz}{z^Tz} + 4\eta\frac{z^TRz}{z^Tz} - 2(\eta-1)\frac{z^TR^2z}{z^TRz} + 2\eta\frac{z^TRz}{z^Tz}\\ &-\; 4\eta\frac{z^TRz}{z^Tz} + 2(\eta-1)\frac{z^TRz}{z^Tz} - 4\eta^2\left(\frac{z^TRz}{z^Tz}\right)^2\\ &+\; 2\eta(\eta-1)\left(\frac{z^TRz}{z^Tz}\right)^2 + 2\eta^2\left(\frac{z^TRz}{z^Tz}\right)^2 - 4\eta(\eta-1)\frac{z^TR^2z}{z^Tz}\\ = \|R-I\|_F^2 \;&-\; 1 + 2\eta\frac{z^TRz}{z^Tz} - 2\eta^2\frac{z^TR^2z}{z^Tz} - 2(\eta-1)\frac{z^TR^2z}{z^TRz}\\ &+\; 2(\eta-1)\frac{z^TR^3z}{z^TRz} + \eta^2\left(\frac{z^TRz}{z^Tz}\right)^2 + (\eta-1)^2\left(\frac{z^TR^2z}{z^TRz}\right)^2.\end{aligned}$$

Stejný výsledek dostaneme roznásobením vztahu uvedeného v lemmatu 26. □

Důsledek 11 *Jsou-li splněny předpoklady lemmatu 26 s* $0 \le \eta \le 1$, *platí* $\|R'_+ - I\|_F \le \|R-I\|_F$.

Důkaz Použijeme-li Schwarzovu nerovnost, dostaneme

$$\begin{aligned}\frac{z^TR^3z}{z^TRz} - \left(\frac{z^TR^2z}{z^TRz}\right)^2 &= \frac{z^TR^3zz^TRz - (z^TR^2z)^2}{(z^TRz)^2} \ge 0,\\ \frac{z^TR^2z}{z^Tz} - \left(\frac{z^TRz}{z^Tz}\right)^2 &= \frac{z^TRzz^Tz - (z^TRz)^2}{(z^Tz)^2} \ge 0,\\ \left(\frac{z^TR^2z}{z^TRz}\right)^2 - \left(\frac{z^TRz}{z^Tz}\right)^2 &= \frac{(z^TR^2zz^Tz)^2 - (z^TRz)^4}{(z^TRzz^Tz)^2}\\ &= \frac{z^TR^2zz^Tz + (z^TRz)^2}{z^TRzz^Tz}\,\frac{z^TR^2zz^Tz - (z^TRz)^2}{z^TRzz^Tz} \ge 0.\end{aligned}$$

Všechny závorky ve vztahu pro $\|R'_+ - I\|_F^2$ v lemmatu 26 jsou tedy nezáporné a jelikož $0 \leq \eta \leq 1$, platí $\|R'_+ - I\|_F^2 \leq \|R - I\|_F^2$. □

Lemma 27 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s proměnnou metrikou z Broydenovy třídy s $\rho_i = 1$, $\gamma_i = 1$ a $0 \leq \eta_i \leq 1$ taková, že $x_i \to x^*$, kde x^* je stacionárním bodem funkce $F : \mathcal{D} \to R$ splňující podmínky (F3)–(F5). Pak platí*

$$\|R_{i+1} - I\|_F = \|R_i - I\|_F(1 + O(\|e_i\|) + O(\|e_i\|).$$

Důkaz Označme

$$\tilde{R}_i = H_i^{1/2}\tilde{G}_i H_i^{1/2} \quad \tilde{R}'_{i+1} = H_{i+1}^{1/2}\tilde{G}_i H_{i+1}^{1/2}.$$

Matice $\tilde{R}_i$ má stejná vlastní čísla jako matice R_i, neboť z $\tilde{G}_i^{1/2} H_i \tilde{G}_i^{1/2} x = \lambda x$, $x \neq 0$, plyne $H_i^{1/2}\tilde{G}_i H^{1/2} y = \lambda y$, $y = H_i^{1/2}\tilde{G}_i^{1/2} x \neq 0$. 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}. Matici R'_{i+1} získáme z matice R_i pomocí aktualizace uvedené v lemmatu 26 (viz důkaz věty 60). Použijeme-li důsledek 11 dostaneme

$$\begin{aligned} \|R_{i+1} - I\|_F &= \|\tilde{R}_{i+1} - I\|_F \leq \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|\tilde{R}'_{i+1} - I\|_F \\ &= \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R'_{i+1} - I\|_F \leq \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R_i - I\|_F. \end{aligned}$$

Stačí tedy dokázat, že $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F = (\|R_i - I\|_F + 1)O(\|e_i\|)$. Použijeme-li definiční vztah pro matici $\tilde{G}_i$ uvedený ve větě 60 a nerovnost (F5), můžeme psát

$$\begin{aligned} \|\tilde{G}_{i+1} - \tilde{G}_i\| &= \|\int_0^1 G(x_{i+1} + \lambda d_{i+1})\mathrm{d}\lambda - \int_0^1 G(x_i + \lambda d_i)\mathrm{d}\lambda\| \\ &\leq \int_0^1 \|G(x_{i+1} + \lambda d_{i+1}) - G(x_i + \lambda d_i)\|\mathrm{d}\lambda \\ &\leq \overline{L}\int_0^1 \|e_{i+1} + \lambda d_{i+1} - e_i - \lambda d_i\|\mathrm{d}\lambda \\ &\leq \overline{L}\left(\|e_{i+1}\| + \|e_i\| + \frac{1}{2}\|d_{i+1}\| + \frac{1}{2}\|d_i\|\right) = O(\|e_i\|), \end{aligned}$$

neboť podle poznámky 31 platí $\|e_{i+1}\| = O(\|e_i\|)$ a $\|d_i\| = O(\|e_i\|)$. Platí tedy

$$\begin{aligned} \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| &\leq \|H_{i+1}^{1/2}(\tilde{G}_{i+1} - \tilde{G}_i)H_{i+1}^{1/2}\| \leq \|H_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \\ &= \|\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2} H_{i+1}\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \leq \|\tilde{G}_i^{-1}\|\|R'_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \\ &\leq \frac{1}{\underline{G}}\|R'_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| = \|R'_{i+1}\|O(\|e_i\|). \end{aligned}$$

Ale

$$\|R'_{i+1}\| = \|I + R'_{i+1} - I\| \leq 1 + \|R'_{i+1} - I\| \leq 1 + \|R'_{i+1} - I\|_F \leq 1 + \|R_i - I\|_F,$$

takže

$$\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F \leq \sqrt{n}\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| = \|R'_{i+1}\|O(\|e_i\|) = (\|R_i - I\|_F + 1)O(\|e_i\|).$$

□

Důsledek 12 *Jsou-li splněny předpoklady lemmatu 27, existují konstanty $\overline{R}$ a $\overline{H}$ takové, že $\|R_i\| \leq \overline{R}$ a $\|H_i\| \leq \overline{H}$ $\forall i \in N$.*

Důkaz Jelikož $\|R_i\|_F \leq \sqrt{n} + \|R_i - I\|_F$ a posloupnost $\|R_i - I\|_F$, $i \in N$, je podle lemmatu 27 a poznámky 142 omezená, je i posloupnost $\|R_i\| \leq \|R_i\|_F$, $i \in N$, omezená. Jelikož

$$\|H_i\| = \|\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2}H_i\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2}\| \leq \|\tilde{G}_i^{-1}\|\|R_i\| \leq \frac{1}{\underline{G}}\|R_i\|,$$

je i posloupnost $\|H_i\|$, $i \in N$, omezená. □

Poznámka 144 Aktualizaci pro R^{-1} dostaneme z aktualizace pro R záměnou $R \to R^{-1}$ a $\eta \to \beta$. Jelikož z $0 \leq \eta \leq 1$ plyne $0 \leq \beta \leq 1$ (poznámka 115), můžeme použít stejné úvahy jako v důkazu lemmatu 26 a lemmatu 27. Existují tedy konstanty $\underline{R}$ a $\underline{H}$ takové, že $\|R_i^{-1}\| \leq 1/\underline{R}$ a $\|H_i^{-1}\| \leq 1/\underline{H}$ $\forall i \in N$.

Věta 67 *(Superlineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s proměnnou metrikou z Broydenovy třídy s $\rho_i = 1$, $\gamma_i = 1$ a $0 \leq \eta_i \leq 1$, přičemž $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2) a (S3). Nechť $x_i \to x^*$, kde x^* je stacionárním bodem funkce $F : \mathcal{D} \to R$ splňující podmínky (F4)–(F6). Pak $x_i \to x^*$ superlineárně.*

Důkaz Z důkazu lemmatu 27 (první nerovnost) víme, že

$$\|R_i - I\|_F - \|R'_{i+1} - I\|_F \leq \|R_i - I\|_F - \|R_{i+1} - I\|_F + \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F,$$

kde $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F = (\|R_i - I\|_F + 1)O(\|e_i\|)$. Jelikož podle důsledku 12 je posloupnost $\|R_i - I\|_F$, $i \in N$, shora omezená, existuje konstanta C taková, že $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F \leq C\|e_i\|$. Použijeme-li větu 66, dostaneme

$$\sum_{i=1}^{\infty} \left(\|R_i - I\|_F - \|R'_{i+1} - I\|_F\right) \leq \|R_1 - I\|_F + C\sum_{i=1}^{\infty}\|e_i\| < \infty,$$

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 $\|R'_{i+1} - I\|_F \leq \|R_i - I\|_F$ jsou omezené, také

$$\lim_{i\to\infty} \left(\|R_i - I\|_F^2 - \|R'_{i+1} - I\|_F^2\right) = 0.$$

Nyní použijeme vztah uvedený v lemmatu 26. Protože poslední tři členy na pravé straně tohoto vztahu mají 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

$$\limsup_{i \overset{N_1}{\to} \infty} \eta_i < 1, \qquad \liminf_{i \overset{N_2}{\to} \infty} \eta_i > 0$$

(například $N_1 = \{i \in N : 0 \leq \eta_i \leq 1/2\}$, $N_2 = \{i \in N : 1/2 < \eta_i \leq 1\}$). 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 12 a poznámky 144 platí $\|R_i\| \leq \overline{R}$ a $\|R_i^{-1}\| \leq 1/\underline{R}$ $\forall i \in N$, můžeme obě tyto limity nahradit jedinou limitou

$$\lim_{i\to\infty} \frac{\|(R_i - I)z_i\|}{\|z_i\|} = \lim_{i\to\infty} \frac{\|(\tilde{G}_i^{1/2} H_i \tilde{G}_i^{1/2} - \tilde{G}_i^{-1/2}\tilde{G}_i^{1/2})z_i\|}{\|\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2} z_i\|} = \lim_{i\to\infty} \frac{\|(\tilde{G}_i^{1/2}(H_i - \tilde{G}_i^{-1})y_i\|}{\|\tilde{G}_i^{-1/2} y_i\|} = 0.$$

Protože $x_i \to x^*$ implikuje $\tilde{G}_i \to G^*$, dostaneme použitím předpokladů (F3) a (F4) vztah

$$\lim_{i\to\infty} \frac{\|(H_i - (G^*)^{-1})y_i\|}{\|y_i\|} = 0,$$

který je, vzhledem k tomu, že $H_i \leq \overline{H}$ a $H_i^{-1} \leq 1/\underline{H}$ $\forall i \in N$ (důsledek 12 a a poznámka 144), ekvivalentní vztahu

$$\lim_{i\to\infty} \frac{\|(B_i - G^*)d_i\|}{\|d_i\|} = 0.$$

Závěr důkazu plyne z věty 16. □

Poznámka 145 Věta 67 předpokládá, že platí $0 \leq \eta_i \leq 1$ (neboli $0 \leq \beta_i \leq 1$) $\forall i \in N$. Tento předpoklad nelze příliš zeslabit. Dá se pouze dokázat, že věta zůstane v platnosti, pokud $\beta_i \leq 1$ $\forall i \in N$ a

$$\sum_{\substack{i=1 \\ \beta_i<0}}^{\infty} \frac{\beta_i}{\beta_i^*} < \infty.$$

Také je nutné, aby platilo $\rho_i = \gamma_i = 1$, v opačném případě nelze použít princip důkazu. Podrobnějším rozborem lze ukázat, že pro $\gamma_i \neq 1$ věta 67 neplatí a to zejména proto, že volba $\alpha_i = 1$ nemá při použití škálování žádné výsadní postavení.

4.7 Aktualizace trojúhelníkového rozkladu

Používáme-li inverzní metody s proměnnou metrikou (133), 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 trojúihelníkový rozklad matice $\bar{B} = B + \sigma z z^T$ z trojúhelníkového rozkladu matice B s použitím $O(n^2)$ operací násobení a sčítání.

Věta 68 *Nechť L, $\bar{L}$ jsou dolní trojúhelníkové matice s jednotkami na hlavní diagonále a D, $\bar{D}$ jsou pozitivně definitní diagonální matice, přičemž*

$$\bar{L}\bar{D}\bar{L}^T = LDL^T + \sigma z z^T. \tag{198}$$

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

$$\begin{aligned} v_i &= z_{ii}, \\ \bar{d}_i &= d_i + \sigma_i v_i^2, \qquad (199) \\ \bar{l}_i &= \frac{d_i}{\bar{d}_i} l_i + \frac{\sigma_i v_i}{\bar{d}_i} z_i \qquad (200) \end{aligned}$$

(z_{ii} je i-tý prvek vektoru z_i), kde $\sigma_1 = \sigma$, $z_1 = z$ a pro $1 \leq i \leq n$ platí

$$\begin{aligned} \sigma_{i+1} &= \frac{d_i}{\bar{d}_i}\sigma_i, \qquad (201) \\ z_{i+1} &= z_i - v_i l_i. \qquad (202) \end{aligned}$$

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

$$\sum_{j=i}^{n} \bar{d}_j \bar{l}_j \bar{l}_j^T = \sum_{j=i}^{n} d_j l_j l_j^T + \sigma_i z_i z_i^T. \tag{203}$$

Zřejmě $\sigma_1 = \sigma$ a $z_1 = z$, neboť rovnost (198) 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, $\bar{l}_j$, $i \le j \le n$, mají prvních $j-1$ prvků nulových, má matice (203) 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

$$\bar{d}_i = d_i + \sigma_i z_{ii}^2, \tag{204}$$

$$\bar{l}_i = \frac{d_i}{\bar{d}_i} l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_i. \tag{205}$$

Vztah (205) můžeme ještě upravit. Položíme-li

$$z_{i+1} = z_i - z_{ii} l_i \tag{206}$$

a použijeme-li (204)–(206), 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

$$\bar{l}_i = l_i + \frac{\sigma_i z_{ii}}{\bar{d}_i} z_{i+1}. \tag{207}$$

Ukážeme nyní, že platí

$$d_i l_i l_i^T - \bar{d}_i \bar{l}_i \bar{l}_i^T + \sigma_i z_i z_i^T = \sigma_{i+1} z_{i+1} z_{i+1}^T, \tag{208}$$

kde číslo σ_{i+1} je určeno vzorcem (201). Použijeme-li vztahy (204)–(207), 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

$$\sigma_{i+1} = \sigma_i - \frac{\sigma_i^2 z_{ii}^2}{\bar{d}_i} = \frac{\bar{d}_i - \sigma_i z_{ii}^2}{\bar{d}_i} \sigma_i = \frac{d_i}{\bar{d}_i} \sigma_i. \tag{209}$$

Platí tedy (208), kde číslo σ_{i+1} je určeno vzorcem (201). Porovnáme-li vztahy (203) a (208), 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 (204) a (205) tedy platí (199) a (200) a vzorce (209) a (206) jsou totožné se vzorci (201) a (202). Zbývá dokázat, že $Lv = z$, kde $v_i = z_{ii}$, $1 \le i \le n$. To je však velmi snadné, neboť podle (202) pro $1 \le i \le n$ 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 13 *Nechť jsou splněny předpoklady věty 68. Pak pro* $1 \le i \le n$ *platí*

$$\begin{aligned} v_i &= z_{ii}, \\ \bar{d}_i &= \frac{\tau_{i+1}}{\tau_i} d_i, \qquad (210) \\ \bar{l}_i &= l_i + \frac{v_i}{\tau_{i+1} d_i} z_{i+1}, \qquad (211) \end{aligned}$$

kde $\tau_1 = 1/\sigma$, $z_1 = z$ *a pro* $1 \le i \le n$ *platí*

$$\begin{aligned} \tau_{i+1} &= \tau_i + \frac{v_i^2}{d_i}, \qquad (212) \\ z_{i+1} &= z_i - v_i l_i \qquad (213) \end{aligned}$$

(přímé rekurence). Nechť vektor v *řešením soustavy rovnic* $Lv = z$. *Pak pro* $n \ge i \ge 1$ *platí*

$$\begin{aligned} \bar{d}_i &= \frac{\tau_{i+1}}{\tau_i} d_i, \qquad (214) \\ \bar{l}_i &= l_i + \frac{v_i}{\tau_{i+1} d_i} z_{i+1}, \qquad (215) \end{aligned}$$

kde $\tau_{n+1} = 1/\sigma + v^T D^{-1} v$, $z_{n+1} = 0$ *a pro* $n \ge i \ge 1$ *platí*

$$\begin{aligned} \tau_i &= \tau_{i+1} - \frac{v_i^2}{d_i}, \qquad (216) \\ z_i &= z_{i+1} + v_i l_i \qquad (217) \end{aligned}$$

(zpětné rekurence).

Důkaz Položme $\tau_i = 1/\sigma_i$, $1 \le i \le n$. Pak podle (199) a (201) 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 (199)–(202) můžeme zapsat ve tvaru (210)–(213) (vzorec (211) plyne ze vzorce (207)). K odvození zpětných rekurencí aplikujeme důsledek 8, na matici (198). 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 (210), 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 (216)–(217) dostaneme obrácením vztahů (212)–(213). □

Poznámka 146 Rekurentní vztahy (199)–(202) 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 (214)–(217). Přímé rekurence (210)–(213), které lze použít pro $\sigma > 0$, mají tu výhodu, že jsou prakticky stejné jako zpětné rekurence (214)–(217), 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 $s_+ = -H_+ g_+$ (tím lze snížit počet operací zhruba o čtvrtinu). Použijeme-li vztahy (102)–(103) a rovnost $d = \alpha s$, 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. Následující věta udává příslušné vzorce.

Věta 69 *Nechť H_+ je matice určená vztahem (116), kde $d = -\alpha Hg$, $\alpha > 0$ je délka kroku a matice H je pozitivně definitní. Pak vektor $s_+ = -H_+ g_+$ můžeme spočítat podle vzorce*

$$\frac{1}{\gamma} s_+ = \frac{\gamma}{\rho}\frac{\delta}{\alpha}\left(\frac{a}{b}d - Hy\right) - \frac{\rho}{\gamma}\frac{d^T g_+}{b} d = \frac{\gamma}{\rho}\frac{\delta}{\alpha}\left(\frac{a}{b}d - Hy\right) + \frac{\rho}{\gamma}\left(\frac{1}{\alpha}\frac{c}{b} - 1\right) d, \tag{218}$$

kde δ je číslo definované vztahem (130).

Důkaz Platí

$$d^T g_+ = d^T(y + g) = d^T y - d^T H^{-1} s = b - \frac{c}{\alpha}, \tag{219}$$

$$y^T H g_+ = y^T H(y + g) = y^T H y - y^T s = a - \frac{b}{\alpha}, \tag{220}$$

takže po dosazení do (116) 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 (130) a (219) dává (218). □

Poznámka 147 Pro metodu DFP, kdy $\eta = 0$, dostaneme

$$s_+^{DFP} = \left[\frac{\gamma}{\alpha}\left(1 + \frac{\rho}{\gamma}\frac{c}{b}\right) - \rho\right] d - \frac{\gamma}{\alpha}\frac{b}{a} Hy. \tag{221}$$

Pro metodu BFGS, kdy $\eta = 1$, dostaneme

$$s_+^{BFGS} = \left[\frac{\gamma}{\alpha}\frac{c}{b}\left(\frac{a}{b} + \frac{\rho}{\gamma}\right) - \rho\right] d - \frac{\gamma}{\alpha}\frac{ac}{b^2} Hy. \tag{222}$$

Poznámka 148 Provádíme-li přesný výběr délky kroku, vymizí v (218) poslední člen, takže všechny aktualizace z Broydenovy třídy aplikované na matici H dávají rovnoběžné směrové vektory s_+. To je v souladu s tvrzením věty 43.

Vzorec (218) lze použít pouze tehdy, když $s = -Hg$. Pokud tento předpoklad neplatí (například u metod s lokálně omezeným krokem, nebo u některých metod s proměnnou metrikou s omezenou pamětí vyšetřovaných v oddílu 8.1), nemusí být splněna podmínka spádovosti $s_+^T g_+ < 0$. Odvodíme ještě jeden vzorec, který má příznivější vlastnosti.

Věta 70 *Označme*

$$p = HVg_+, \qquad V = I - \frac{1}{b} y d^T. \tag{223}$$

Nechť H_+ je matice určená vztahem (116), kde $d = -\alpha Hg$, $\alpha > 0$ je délka kroku a matice H je pozitivně definitní. Pak vektor $s_+ = -H_+ g_+$ můžeme spočítat podle vzorce

$$-\frac{1}{\gamma} s_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \frac{b + \eta\alpha y^T p}{b + \alpha y^T p} V^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \delta \frac{\gamma b}{\rho c} V^T p, \tag{224}$$

kde $\alpha y^T p \geq 0$ a δ je číslo určené vztahem (130).

Důkaz (a) Použijeme-li rovnosti (219) a (223), dostaneme

$$\begin{aligned} p &= HVg_+ = H\left(I - \frac{1}{b} y d^T\right) g_+ = Hg + Hy - \frac{d^T g_+}{b} Hy \\ &= -\frac{1}{\alpha} d + Hy - \left(1 - \frac{c}{\alpha b}\right) Hy = \frac{1}{\alpha}\left(\frac{c}{b} Hy - d\right), \end{aligned} \tag{225}$$

takže $Hy = (\alpha p + d) b/c$ a jelikož $V^T d = 0$, platí

$$V^T Hy = \alpha \frac{b}{c} V^T p. \tag{226}$$

(b) Použijeme-li vyjádření (122), můžeme psát dostaneme

$$\begin{aligned} -\frac{1}{\gamma} s_+ &= \frac{1}{\gamma} H_+ g_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + V^T \left(H + \frac{\eta - 1}{a} H y y^T H\right) V g_+ \\ &= \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + V^T p + \frac{\eta - 1}{a} V^T H y y^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \left(1 + \alpha \frac{\eta - 1}{a} \frac{b}{c} y^T p\right) V^T p. \end{aligned}$$

Ale $a = y^T Hy = y^T(\alpha p + d) b/c = (b + \alpha y^T p) b/c$, takže

$$-\frac{1}{\gamma} s_+ = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \left(1 + \frac{\eta - 1}{b + \alpha y^T p} \alpha y^T p\right) V^T p = \frac{\rho}{\gamma} \frac{d^T g_+}{b} d + \frac{b + \alpha y^T p}{b + \alpha y^T p} V^T p.$$

(c) Použijeme-li vztah (225) 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 (150) a použijeme-li (130), 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 149 Jelikož $V^T p = V^T HVg_+$, můžeme vzorec (224) 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 150 Neplatí-li $s = -Hg$, můžeme vzorec (224) upravit tak, že místo čísla $\alpha y^T p$ použijeme číslo $\tau = \max(\alpha y^T p, 0)$. Pak podle (224) 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 $Vg_+ = g_+$, takže $g_+^T s_+ = -g_+^T H g_+(b + \eta\alpha\tau)/(b + \alpha\tau) < 0$, neboť matice H je pozitivně definitní a $\tau \geq 0$.

Nyní popíšeme modifikaci metod s proměnnou metrikou v součinovém tvaru, která používá ortogonální transformace umožňující značně zjednodušit použité aktualizace. 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 71 *Nechť $H = SS^T$, $[d, Hy] = S[\tilde{d}, \tilde{y}]$, $a = \tilde{y}^T\tilde{y} > 0$, $b = \tilde{y}^T\tilde{d} > 0$, $c = \tilde{d}^T\tilde{d} > 0$. 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*

$$\begin{aligned}
\frac{1}{\sqrt{\gamma}} S_+ e_1 &= \sqrt{\frac{\rho}{\gamma b}}\, d \\
\frac{1}{\sqrt{\gamma}} S_+ e_2 &= \sqrt{\frac{(y^T\bar{s}_1)^2 + \eta(y^T\bar{s}_2)^2}{(y^T\bar{s}_1)^2 + (y^T\bar{s}_2)^2}} \left(\bar{s}_2 - \frac{y^T\bar{s}_2}{b} d\right) \\
\frac{1}{\sqrt{\gamma}} S_+ e_j &= \bar{s}_j, \qquad 3 \leq j \leq n,
\end{aligned}$$

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 (116).

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

$$\begin{aligned}
\left(I - \frac{dy^T}{y^T d}\right)\bar{s}_1 &= 0, \\
\left(I - \frac{dy^T}{y^T d}\right)\bar{s}_2 &= \bar{s}_2 - \frac{y^T\bar{s}_2}{y^T d} d, \\
\left(I - \frac{dy^T}{y^T d}\right)\bar{s}_j &= \bar{s}_j, \quad 3 \leq j \leq n
\end{aligned}$$

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 dá vá

$$\begin{aligned}
\frac{Hy}{y^T Hy} - \frac{d}{y^T d} &= \frac{y^T \bar{s}_1 \, \bar{s}_1 + y^T \bar{s}_2 \, \bar{s}_2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} - \frac{\bar{s}_1}{y^T \bar{s}_1} \\
&= \frac{y^T \bar{s}_2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} \bar{s}_2 + \frac{y^T \bar{s}_1}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} \bar{s}_2 - \frac{1}{y^T \bar{s}_1} \bar{s}_1 \\
&= \frac{y^T \bar{s}_2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T \bar{s}_1} \bar{s}_1 \right).
\end{aligned}$$

Nyní použijeme vztah (123), podle kterého platí

$$\begin{aligned}
\frac{1}{\gamma} H_+ &= \left(I - \frac{dy^T}{y^T d} \right) \bar{S}\bar{S}^T \left(I - \frac{dy^T}{y^T d} \right)^T + \frac{\rho}{\gamma} \frac{dd^T}{y^T d} \\
&\quad + y^T Hy (\eta - 1) \left(\frac{d}{y^T d} - \frac{Hy}{y^T Hy} \right) \left(\frac{d}{y^T d} - \frac{Hy}{y^T Hy} \right)^T \\
&= \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right) \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right)^T + \sum_{j=3}^{n} \bar{s}_j \bar{s}_j^T + \frac{\rho}{\gamma} \frac{dd^T}{y^T d} \\
&\quad + (\eta - 1) \frac{(y^T \bar{s}_2)^2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right) \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right)^T \\
&= \sum_{j=3}^{n} \bar{s}_j \bar{s}_j^T + \frac{(y^T \bar{s}_1)^2 + \eta (y^T \bar{s}_2)^2}{(y^T \bar{s}_1)^2 + (y^T \bar{s}_2)^2} \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right) \left(\bar{s}_2 - \frac{y^T \bar{s}_2}{y^T d} d \right)^T + \frac{\rho}{\gamma} \frac{dd^T}{y^T d} \\
&= \frac{1}{\gamma} \sum_{j=1}^{n} S_+ e_j (S_+ e_j)^T = \frac{1}{\gamma} S_+ S_+^T.
\end{aligned}$$

□

Poznámka 151 Z věty 71 plyne, že stačí aktualizovat pouze dva sloupce matice $\bar{S}$. 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 zobecněou 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 141) a také ke korekci kvadratického modelu.

Poznámka 152 Nechť $B_+ d = \tilde{y}$, kde $\tilde{y} = y + \tau d$. Pak $d^T B_+ d = d^T \tilde{y} = d^T y + \tau \|d\|^2$, takže $d^T B_+ d = (1/\rho) d^T y$ právě tehdy, když

$$\tau = \left(\frac{1}{\rho} - 1 \right) \frac{d^T y}{\|d\|^2}.$$

Abychom dostali korekci jako v poznámce 134, stačí položit $\tau \|d\|^2 = 2(F - F_+) + d^T g_+ + d^T g$. Abychom dostali korekci jako v poznámce 135, stačí položit $\tau \|d\|^2 = 6(F - F_+) + 3(d^T g_+ + d^T g)$.

Jiná možnost spočívá v použití hodnot a gradientů funkce F ve více bodech, například v bodech x_-, x, x_+, kde x_- je vektor z předchozího iteračního kroku. Zobecněná kvazinewtonovská podmínka se odvodí

pomocí vhodné interpolace. Všechny tyto modifikace (včetně té uvedené v poznámce 152) však příliš nezlepšují (při použití škálování dokonce zhoršují) účinnost metod s proměnou metrikou. Proto se jimi dále zabývat nebudeme.

Zatím jsme se zabývali modifikacemi metod z Broydenovy třídy. Nyní popíšeme Davidonovu třídu metod s proměnnou metrikou. Tato třída používá vztahy (102)–(104), kde $U_i = [u_i, d_i - H_i y_i]$ a u_i je vektor, který se konstruuje rekurentně tak, aby platilo

$$u_{i+1} \in \mathcal{L}(U_i), \qquad u_{i+1}^T y_i = 0 \tag{227}$$

(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 například vektor

$$u_{i+1} = y_i^T (d_i - H_i y_i) u_i - y_i^T u_i (d_i - H_i y_i). \tag{228}$$

Věta 72 *(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 (40). Opět se indukcí pro $1 \leq j < i \leq n+1$ dokazují vztahy (105)–(108) a navíc vztah

$$u_i^T y_j = 0, \tag{229}$$

který se používá v části (a) k důkazu toho, že $U_i^T y_j = 0$, $1 \leq j \leq i$. Indukční krok pro (229) je velmi jednoduchý. Jelikož $u_{i+1} \in \mathcal{L}(U_i)$ a $U_i^T y_j = 0$, $1 \leq j < i$, platí $u_{i+1}^T y_j = 0$, $1 \leq j < i$. Protože podle (227) platí $u_{i+1}^T y_i = 0$, dostaneme $u_{i+1}^T y_j = 0$, $1 \leq j \leq i$. □

Věta 72 neposkytuje nic nového, co by nesplňovala i jednodušší Broydenova třída metod s proměnnou metrikou. Následující věta však ukazuje, že za jistých předpokladů má Davidonova třída vlastnost kvadratického ukončení i bez přesného výběru délky kroku.

Věta 73 *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 $\rho_i = 1$, $\gamma_i = 1$ a $y_i^T u_i \neq 0$, $1 \leq i \leq n$, 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 nějký index $1 < i \leq n$ platí $\dim \mathcal{Z}_i \geq i - 1$, (platí to pro i = 2, neboť s použitím (103) (104) a (227) dostaneme $GH_2 y_1 = G d_1 = y_1$ 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 $U_i^T z = 0$. Podle (103) a (227) 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 (104) a (227) dostaneme $GH_{i+1} y_i = G d_i = y_i$ 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 153 Ve větě 73 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 $H_{i+1} = G^{-1}$.

Poznámka 154 Dá se snadno ukázat, že aktualizace $H_+ = H + UMU^T$, kde $U = [u, d - Hy]$, splňuje kvazinewtonovskou podmínku $H_+y = d$, pokud

$$H_+ = H + \frac{(d - Hy)(d - Hy)^T}{y^T(d - Hy)} - \frac{\varphi\, u_+ u_+^T}{y^T(d - Hy)}, \tag{230}$$

kde $\varphi = -\det M$ je volný parametr a u_+ je vektor určený vztahem (228). Zvolíme-li $\varphi = 0$, dostaneme metodu hodnosti 1 (vzorec (119)), která patří do Davidonovy třídy. Davidonovu třídu lze tedy chápat jako zobecnění metody hodnosti 1. Podstatné je, že Davidonova třída obsahuje aktualizace, pro něž je matice H_+ vždy pozitivně definitní.

Vztah (230) je velmi jednoduchý, neposkytuje však jednoduché výrazy pro vhodné hodnoty parametru φ dávající pozitivně definitní matici H_+. Proto se používá aktualizace $H_+ = H + \bar{U}\bar{M}\bar{U}$, kde $\bar{U} = P[d, Hy]$ a $P = U(U^T H^{-1} U)^{-1} U^T H^{-1}$ (P je matice projekce do $\mathcal{L}(U)$).

Věta 74 *Nechť $H_+ = H + \bar{U}\bar{M}\bar{U}^T$, kde H je symetrická pozitivně definitní matice a $\bar{U} = P[d, Hy]$, kde $P = U(U^T H^{-1} U)^{-1} U^T H^{-1}$ a $U = [u, d - Hy]$. Pak $H_+ y = d$ platí právě tehdy, jestliže*

$$\bar{M} = \begin{bmatrix} \frac{1}{\bar{b}}\left(\bar{\eta}\frac{\bar{a}}{\bar{b}} + 1\right), & -\frac{\bar{\eta}}{\bar{b}} \\ -\frac{\bar{\eta}}{\bar{b}}, & \frac{\bar{\eta} - 1}{\bar{a}} \end{bmatrix},$$

kde $\bar{\eta}$ je volný parametr a kde

$$\bar{a} = y^T P H y, \qquad \bar{b} = y^T P d, \qquad \bar{c} = d^T H^{-1} P d.$$

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

$$\begin{aligned} P^T H^{-1} P &= H^{-1} U (U^T H^{-1} U)^{-1} U^T H^{-1} U (U^T H^{-1} U)^{-1} U^T H^{-1} \\ &= H^{-1} U (U^T H^{-1} U)^{-1} U^T H^{-1} = H^{-1} P = P^T H^{-1} \end{aligned} \tag{231}$$

takže $\bar{U}^T P^T y = \bar{U}^T P^T H^{-1} H y = \bar{U}^T P^T H^{-1} P H y = \bar{U}^T H^{-1} P H y$, což po dosazení do kvazinewtonovské podmínky dává

$$[Pd, PHy] \begin{bmatrix} m_1, & m_2, \\ m_2, & m_3 \end{bmatrix} \begin{bmatrix} \bar{b} \\ \bar{a} \end{bmatrix} = Pd - PHy,$$

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

$$\begin{aligned} m_1\bar{b} + m_2\bar{a} &= 1, \\ m_2\bar{b} + m_3\bar{a} &= -1. \end{aligned}$$

Jeden parametr je nadbytečný. Zvolíme $m_2 = -\bar{\eta}/b$ a zbylé prvky m_1, m_3 určíme řešením uvedených rovnic. Tím dostaneme matici $\bar{M}$ uvedenou ve větě 74. □

Poznámka 155 Vztah $H_+ = H + \bar{U}\bar{M}\bar{U}^T$ můžeme roznásobit. Platí

$$H_+ = H + \frac{1}{\bar{b}}Pd(Pd)^T - \frac{1}{\bar{a}}PHy(PHy)^T + \frac{\bar{\eta}}{\bar{a}}\left(\frac{\bar{a}}{\bar{b}}Pd - PHy\right)\left(\frac{\bar{a}}{\bar{b}}Pd - PHy\right)^T. \tag{232}$$

Pro tuto aktualizaci platí stejné úvahy jako pro (116). Matice H_+ je pozitivně definitní, pokud $\bar{\eta} > \bar{\eta}^*$, kde $\bar{\eta}^* = -\bar{b}^2/(\bar{a}\bar{c} - \bar{b}^2)$. Pro $\bar{\eta} = 1$ dostaneme analogii metody BFGS.

Poznámka 156 Aktualizaci (232) můžeme vyjádřit v inverzním tvaru. Použije se k tomu stejný postup jako v důkazu věty 46 a rovnost $BP = P^TB$, která plyne z (231). Platí

$$B_+ = B + \frac{1}{\bar{b}}P^Ty(P^Ty)^T - \frac{1}{\bar{c}}P^TBd(P^TBd)^T + \frac{\bar{\beta}}{\bar{c}}\left(\frac{\bar{c}}{\bar{b}}P^Ty - P^TBd\right)\left(\frac{\bar{c}}{\bar{b}}P^Ty - P^TBd\right)^T, \tag{233}$$

kde

$$\bar{\beta}\bar{\eta}(\bar{a}\bar{c} - \bar{b}^2) + (\bar{\beta} + \bar{\eta})\bar{b}^2 = \bar{b}^2.$$

Poznámka 157 Inverzí vztahu $H_+ = H + UMU^T$ podle důsledku 8 dostaneme (tak jako v důkazu věty 46) $B_+ = B + BUKBU^T$. Položíme-li $v = -Bu$ a $V = [v, y - Bd]$, můžeme psát $B_+ = B + VKV^T$. Dá se snadno ukázat, že tato aktualizace splňuje kvazinewtonovskou podmínku $B_+d = y$, pokud

$$B_+ = B + \frac{(y - Bd)(y - Bd)^T}{d^T(y - Bd)} - \frac{\psi\, v_+v_+^T}{d^T(y - Bd)}, \tag{234}$$

kde $\psi = -\det K$ je volný parametr a

$$v_+ = d^T(y - Bd)v - d^Tv(y - Bd). \tag{235}$$

Vztah (234) je duální ke vztahu (230) a platí $\psi = \varphi/\bar{\delta}$, kde $\bar{\delta} = \det H_+/\det H = \det B/\det B_+$.

Poznámka 158 Nechť $v = -Bu$ a v_+ je vektor určený vztahem (235). Pak platí

$$-B_+u_+ = \frac{v_+}{\bar{\delta}}.$$

Jelikož nezáleží na normě vektoru u_+, můžeme v dalším iteračním kroku použít vektor v_+ místo vektoru B_+u_+. V prvním iteračním kroku lze volit vektory u a v libovolně.

Poznámka 159 Ke konstrukci matice $P = U(U^TH^{-1}U)^{-1}U^TH^{-1}$ je zapotřebí matice $H^{-1} = B$, která je k disposici, používáme-li aktualizaci (233). V případě aktualizace (232) můžeme použít vztah

$$P = U(U^TH^{-1}U)^{-1}U^TH^{-1} = HBU(U^TBHBU)^{-1}U^TB = HV(V^THV)^{-1}V,$$

ve kterém je místo vektoru u použit vektor v, určovaný rekurentně podle vzorce (235). Ve vztahu (235) využíváme toho, že pro metody spádových směrů platí $Bd = -\alpha g$ (poznámka 86).

Z modifikací, které jsme uvedli, má praktický význam pouze eliminace maticového násobení užitím věty 69. Ostatní modifikace lze překonat vhodným škálováním (parametr γ) a korekcí (parametr ρ). Metody z Davidonovy třídy jsou efektivní co se týče počtu vyčíslení hodnot a gradientů minimalizované funkce. Vyžadují však delší čas výpočtu, který se spotřebovává na určení projekcí nebo parametrů φ a ψ v (230) a (234).

Poznámka 160 Závěrem uvedeme několik poznámek k implementaci metod s proměnnou metrikou.

- 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)

$$\alpha = \min\left(1, \frac{4(\underline{F} - F_i)}{s_i^T g_i}\right).$$

- Korekce (parametr ρ): Vyplácí se, zejména ve spojení se škálováním, používat parametr ρ určený zpětným použitím věty o střední hodnotě (poznámka 134) nebo použitím homogenního modelu (poznámka 137) a upravený tak, aby platilo $\underline{\rho} \leq \rho \leq \underline{\rho}$.
- Škálování (parametr γ): Vhodné škálování značně zvyšuje účinnost omezených metod s proměnnou metrikou (kdy $0 \leq \eta \leq 1$), pokud volíme parametr γ tak, aby platilo $b/c \leq \rho/\gamma \leq a/b$. 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 škálování popsané v poznámce 139.
- Výběr konkretni 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, korekce a škálování rozdíly mezi nimi stírají (s celkovým zlepšením účinnosti), takže lze doporučit korigovanou a škálovanou metodu BFGS.

Algoritmus metody s proměnnou metrikou lze popsat zhruba takto:

Algoritmus 4 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 $\|g_i\| \leq \underline{\varepsilon}$ 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 zpětným použitím věty o střední hodnotě (poznámka 134) nebo použitím homogenního modelu (poznámka 137). Jestliže $\rho_i < \underline{\rho}$ nebo $\rho_i > \overline{\rho}$ položíme $\rho_i := 1$. Použijeme řízené škálování (poznámka 139) 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 (116) (pro metodu BFGS volíme $\eta_i = 1$).

Krok 4 Zvětšíme i o 1 a přejdeme na krok 2.

Následující tabulka ukazuje srovnání několika metod s proměnnou metrikou a jejich porovnání s metodou sdružených gradientů pomocí souboru 92 testovacích problémů s 50 a 200 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a selhání F, jakož i celkový čas výpočtu). V tabulce je kromě metody DFP (117), metody BFGS (118) a Hoshinovy metody (120) uvedena metoda, která používá hodnoty

$$\begin{aligned} \eta^N &= \frac{\max(0, \sqrt{c/a} - b^2/(ac))}{1 - b^2/(ac)}, \qquad b^2/(ac) < 1, \\ \eta^N &= 1, \qquad\qquad\qquad\qquad\qquad\quad b^2/(ac) \geq 1 \end{aligned}$$

(poznámka 133). Tuto metodu uvádíme, abychom demonstrovali, že efektivita metody BFGS může být překonána vhodnou volbou parametru η. Označení typu škálování v prvním sloupci tabulky má stejný význam jako v poznámce 139. Pro NS a PS se používá hodnota $\rho = 1$. Pro CS se používá parametr ρ určený použitím homogenního modelu (poznámka 137). První sada sloupců odpovídá dimenzi $n = 50$ a

druhá dimenzi $n = 200$. Pro srovnání jsou též uvedeny výsledky pro metodu sdružených gradientů CG (je použita verze Hestenese a Stiefela a silná Wolfeho podmínka pro výběr délky kroku).

Metoda	NIT	NFV	F	Čas	NIT	NFV	F	Čas
DFP + NS	79799	84566	36	4.79	125737	138199	34	53.06
DFP + PS	89241	90993	39	4.76	143663	146872	42	54.94
DFP + CS	12857	15266	1	0.88	36030	43824	4	13.50
BFGS + NS	13192	21453	1	1.47	32492	57518	1	25.20
BFGS + PS	15193	16840	1	1.13	34724	38444	3	11.95
BFGS + CS	8024	9527	-	0.69	19439	22783	-	7.30
H + NS	15928	21565	1	1.42	36475	51893	1	17.03
H + PS	18561	19600	1	1.11	39889	42024	2	13.08
H + CS	8225	9577	-	0.64	20716	24085	-	7.70
N + NS	11493	17537	1	1.22	29760	45904	1	16.42
N + PS	11895	13891	1	0.79	27048	29608	2	9.30
N + CS	7814	8796	-	0.63	18366	20426	-	6.99
CG	163098	311093	5	10.25	227547	417467	9	18.61

Poznámka 161 Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Metoda DFP je velmi neefektivní (používáme-li standardní výběr délky kroku založený na použití slabé Wolfeho podmínky).
- Řízené škálování velmi zvyšuje efektivitu metod s proměnnou metrikou (podobnou vlastnost má i intervalové škálování).
- Efektivita metody BFGS může být překonána vhodnou volbou parametru η (například volbou $\eta = \eta^N$).
- 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.

5 Metody s lokálně omezeným krokem

5.1 Základní vlastnosti metod s lokálně omezeným krokem

Poznámka 162 Při výkladu metod s lokálně omezeným krokem budeme používat označení

$$Q_i(s) = g_i^T s + \frac{1}{2} s^T B_i s$$

pro kvadratickou funkci, která lokálně aproximuje rozdíl $F(x_i + s) - F(x_i)$ a označení

$$\omega_i(s) = (B_i s + g_i)/\|g_i\|$$

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

$$\rho_i(s) = \frac{F(x_i + s) - F(x_i)}{Q_i(s)}$$

pro podíl skutečného a předpověděného poklesu funkce F.

Definice 25 *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*

$$\|s_i\| \leq \overline{\delta} \Delta_i, \tag{T1a}$$

$$\|s_i\| < \underline{\delta} \Delta_i \Rightarrow \|\omega_i(s_i)\| \leq \overline{\omega}_i \leq \overline{\omega}, \tag{T1b}$$

$$-Q_i(s_i) \geq \underline{\sigma} \|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|), \tag{T1c}$$

kde $0 < \underline{\delta} \leq 1 \leq \overline{\delta}$, $0 < \underline{\sigma} < 1$ a $0 \leq \overline{\omega} < 1$, délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq \underline{\rho} \;\Rightarrow\; \alpha_i = 0, \tag{T2a}$$

$$\rho_i(s_i) > \underline{\rho} \;\Rightarrow\; \alpha_i = 1, \tag{T2b}$$

kde $\underline{\rho} \geq 0$, a čísla $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se volí tak, že

$$\rho_i(s_i) < \overline{\rho} \;\Rightarrow\; \underline{\beta} \|s_i\| \leq \Delta_{i+1} \leq \overline{\beta} \|s_i\|, \tag{T3a}$$

$$\rho_i(s_i) \geq \overline{\rho} \;\Rightarrow\; \Delta_i \leq \Delta_{i+1} \leq \min(\underline{\gamma} \Delta_i, \overline{\Delta}), \tag{T3b}$$

kde $0 < \underline{\beta} \leq \overline{\beta} < 1 < \underline{\gamma} < \infty$ a $0 \leq \underline{\rho} < \overline{\rho} < 1$, přičemž $\overline{\beta}\,\overline{\delta} < 1$.

Poznámka 163 Jestliže $\overline{\omega} = 0$ 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 $\underline{\sigma}$ není vnějším parametrem metody. Jeho existence musí být zaručena, ale jeho velikost závisí na zvolené metodě (obvykle $\underline{\sigma} = 1/2$). V (T2) obvykle volíme $\underline{\rho} = 0$. Pokud $\underline{\rho} > 0$, dostaneme silnější tvrzení o globální konvergenci (věta 76). Číslo $\overline{\Delta} > 0$ slouží k omezení délky kroku, abychom se nedostali mimo definiční obor $\mathcal{D}_F$ funkce $F : \mathcal{D}_F \to R$. Při vyšetřování metod s lokálně omezeným krokem budeme předpokládat, že $\mathcal{D}_F(\overline{F}) + B(0, \overline{\Delta}) \subset \mathcal{D}$. Pak $x_i + s_i \in \mathcal{D}$ $\forall i \in N$.

Poznámka 164 V podmínce (T1c) se někdy používá $\|s_i\|$ místo Δ_i, což je možné, neboť podle (T1a) platí $\Delta_i \geq \|s_i\|/\overline{\delta}$. Navíc Δ_i se v podmínce (T1c) uplatňuje většinou tehdy, když $\|s_i\| \geq \underline{\delta} \Delta_i$ (viz důkaz věty 80).

Poznámka 165 Normy v (T1) a (T3) mohou být i jiné než euklidovské. V tomto případě se využívá ekvivalence norem. Pro libovolnou vektorovou normu $\|s\|_*$ platí $\underline{\nu}\|s\| \leq \|s\|_* \leq \overline{\nu}\|s\|$ a podíl $\overline{\nu}/\underline{\nu}$ pak vystupuje v odpovídajících vzorcích.

Poznámka 166 Při vyšetřování metod s lokálně omezeným krokem budeme používat označení

$$N_1 = \{i \in N : \|s_i\| < \underline{\delta}\Delta_i\},$$

$$N_2 = \{i \in N : \rho_i(s_i) > \underline{\rho}\},$$

$$N_3 = \{i \in N : \rho_i(s_i) \geq \overline{\rho}\}.$$

Jelikož $0 \leq \underline{\rho} < \overline{\rho}$, platí $N_3 \subset N_2$.

Lemma 28 *Aplikujeme-li metodu s lokálně omezeným krokem (T1)–(T3) na funkci $F : \mathcal{D} \to R$, která splňuje podmínku (F3), existuje konstanta $0 < \underline{c} < 1$ taková, že*

$$\|s_i\| \geq \underline{c} m_i / M_i, \tag{236}$$

kde

$$m_i = \min_{1 \leq j \leq i} \|g_j\|,$$

$$M_i = \max_{1 \leq j \leq i} \|B_j\|.$$

Důkaz (a) Nechť $i \in N_1$. Pak podle (T1b) platí

$$|\|B_i s_i\| - \|g_i\|| \leq \|B_i s_i + g_i\| = \|\omega_i(s_i)\|\|g_i\| \leq \overline{\omega}\|g_i\|,$$

takže buď $\|B_i s_i\| \geq \|g_i\|$ nebo $\|B_i s_i\| < \|g_i\|$ a $\|B_i s_i\| \geq (1-\overline{\omega})\|g_i\|$. Spojením těchto nerovností dostaneme $\|B_i\|\|s_i\| \geq \|B_i s_i\| \geq (1-\overline{\omega})\|g_i\|$, což dává $\|s_i\| \geq (1-\overline{\omega})m_i/M_i$.

(b) Nechť $i \notin N_1$ a $i \notin N_3$. Pak podle definice množiny N_3 a funkce $Q_i(s)$ platí

$$F(x_i + s_i) - F(x_i) \geq \overline{\rho} Q_i(s_i) = \overline{\rho}\left(g_i^T s_i + \frac{1}{2} s_i^T B_i s_i\right) \geq \overline{\rho}\left(g_i^T s_i - \|B_i\|\|s_i\|^2\right).$$

Z druhé strany použitím (1) dostaneme

$$F(x_i + s_i) - F(x_i) \leq g_i^T s_i + \overline{G}\|s_i\|^2,$$

což dohromady dává

$$(\overline{G} + \overline{\rho}\|B_i\|)\|s_i\|^2 \geq (\overline{\rho} - 1) g_i^T s_i.$$

Podle (T1c) platí

$$-\underline{\sigma}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|) \geq Q_i(s_i) \geq g_i^T s_i - \|B_i\|\|s_i\|^2,$$

což spolu s předchozí nerovností dává

$$\begin{aligned}(\overline{G} + \overline{\rho}\|B_i\|)\|s_i\|^2 &\geq (\overline{\rho} - 1) g_i^T s_i \geq (\overline{\rho} - 1)\|B_i\|\|s_i\|^2 - \\ &- \underline{\sigma}(\overline{\rho} - 1)\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|),\end{aligned}$$

neboli

$$(\overline{G} + \|B_i\|)\|s_i\|^2 \geq \underline{\sigma}(1-\overline{\rho})\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|),$$

takže buď $\|s_i\| \geq \underline{\delta}\Delta_i \geq \underline{\delta}\|g_i\|/\|B_i\| \geq \underline{\delta} m_i/M_i$, nebo

$$(\overline{G} + \|B_1\|)\frac{M_i}{\|B_1\|}\|s_i\|^2 \geq (\overline{G} + \|B_i\|)\|s_i\|^2 \geq \underline{\sigma}(1-\overline{\rho})\|g_i\|\Delta_i \geq \frac{\underline{\sigma}(1-\overline{\rho})}{\overline{\delta}}\|g_i\|\|s_i\|,$$

což dává $\|s_i\| \geq \left(\underline{\sigma}(1-\overline{\rho})\|B_1\|/\left(\overline{\delta}(\overline{G} + \|B_1\|)\right)\right) m_i/M_i$.

(c) Nechť $i = 1$. Pokud $\|g_1\| = 0$, platí zřejmě $\|s_1\| \geq \|g_1\|/\|B_1\| \geq m_1/M_1$. Pokud $\|g_1\| \neq 0$ můžeme psát

$$\|s_1\| = \frac{\|s_1\|\|B_1\|}{\|g_1\|}\frac{\|g_1\|}{\|B_1\|},$$

takže $\|s_1\| \geq (\|s_1\|\|B_1\|/\|g_1\|)m_1/M_1$

(d) Nechť $i \notin N_1$, $i \in N_3$ a $i \neq 1$. Nechť $k < i$ je největší index pro který neplatí současně $k \notin N_1$, $k \in N_3$ a $k \neq 1$. Pak podle (T3) a (T1a) platí

$$\|s_i\| \geq \underline{\delta}\Delta_i \geq \underline{\delta}\Delta_{i-1} \geq \ldots \geq \underline{\delta}\Delta_{k+1} \geq \underline{\beta}\,\underline{\delta}\|s_k\|,$$

takže podle (a)–(c) platí

$$\|s_i\| \geq \underline{\beta}\,\underline{\delta}\|s_k\| \geq \underline{c}\, m_k/M_k \geq \underline{c}\, m_i/M_i,$$

kde

$$\underline{c} = \underline{\beta}\,\underline{\delta}\min\left((1-\overline{\omega}),\, \underline{\delta},\, \frac{\underline{\sigma}(1-\overline{\rho})\|B_1\|}{\overline{\delta}(\overline{G} + \|B_1\|)},\, \frac{\|s_1\|\|B_1\|}{\|g_1\|}\right).$$

□

Lemma 29 *(Powell) Nechť Δ_i, $i \in N$, a M_i, $i \in N$, jsou dvě posloupnosti kladných čísel a $N_3 \subset N$. Nechť*

$$\Delta_i \geq \frac{\mu}{M_i} > 0, \quad i \in N, \tag{237}$$

kde $\mu > 0$,

$$\begin{aligned} \Delta_{i+1} &\leq \gamma\Delta_i, \quad i \in N_3, && (238)\\ \Delta_{i+1} &\leq \beta\Delta_i, \quad i \notin N_3, && (239)\\ M_{i+1} &\geq M_i, \quad i \in N, && (240) \end{aligned}$$

kde $0 < \beta < 1 < \gamma$, a

$$\sum_{i \in N_3} \frac{1}{M_i} < \infty. \tag{241}$$

Pak

$$\sum_{i \in N} \frac{1}{M_i} < \infty. \tag{242}$$

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 $[1, i] = \{1, \ldots, i\}$, které jsou prvky množiny N_3 (čili $p(i)$ je mohutnost množiny $[1, i] \cap N_3$). Nechť $i \in N_4$, kde

$$N_4 = \{i \in N : rp(i) < i\}.$$

Pak podle (238) a (239) platí

$$\Delta_i \le \gamma^{p(i-1)}\beta^{i-1-p(i-1)}\Delta_1 \le \gamma^{(i-1)/r}\beta^{(r-1)(i-1)/r}\Delta_1 \le \left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1.$$

Protože podle předpokladu je $\gamma\beta^{r-1} < 1$, můžeme psát

$$\sum_{i\in N_4}\Delta_i \le \sum_{i\in N_4}\left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1 \le \sum_{i=1}^{\infty}\left(\gamma\beta^{(r-1)}\right)^{(i-1)/r}\Delta_1 = \frac{\Delta_1}{1-\left(\gamma\beta^{(r-1)}\right)^{1/r}} < \infty.$$

Použijeme-li nyní (237), dostaneme

$$\sum_{i\in N_4}\frac{1}{M_i} \le \frac{1}{\mu}\sum_{i\in N_4}\Delta_i < \infty.$$

(b) Nyní stačí dokázat, že

$$\sum_{i\in N_5}\frac{1}{M_i} < \infty,$$

kde $N_5 = N \setminus N_4$, takže $N_5 = \{i \in N : rp(i) \ge i\}$. Označme

$$N_3 = \{i_1, i_2, i_3 \ldots\}, \qquad N_5 = \{k_1, k_2, k_3 \ldots\}$$

(předpokládáme uspořádání prvků podle velikosti) a sestrojme množinu

$$N_6 = \{l_1, l_2, l_3 \ldots\} = \{\underbrace{i_1, \ldots, i_1}_{r-\text{krát}}, \underbrace{i_2, \ldots, i_2}_{r-\text{krát}}, \underbrace{i_3, \ldots, i_3}_{r-\text{krát}}, \ldots\}.$$

Z konstrukce množiny N_5 plyne, že

$$rp(k_j) \ge k_j \ge j \quad \forall j \in N,$$

takže podle definice množiny N_6 dostaneme

$$l_j \le l_{rp(k_j)} \le i_{p(k_j)} \le k_j \quad \forall j \in N,$$

neboť $i_{p(k_j)}$ je poslední prvek množiny $[1, k_j] \cap N_3$. Podle (240) tedy platí $M_{l_j} \le M_{k_j}$ $\forall j \in N$, takže podle (241) dostaneme

$$\sum_{i\in N_5}\frac{1}{M_i} = \sum_{j=1}^{\infty}\frac{1}{M_{k_j}} \le \sum_{j=1}^{\infty}\frac{1}{M_{l_j}} = \sum_{i\in N_6}\frac{1}{M_i} = r\sum_{i\in N_3}\frac{1}{M_i} < \infty,$$

což spolu s (a) dává (242). □

Věta 75 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3) taková, že*

$$\sum_{i=1}^{\infty}\frac{1}{M_i} = \infty, \tag{243}$$

kde M_i, $i \in N$, jsou čísla definovaná v lemmatu 28. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak platí

$$\liminf_{i\to\infty}\|g_i\| = 0. \tag{244}$$

Důkaz (a) Předpokládejme, že existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$. Pak podle (T1a) a lemmatu 28 platí

$$\Delta_i \geq \frac{1}{\overline{\delta}}\|s_i\| \geq \frac{\underline{c}\,\underline{\varepsilon}}{\overline{\delta} M_i} \triangleq \frac{\mu}{M_i} \tag{245}$$

$\forall i \in N$. Protože $N_3 \subset N_2$, můžeme s použitím (245) psát

$$F_i - F_{i+1} = F(x_i) - F(x_i + s_i) \geq -\overline{\rho} Q_i(s_i) \geq \overline{\rho}\,\underline{\sigma}\,\underline{\varepsilon} \min\left(\Delta_i, \frac{\underline{\varepsilon}}{M_i}\right) \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{\varepsilon}^2 \underline{c}}{\overline{\delta}} \frac{1}{M_i}$$

$\forall i \in N_3$, takže

$$F_1 - \underline{F} \geq \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i\in N_3}(F_i - F_{i+1}) \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{\varepsilon}^2 \underline{c}}{\overline{\delta}} \sum_{i\in N_3} \frac{1}{M_i}.$$

Platí tedy

$$\sum_{i\in N_3} \frac{1}{M_i} < \infty.$$

(b) Položíme-li $\beta = \overline{\beta}\,\overline{\delta} < 1$ a $\gamma = \underline{\gamma} > 1$, jsou splněny předpoklady lemmatu 29, takže platí (242) což je ve sporu s předpokladem věty. □

Poznámka 167 Předpoklady věty 75 jsou splněny například tehdy, jsou-li matice B_i, $i \in N$, stejnoměrně omezené, kdy platí

$$\|B_i\| \leq \overline{B} \quad \forall i \in N.$$

Důkaz tohoto dílčího tvrzení je velmi jednoduchý. Stačí část (a) důkazu věty 75 pozměnit tak, že

$$F_1 - \underline{F} \geq \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i\in N_3}(F_i - F_{i+1}) \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{\varepsilon}^2 \underline{c}}{\overline{\delta}} \sum_{i\in N_3} \frac{1}{\overline{B}}.$$

Je-li množina N_3 nekonečná, dojdeme ihned ke sporu. Je-li množina N_3 konečná, musí podle (T3a) platit $\Delta_i \to 0$, což je ve sporu s (245), neboť $M_i \leq \overline{B}$.

Poznámka 168 Předpoklady věty 75 jsou splněny také tehdy, jsou-li matice B_i, $i \in N$, dostatečně omezené, kdy platí

$$\|B_i\| \leq C_i \quad \forall i \in N$$

a čísla C_i vyhovují rekurentním nerovnostem

$$C_{i+1} \leq C_i + \overline{C}\|s_i\| \leq C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}\,i,$$

kde $C_1 > 1$ a $\overline{C} \geq 0$ jsou vhodné konstanty. V tomto případě platí

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \geq \frac{1}{C_1} + \sum_{i=1}^{\infty} \frac{1}{C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}\,i} \geq \frac{1}{C_1} + \frac{1}{C_1 + \overline{C}\,\overline{\delta}\,\overline{\Delta}} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

neboť harmonická řada je divergentní.

V poznámce 25 jsme ukázali, že pro metody stejnoměrně spádových směrů platí $\|g_i\| \to 0$. Nyní dokážeme, že totéž platí pro metody s lokálně omezeným krokem, jsou-li matice B_i, $i \in N$, stejnoměrně omezené a platí-li $\underline{\rho} > 0$.

Věta 76 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3) takovou, že $\|B_i\| \leq \overline{B}$ $\forall i \in N$ a $\underline{\rho} > 0$. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak platí*

$$\lim_{i\to\infty} \|g_i\| = 0.$$

Důkaz V poznámce 167 jsme ukázali, že platí

$$\liminf_{i\to\infty} \|g_i\| = 0.$$

Předpokládejme, že

$$\limsup_{i\to\infty} \|g_i\| > \underline{\varepsilon} > 0.$$

Za tohoto předpokladu musí být množina N_2 nekonečná a musí obsahovat nekonečnou podmnožinu $\overline{N}_2 \subset N_2$ takovou, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in \overline{N}_2$ (pokud by byla N_2 konečná, existoval by index $k \in N$ takový, že $x_i = x_k$ $\forall i \geq k$ a tudíž $\|g_i\| = 0$ $\forall i \geq k$). Předpokládejme pro jednoduchost, že $N_2 = N$ (v opačném případě můžeme posloupnost N_2 přečíslovat) a označme

$$\overline{N}_2 = \{k_1, k_2, k_3, \ldots\}.$$

Jelikož posloupnost $F(x_{k_j})$, $j \in N$, je podle (T2) nerostoucí a podle (F1) zdola omezená, má tato posloupnost limitu a existuje tedy index $m \in N$ takový, že

$$F(x_{k_j}) - F(x_{k_{j+1}}) < \underline{\rho}\, \frac{\underline{\sigma}\, \underline{\varepsilon}^2}{4\overline{\delta}^2} \min\left(\frac{1}{\overline{B}},\ \frac{1}{\overline{G}}\right), \quad \forall j \geq m.$$

Nechť l_j je nevětší index takový, že $k_j \leq l_j < k_{j+1}$ a $\|g_l\| \geq \underline{\varepsilon}/(2\overline{\delta})$ $\forall k_j \leq l \leq l_j$. Pak podle (T1) a (T2) platí

$$F(x_l) - F(x_{l+1}) > \underline{\rho}\, \underline{\sigma} \|g_l\| \min\left(\Delta_l, \frac{\|g_l\|}{\|B_l\|}\right) \geq \underline{\rho}\, \frac{\underline{\sigma}\, \underline{\varepsilon}}{2\overline{\delta}^2} \min\left(\|s_l\|, \frac{\underline{\varepsilon}}{2\overline{B}}\right), \quad \forall k_j \leq l \leq l_j,$$

takže

$$\begin{aligned} \underline{\rho}\, \frac{\underline{\sigma}\, \underline{\varepsilon}^2}{4\overline{\delta}^2} \min\left(\frac{1}{\overline{B}},\ \frac{1}{\overline{G}}\right) &> F(x_{k_j}) - F(x_{k_{j+1}}) \geq F(x_{k_j}) - F(x_{l_j+1}) \\ &= \sum_{l=k_j}^{l_j} (F(x_l) - F(x_{l+1})) > \underline{\rho}\, \frac{\underline{\sigma}\, \underline{\varepsilon}}{2\overline{\delta}^2} \sum_{l=k_j}^{l_j} \min\left(\|s_l\|, \frac{\underline{\varepsilon}}{2\overline{B}}\right). \end{aligned}$$

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 mensí 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 (20), 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 N_2 je nekonečná a platí $\liminf_{i\to\infty} \|g_i\| = 0$, musí existovat index $j \geq m$ takový, že $l_j + 1 < k_{j+1}$. 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 2). Abychom nemuseli stále ověřovat zda pro daný index $i \in N$ již platí $x_i \in \mathcal{B}(x^*, \varepsilon)$, nahradíme předpoklady věty 2 silnějšími předpoklady (F4) a (F5). Tím se formálně zjednoduší a zpřehlední většina důkazů aniž by došlo k újmě na obecnosti. Nejprve ukážeme, že jsou-li matice B_i, $i \in N$, dostatečně omezené a platí-li (F4), (F5) a

$$\Delta_{i+1} \leq \overline{\gamma}\|s_i\| \qquad \forall i \in N_3, \tag{246}$$

kde $\overline{\gamma}\,\underline{\delta} > 1$, jsou tyto matice stejnoměrně omezené.

Věta 77 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3), pro kterou platí (246). 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 podmínkám (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 Nechť $k \in N_3$ a $l \in N_3$ jsou dva indexy takové že $j \notin N_3\ \forall k < j < l$. Pak podle (T1a), (T3) a (246) platí

$$\|s_j\| \leq \overline{\delta}\Delta_j \leq \overline{\beta}\,\overline{\delta}\|s_{j-1}\| \leq \ldots \leq (\overline{\beta}\,\overline{\delta})^{j-k-1}\|s_{k+1}\| \leq \frac{1}{\overline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\Delta_{k+1} \leq \frac{\overline{\gamma}\,\overline{\gamma}}{\overline{\beta}}(\overline{\beta}\,\overline{\delta})^{j-k}\|s_k\|$$

$\forall k < j < l$, neboli

$$\sum_{j=k}^{l-1}\|s_j\| \leq \frac{\overline{\gamma}\,\overline{\gamma}}{\overline{\beta}}\|s_k\|\sum_{j=k}^{l-1}(\overline{\beta}\,\overline{\delta})^{j-k} \leq \frac{\underline{\gamma}\,\overline{\gamma}}{\overline{\beta}(1-\overline{\beta}\,\overline{\delta})}\|s_k\| = \overline{D}\|s_k\|,$$

kde $\overline{D} = \underline{\gamma}\,\overline{\gamma}/(\overline{\beta}(1-\overline{\beta}\,\overline{\delta})) > 1$, takže pro libovolný index $i \in N$ platí

$$C_1 + \sum_{j=1}^{i}\overline{C}\|s_j\| \leq \overline{D}(C_1 + \sum_{j\in N_3}^{i}\overline{C}\|s_j\|)$$

(předpokládáme bez újmy na obecnosti, že $1 \in N_3$). Nyní můžeme postupovat podobně jako v důkazu věty 13. Použijeme-li (T1), dostaneme

$$\frac{F_i - F_{i+1}}{\|g_i\|} \geq -\overline{\rho}\frac{Q_i(s_i)}{\|g_i\|} \geq \overline{\rho}\,\underline{\sigma}\min\left(\Delta_i, \frac{\|g_i\|}{C_i}\right) \geq \frac{\overline{\rho}\,\underline{\sigma}}{\overline{\delta}}\min\left(\|s_i\|, \frac{\|g_i\|}{C_i}\right) \geq \frac{\overline{\rho}\,\underline{\sigma}}{\overline{\delta}}\frac{\|g_i\|\|s_i\|}{\|g_i\| + C_i\|s_i\|}$$

$\forall i \in N_3$, neboť pro libovolná kladná čísla a, b platí $\min(a,b) \geq ab/(a+b)$. Dále podle (F4) platí

$$0 \geq F_{i+1} - F_i \geq s_i^T g_i + \frac{1}{2}\underline{G}\|s_i\|^2 \geq -\|s_i\|\|g_i\| + \frac{1}{2}\underline{G}\|s_i\|^2,$$

neboli

$$\|s_i\| \leq \frac{2}{\underline{G}}\|g_i\|$$

$\forall i \in N_3$ (bez újmy na obecnosti budeme předpokladat, že $\underline{G} \leq C_1$, takže $\underline{G} \leq C_i\ \forall i \in N_3$). Vrátíme-li se k původní nerovnosti, můžeme psát

$$\frac{F_i - F_{i+1}}{\|g_i\|} \geq \frac{\overline{\rho}\,\underline{\sigma}}{\overline{\delta}} \frac{\|g_i\|\|s_i\|}{\|g_i\| + \dfrac{2}{\underline{G}} C_i \|g_i\|} \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{G}}{3\overline{\delta}} \frac{\|s_i\|}{C_i} \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{G}}{3\overline{\delta}\,\overline{C}} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|s_j\|} \geq \frac{\overline{\rho}\,\underline{\sigma}\,\underline{G}}{3\overline{\delta}\,\overline{C}\,\overline{D}} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j\in N_3}^{i} \overline{C}\|s_j\|}$$

$\forall i \in N_3$, takže jako v důkazu věty 13 platí

$$\frac{\overline{\rho}\,\underline{\sigma}\,\underline{G}}{3\overline{\delta}\,\overline{C}\,\overline{D}} \sum_{i\in N_3} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j\in N_3}^{i} \overline{C}\|s_j\|} \leq \sum_{i\in N_3} \frac{F_i - F_{i+1}}{\|g_i\|} \leq \sum_{i=1}^{\infty} \frac{F_i - F_{i+1}}{\|g_i\|} \leq \frac{\sqrt{2\overline{\overline{G}}}}{\underline{G}} \sqrt{F_1 - F^*}.$$

Existuje tedy číslo $\underline{C}$, takové, že

$$C_k \leq C_1 + \sum_{j=1}^{k-1} \overline{C}\|s_j\| \leq \overline{D}(C_1 + \sum_{j\in N_3}^{k-1} \overline{C}\|s_j\|) \leq \overline{D}C_1/\underline{C}$$

$\forall k \in N$ (viz důkaz věty 10), takže $\|B_k\| \leq C_k \leq \overline{B}\ \forall k \in N$, kde $\overline{B} = \overline{D}C_1/\underline{C}$. Z toho že $C_1 + \sum_{j=1}^{k-1} \overline{C}\|s_j\| \leq \overline{B}\ \forall k \in N$ plyne nerovnost

$$\sum_{i=1}^{\infty} \|s_i\| < \infty.$$

□

Poznámka 169 Podmínku (246) splníme snadno tak, že položíme $\Delta_{i+1} = \overline{\gamma}\|s_i\|$, pokud v (T3c) vyjde $\Delta_{i+1} > \overline{\gamma}\|s_i\|$. Poznamenejme, že jelikož $\overline{\gamma}\,\underline{\delta} > 1$, může tento případ nastat pouze tehdy, když $i \in N_1$. Můžeme se přesvědčit, že se touto úpravou neporuší platnost žádné z vět o globální konvergenci, ani platnost věty 79 o superlineární konvergenci. Věta 77 má význam při vyšetřování superlineární konvergence metod s proměnnou metrikou pro řídké úlohy (věta 151)).

Věta 78 *(lineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3) 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 $F \in \mathcal{C}^2 : \mathcal{D} \to R$, která vyhovuje podmínkám (F4) a (F5). Pak platí*

$$\sum_{i=1}^{\infty} \|x_{i+1} - x_i\| < \infty.$$

Důkaz (a) Dokážeme nejprve, že posloupnost x_i, $i \in N_3$, je lineárně konvergentní. Důkaz tohoto dílčího tvrzení je velmi podobný důkazu věty 14. Nechť $i \in N_3$. Podle lemmatu 28 existuje index $k \leq i$ takový, že $\|s_i\| \geq (\underline{c}/\overline{B})\|g_k\|$. Jelikož posloupnost $F(x_i)$, $i \in N$, je nerostoucí, platí

$$1 \geq \frac{F(x_i) - F(x^*)}{F(x_k) - F(x^*)} \geq \frac{\underline{G}^2}{2} \frac{\|x_i - x^*\|^2}{\|g_k\|^2} \geq \frac{\underline{G}^2}{2\overline{G}^2} \frac{\|g_i\|^2}{\|g_k\|^2}$$

(používáme vztahy (18)–(24)), takže

$$\|s_i\| \geq \frac{1}{\sqrt{2}} \frac{\underline{c}\,\underline{G}}{\overline{B}\,\overline{G}} \|g_i\|. \tag{247}$$

Jelikož $i \in N_3$, platí $\rho_i(s_i) \geq \overline{\rho}$, což spou s (T1c) dává

$$F_i - F_{i+1} \geq \overline{\rho}\,\underline{\sigma}\|g_i\|^2 \min\left(\frac{1}{\sqrt{2}} \frac{\underline{c}\,\underline{G}}{\overline{\delta}\,\overline{B}\,\overline{G}}, \frac{1}{\overline{\overline{B}}}\right) = \frac{\overline{\rho}}{\sqrt{2}} \frac{\underline{\sigma}\,\underline{c}\,\underline{G}}{\overline{\delta}\,\overline{B}\,\overline{G}} \|g_i\|^2 \geq \frac{c}{\overline{\overline{G}}} \|g_i\|^2.$$

kde

$$c = \frac{\overline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{G}}{2\overline{\delta}\,\overline{B}} < \frac{\overline{\rho}\,\underline{\sigma}\,\underline{G}}{\overline{\delta}\,\overline{G}} < 1,$$

neboť podle lemmatu 28 platí

$$\underline{c} \leq \frac{2\underline{\sigma}(1-\overline{\rho})\|B_1\|}{\overline{\delta}(\overline{G}+\|B_1\|)} < 2\frac{\overline{B}}{\overline{\overline{G}}}.$$

Použijeme-li ještě jednou vztah (24), dostaneme

$$F_{i+1} - F^* \leq \left(1 - c\frac{\underline{G}}{\overline{\overline{G}}}\right)(F_i - F^*) \quad \forall i \in N_3,$$

takže posloupnost x_i, $i \in N_3$, konverguje k bodu x^* R-lineárně, což dává

$$\sum_{i \in N_3} \|x_{i+1} - x_i\| < \infty.$$

(b) Jelikož $x_{i+1} = x_i$, pokud $i \notin N_2$, budeme bez újmy na obecnosti předpokládat, že $N_2 = N$ (v opačném případě můžeme posloupnost N_2 přečíslovat). Dále budeme předpokládat, že $1 \in N_3$ (důkaz pro $1 \notin N_3$ není principiálně složitější, jen se prodlouží). Nechť $i \in N_3$ a $k > i$ je index takový, že $j \notin N_3$ $\forall i < j \leq k$. Pak platí

$$\|s_j\| \leq \overline{\delta}\Delta_j \leq \overline{\beta}\,\overline{\delta}\|s_{j-1}\| \leq \ldots \leq (\overline{\beta}\,\overline{\delta})^{j-i-1}\|s_{i+1}\| \tag{248}$$

$\forall i < j \leq k$. Podle (247) a (21) platí $\|s_i\| \geq c_0\|g_i\| \geq c_0\underline{G}\|e_i\|$, kde $c_0 = (1/\sqrt{2})(\underline{c}\,\underline{G})/(\overline{B}\,\overline{G})$. Jelikož posloupnost $F(x_i)$, $i \in N$, je nerostoucí, pak pro libovolný index $j > i$ platí

$$1 \geq \frac{F(x_j) - F(x^*)}{F(x_i) - F(x^*)} \geq \frac{\underline{G}}{\overline{\overline{G}}}\frac{\|e_j\|}{\|e_i\|},$$

neboli $\|e_j\| \leq \sqrt{\overline{G}/\underline{G}}\,\|e_i\|$. Můžeme tedy psát

$$\|s_{i+1}\| = \|x_{i+2} - x_{i+1}\| \leq \|e_{i+2}\| + \|e_{i+1}\| \leq 2\sqrt{\frac{\overline{G}}{\underline{G}}}\|e_i\| \leq \frac{2}{c_0\underline{G}}\sqrt{\frac{\overline{G}}{\underline{G}}}\|s_i\| \triangleq \overline{c}\|s_i\|,$$

což po dosazení do (248) dává

$$\sum_{j=i}^{k}\|s_j\| \leq \|s_i\| + \sum_{j=i+1}^{k}\overline{c}(\overline{\beta}\,\overline{\delta})^{j-i-1}\|s_i\| \leq \left(1 + \frac{\overline{c}}{1-\overline{\beta}\,\overline{\delta}}\right)\|s_i\|,$$

takže podle (a) platí

$$\sum_{i=1}^{\infty}\|x_i - x^*\| = \sum_{i \in N_2}\|s_i\| \leq \left(1 + \frac{\overline{c}}{1-\overline{\beta}\,\overline{\delta}}\right)\sum_{i \in N_3}\|s_i\| < \infty.$$

□

Poznámka 170 Ve větě 78 nepotřebujeme, aby byla splněna podmínka (246). Vyžadujeme však aby matice B_i, $i \in N$, byly stejnoměrně omezené. Platí-li (246), můžeme použít silnější tvrzení věty 77.

Věta 79 *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3) taková, že $x_i \to x^*$. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F4) a (F5). Nechť*

$$\lim_{i\to\infty}\overline{\omega}_i = 0, \tag{249}$$

$$\lim_{i\to\infty}\frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0. \tag{250}$$

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 $k_1 \in N$ takový, že

$$\|g_i\| \geq \frac{1}{2}\underline{G}\|s_i\|$$

a

$$-Q_i(s_i) \geq \frac{\underline{\sigma}\,\underline{G}^2}{4\overline{G}}\|s_i\|^2$$

$\forall i \geq k_1$. 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

$$\|B_i s_i\| \leq \overline{\lambda}(G_i)\|s_i\| + \|\vartheta_i\|\|s_i\|,$$

$$s_i^T B_i s_i \geq \underline{\lambda}(G_i)\|s_i\|^2 - \|\vartheta_i\|\|s_i\|^2$$

a jelikož $\|\vartheta_i\| \to 0$ (podle (250)) 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) a přihlédneme-li k poznámce **??**, můžeme psát

$$-Q_i(s_i) \geq \underline{\sigma}\|g_i\| \min(\|s_i\|, \|g_i\|/\overline{G}) \geq \frac{\underline{\sigma}\,\underline{G}}{2}\min(1, \frac{\underline{G}}{2\overline{G}})\|s_i\|^2 = \frac{\underline{\sigma}\,\underline{G}^2}{4\overline{G}}\|s_i\|^2.$$

(b) Ukážeme, že existuje index $k_2 \geq k_1$ tak, že $i \in N_3$ $\forall i \geq k_2$. Podle věty 3 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 (a) však platí

$$\left|\frac{s_i^T(G_i - B_i)s_i + o(\|s_i\|^2)}{2Q_i(s_i)}\right| \leq \frac{2\overline{G}}{\underline{\sigma}\,\underline{G}^2}\frac{\|\vartheta_i\|\|s_i\|^2 + o(\|s_i\|^2)}{\|s_i\|^2} \to 0,$$

neboť $\|\vartheta_i\| \to 0$. Platí tedy $\rho_i(s_i) \to 1$ a jelikož $\overline{\rho} < 1$, existuje index $k_2 \geq k_1$ takový, že $\rho_i(s_i) \geq \overline{\rho}$ $\forall i \geq k_2$.

(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 $i \in N_3$ $\forall i \geq k \geq k_2$. To je však spor, neboť podle (a) platí $\|s_i\| \leq 2\|g_i\|/\underline{G}$, takže $\|g_i\| \to 0$ 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 (249), (250) a (T1b) platí $\|\omega_i\| \overset{N_1}{\to} 0$ a $\|\vartheta_i\| \overset{N_1}{\to} 0$, takže stejným způsobem jako v důkazu věty 16 se dá ukázat, že existuje index $k_3 \geq k_2$, $k_3 \in N_1$, takový, že

$$\underline{G}\|s_i\| \leq \|g_i\| \leq \overline{G}\|s_i\|$$

$\forall i \geq k_3$, $i \in N_1$. Použijeme-li větu 3, 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) $i \in N_3 \subset N_2$ pokud $i \geq k_3 \geq k_2$. Označme

$$\lambda_i = \frac{g_{i+1} - g_i - B_i s_i}{\|g_i\|} = -\frac{\vartheta_i \|s_i\| + o(\|s_i\|)}{\|g_i\|}$$

(pro $i \geq k_3$). Pak z nerovnosti $\underline{G}\|s_i\| \leq \|g_i\|$, platící pro $i \geq k_3$, $i \in N_1$, plyne, že $\|\lambda_i\| \leq \|\vartheta_i\| + o(1))/\underline{G} \overset{N_1}{\to} 0$. Jelikož zároveň $\|\omega_i\| \overset{N_1}{\to} 0$, existuje index $k \geq k_3$, $k \in N_1$ takový, že $\|\lambda_i\| < (\underline{G}/\overline{G})/(2\overline{\delta})$ a $\|\omega_i\| < (\underline{G}/\overline{G})/(2\overline{\delta})$ $\forall i \geq k$, $i \in N_1$. Pak pro $i \geq k$ 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) $i \in N_3 \subset N_2$, pokud $i \geq k$, 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) Superlineární konvergence. 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

$$\lim_{i\to\infty} \frac{\|x_{i+1} - x^*\|}{\|x_i - x^*\|} \leq \frac{\overline{G}}{\underline{G}} \lim_{i\to\infty} \frac{\|g_{i+1}\|}{\|g_i\|} = 0.$$

□

5.2 Metody s optimálním lokálně omezeným krokem

Definice 26 *Metody s optimálním lokálně omezeným krokem používají směrový vektor*

$$s_i^* = \arg \min_{\|s\| \leq \Delta_i} Q_i(s), \tag{251}$$

přičemž $\|s_i^*\| = \Delta_i$, pokud toto minimum není jediné.

Vektor s_i^* určený podle (251) 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

$$s_i(\alpha^*) = \arg \min_{\|s(\alpha)\| \leq \Delta_i} Q_i(s(\alpha)), \tag{252}$$

kde $s(\alpha) = -\alpha g$.

Lemma 30 *Směrový vektor $s_i(\alpha^*) \in R^n$ určený podle (252), který lze vyjádřit ve tvaru*

$$s_i(\alpha^*) = -\frac{g_i^T g_i}{g_i^T B_i g_i} g_i, \quad g_i^T B_i g_i \geq \|g_i\|^3/\Delta_i, \tag{253}$$

$$s_i(\alpha^*) = -\frac{\Delta_i}{\|g_i\|} g_i, \qquad g_i^T B_i g_i < \|g_i\|^3/\Delta_i, \tag{254}$$

vyhovuje podmínce (T1c) s $\underline{\sigma} = 1/2$.

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 (252). Navíc platí

$$-Q_i(s_i(\alpha^*)) = \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} - \frac{1}{2}\frac{(g_i^T g_i)^2 g_i^T B_i g_i}{(g_i^T B_i g_i)^2} = \frac{1}{2}\frac{(g_i^T g_i)^2}{g_i^T B_i g_i} \geq \frac{1}{2}\|g_i\|^2/\|B_i\|.$$

(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 (252) 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 171 Lemma 30 zdůvodňuje volbu podmínky (T1c), neboť ukazuje, že lze najít vektor, který této podmínce vyhovuje.

Poznámka 172 Vektor $s_i(\alpha^*)$, který je řešením úlohy (252), se často nazývá Cauchyovým krokem.

Věta 80 *Směrový vektor $s_i^* \in R^n$ určený podle (251) vyhovuje podmínkám (T1a)–(T1c) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\sigma} = 1/2$.*

Důkaz (a) Podmínka (T1a) je přímo součástí podmínky (251). Předpokládejme, že $s_i^* \in R^n$ je řešením úlohy (251), 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

$$-Q_i(s_i^*) = g_i^T B_i^{-1} g_i - \frac{1}{2} g_i^T B_i^{-1} g_i = \frac{1}{2} g_i^T B_i^{-1} g_i \geq \frac{1}{2}\|g_i\|^2/\|B_i\|.$$

(b) Nechť $\|s_i^*\| = \Delta_i$. Podle (251) musí být $Q_i(s_i^*) \leq Q_i(s_i(\alpha^*))$, takže nutně

$$-Q_i(s_i^*) \geq -Q_i(s_i(\alpha^*)) \geq \frac{1}{2}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|).$$

□

Nyní uvedeme důležitou větu, která charakterizuje řešení úlohy (251).

Věta 81 *Vektor $s_i^* \in R^n$ je řešením úlohy (251) právě tehdy, jestliže $\|s_i^*\| \leq \Delta_i$ a jestliže 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 (251) pouze přibližně. Proto zavádíme pojem kvazioptimálních metod s lokálně omezeným krokem.

Definice 27 *Metody s kvazioptimálním lokálně omezeným krokem používají místo podmínky (T1c) podmínku*

$$Q_i(s_i) \leq \underline{\nu} Q_i(s_i^*) \tag{255}$$

s $0 < \underline{\nu} \leq 1$, kde vektor s_i^* je řešením úlohy (251).

Poznámka 173 Podle definice 27 a věty 80 splňuje metoda s kvazioptimálním lokálně omezeným krokem podmínku (T1c) s $\underline{\sigma} = \underline{\nu}/2$.

5.3 Newtonova metoda s lokálně omezeným krokem

Newtonova metoda používá matice $B_i = G(x_i), i \in N$, takže z (F4) plyne $\|B_i\| = \|G(x_i)\| \leq \overline{G}, i \in N$.

Věta 82 *Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F4). Pak Newtonova metoda realizovaná jako metoda s lokálně omezeným krokem je globálně konvergentní. Je-li navíc splněna podmínka (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 75 (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 79). □

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 zejéna 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 83 *Nechť x_i, $i \in N$, je posloupnost určená Newtonovou metodou s kvazioptimálním lokálně omezeným krokem. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1), (F2) a (F4). Pak existuje hromadný bod $x^* \in R^n$ posloupnosti x_i, $i \in N$, takový, že $g(x^*) = 0$ a $G(x^*) \succeq 0$. Nechť navíc bod $x^* \in R^n$ vyhovuje postačujícím podmínkám pro extrém ($g(x^*) = 0$ a $G(x^*) \succ 0$). Pak $x^* \in R^n$ je 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

$$\Delta_i \to 0 \quad \text{a} \quad i \notin N_3 \quad \forall i \in M \tag{256}$$

(neboť $\overline{\gamma} = \infty$). Ve druhém případě existuje podposloupnost x_i, $i \in M \subset N$, taková, že

$$\Delta_i \geq \underline{\Delta} \quad \text{a} \quad i \in N_3 \quad \forall i \in M, \tag{257}$$

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 (256) 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

$$|Q_i(s_i)| \geq \underline{\sigma}\|g_i\| \min\left(\Delta_i, \frac{\|g_i\|}{\|G_i\|}\right) \geq \underline{\sigma}\frac{\|g^*\|}{2}\Delta_i \geq \frac{\underline{\sigma}\|g^*\|}{2\overline{\delta}}\|s_i\| \quad \forall i \in M,\ i \geq k,$$

což s použitím věty 3 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 $i \notin N_3$.

(c) Předpokládejme, že platí (256) 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 (255) a (251) platí

$$|Q_i(s_i)| \geq \underline{\nu}|Q_i(s_i^*)| \geq \underline{\nu}|Q_i(v_i)| = -\underline{\nu}(v_i^T g_i + \frac{1}{2}v_i^T G_i v_i) \geq -\frac{\underline{\nu}}{2}\underline{\lambda}_i\Delta_i^2 \geq \frac{\underline{\nu}}{4\overline{\delta}^2}|\underline{\lambda}^*|\|s_i\|^2 \quad \forall i \in M,\ i \geq k,$$

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 $i \notin N_3$.

(d) Předpokládejme, že platí (257). 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 $M \subset N_3$, také $Q_i(s_i) \overset{M}{\to} 0$. Nechť

$$s^* = \arg \min_{\|s\|\leq \underline{\Delta}/2} Q^*(s), \tag{258}$$

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 (258), 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 ($\overline{\text{F4}}$) 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 (256) nebo (257)). 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 splněna podmínka (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

$$|Q_i(s_i)| \geq \frac{\underline{\sigma}\,\underline{G}^2}{4\overline{\delta}\,\overline{G}}\|s_i\|^2.$$

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 $i \in N_3$, pokud $i \in M$, $i \geq k_1$. Tím jsme eliminovali případ (256).

(f) Nechť $g(x^*) = 0$ a $G(x^*) \succ 0$ a nechť platí (257). 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 $i \in N_1 \cap N_3$ a použijeme-li větu 3 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 (20) a (21) z důkazu věty 12) 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 (257). 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 vyniká v případě úloh s řídkými Hessovými maticemi jak bude ukázáno v sedmé kapitole.

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. 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 28 *Nemonotonní metody s lokálně omezeným krokem se liší od standardních metod s lokálně omezeným krokem (definice 25) pouze tím, že podmínku (T2) nahradíme podmínkou*

$$\overline{\rho}_i(s_i) \leq \underline{\rho} \Rightarrow \alpha_i = 0, \tag{T2c}$$

$$\overline{\rho}_i(s_i) > \underline{\rho} \Rightarrow \alpha_i = 1, \tag{T2d}$$

kde

$$\overline{\rho}_i(s_i) = \frac{F(x_i + s) - \overline{F}_i}{Q_i(s)}$$

a $0 < \underline{\rho} < \overline{\rho} < 1$ *(podmínky (T1) a (T3) zůstanou zachovány). Množinu indexů, pro které platí (T2d) označíme* $\overline{N}_2$.

Nejprve vyšetříme jednoduchou nemonotonní metodu s lokálně omezeným krokem, pro kterou platí

$$\overline{F}_i = \max\{F_j : i - \min(m, i) + 1 \leq j \leq i\}, \tag{259}$$

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

Věta 84 *(Globální konvergence metody (259)) Nechť* $x_i \in R^n$, $i \in N$, *je posloupnost generovaná nemonotonní metodou s lokálně omezeným krokem (259) taková, že* $\|B_i\| \leq \overline{B}$ $\forall i \in N$. *Nechť funkce* $F : \mathcal{D} \to R$ *splňuje podmínky (F1) a (F3). Pak platí*

$$\liminf_{i \to \infty} \|g_i\| = 0. \tag{260}$$

Důkaz Jelikož akceptujeme podmínky (T1) a (T3), můžeme použít lemma 28, podle kterého platí

$$\Delta_i \geq \frac{1}{\overline{\delta}}\|s_i\| \geq \frac{\underline{c}\, m_i}{\overline{\delta}\,\overline{B}}, \quad m_i = \min_{1\leq j\leq i}\|g_j\|.$$

(a) Předpokládejme, že není splněna podmínka (260). Pak musí existovat číslo $\underline{\varepsilon} > 0$ takové, že $m_i \geq \underline{\varepsilon}$ $\forall i \in N$, což spolu s předchozí nerovností dává

$$\Delta_i \geq \frac{\underline{c}\,\underline{\varepsilon}}{\overline{\delta}\,\overline{B}} > 0, \quad \forall i \in N. \tag{261}$$

Musí tedy existovat nekonečná podmnožina $N_3 \subset N$ indexů, pro které platí (T3). Zřejmě $N_3 \subset \overline{N}_2$, neboť $\overline{\rho}_i \geq \rho_i$ a $0 < \underline{\rho} < \overline{\rho} < 1$. Množina $\overline{N}_2$ je tedy také nekonečná a existuje její nekonečná podmnožina $N_4 = \{i_1, i_2, i_3, \ldots\} \subset \overline{N}_2$ taková, že $i_{k+1} - i_k \geq m$ $\forall k \in N$.

(b) Ukážeme, že není-li splněna podmínka (260), platí

$$F_{i_k+j} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}} \quad \forall j \in N, \tag{262}$$

kde $\underline{\varepsilon}$ je číslo použité v (261). Důkaz provedeme indukcí. Nechť $i_k \in N_4$ a $j \in N$. Budeme předpokládat, že buď $i_k + j - 1 \in N_4$ (což je splněno například pro $j = 1$), nebo $i_k + j - 1 \notin N_4$ a platí (262) s $j - 1$ místo j (indukční předpoklad). V prvním případě použitím (T1c), (T2d) a (261) dostaneme

$$\begin{aligned} F_{i_k+j} &\leq \overline{F}_{i_k+j-1} + \underline{\rho} Q_{i_k+j-1} \\ &\leq \overline{F}_{i_k} - \underline{\rho}\,\underline{\sigma}\|g_{i_k+j-1}\| \min\left(\Delta_{i_k+j-1}, \frac{\|g_{i_k+j-1}\|}{\|B_{i_k+j-1}\|}\right) \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}}, \end{aligned}$$

neboť posloupnost $\overline{F}_i$, $i \in N$, je nerostoucí (viz důkaz věty 20). Ve druhém případě podle indukčního předpokladu platí

$$F_{i_k+j} = F_{i_k+j-1} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}}.$$

Tím je indukční krok proveden.

(c) Není-li splněna podmínka (260), platí

$$\overline{F}_{i_{k+1}} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}}$$

podle (259) a (262), neboť $i_{k+1} - i_k \geq m$. Můžeme tedy psát

$$\overline{F}_{i_1} - \underline{F} \geq \overline{F}_{i_1} - \lim_{k\to\infty} \overline{F}_{i_{k+1}} = \sum_{k=1}^{\infty} (\overline{F}_{i_k} - \overline{F}_{i_{k+1}}) \geq \sum_{k=1}^{\infty} \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}} = \infty,$$

což je spor, neboť výraz na levé straně této nerovnosti je podle (F1) konečný. □

Poznámka 174 Hodnotu (259) 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

$$\overline{\rho}_i(s_i) \leq \underline{\rho} \quad \Rightarrow \quad \tilde{\mathcal{F}}_{i+1} = \mathcal{F}_i, \tag{263}$$

$$\overline{\rho}_i(s_i) > \underline{\rho} \quad \Rightarrow \quad \tilde{\mathcal{F}}_{i+1} = \mathcal{F}_i \cup \{F_{i+1}\}. \tag{264}$$

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 (263)–(264) je také globálně konvergentní. Důkaz tohoto tvrzení je podobný důkazu věty 84 a přenecháme ho čtenáři.

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

$$\overline{\rho}_i(s_i) \leq \underline{\rho} \quad \Rightarrow \quad \overline{n}_{i+1} = \overline{n}_i, \qquad \overline{F}_{i+1} = \overline{F}_i \tag{265}$$

$$\overline{\rho}_i(s_i) > \underline{\rho} \quad \Rightarrow \quad \overline{n}_{i+1} = \lambda \overline{n}_i + 1, \quad \overline{F}_{i+1} = \frac{\lambda \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \tag{266}$$

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

Věta 85 *(Globální konvergence metody (265)–(266)) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná nemonotonní metodou s lokálně omezeným krokem (265)–(266) taková, že $\|B_i\| \leq \overline{B}$ $\forall i \in N$. Nechť funkce $F : \mathcal{D} \to R$ splňuje podmínky (F1) a (F3). Pak platí (260).*

Důkaz Předpokládejme, že není splněna podmínka (260). Podobně jako v části (a) důkazu věty 84 z toho plyne, že množina $\overline{N}_2 = \{i_1, i_2, i_3, \ldots\}$ je nekonečná a existuje číslo $\underline{\varepsilon} > 0$ takové, že platí $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$ a (261) a podobně jako v části (b) důkazu věty 84 dostaneme nerovnost

$$F_{i_k+1} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}}, \quad \forall k \in N,$$

která spolu s (T2c)–(T2d) a (265)–(266) dává

$$\overline{F}_{i_{k+1}} = \frac{\lambda \overline{n}_{i_k} \overline{F}_{i_k} + F_{i_k+1}}{\overline{n}_{i_{k+1}}} \leq \overline{F}_{i_k} - \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}}.$$

Můžeme tedy psát

$$\overline{F}_{i_1} - \underline{F} \geq \overline{F}_{i_1} - \lim_{k\to\infty} \overline{F}_{i_{k+1}} = \sum_{k=1}^{\infty} (\overline{F}_{i_k} - \overline{F}_{i_{k+1}}) \geq \sum_{k=1}^{\infty} \frac{\underline{\rho}\,\underline{\sigma}\,\underline{c}\,\underline{\varepsilon}^2}{\overline{\delta}\,\overline{B}} = \infty,$$

což je spor, neboť výraz na levé straně této nerovnosti je podle (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

$$\lambda_i = 1, \qquad \|H_i g_i\| \leq \Delta_i, \tag{267}$$

$$\lambda_i = \Delta_i / \|H_i g_i\|, \quad \|H_i g_i\| > \Delta_i, \tag{268}$$

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 31 *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 (267)–(268). Pak platí*

$$-Q_i(s_i) \geq \frac{1}{2}\|g_i\|\Delta_i \frac{g_i^T H_i g_i}{\|g_i\|\|H_i g_i\|} = \frac{1}{2\sqrt{\kappa(H_i)}}\|g_i\|\Delta_i \tag{269}$$

Jestliže $\kappa(H_i) \leq \overline{\kappa}$, je splněna podmínka (T1c) s $\underline{\sigma} = 1/(2\sqrt{\overline{\kappa}})$.

Důkaz Jelikož $H_i = B_i^{-1}$, platí

$$Q_i(s_i) = \frac{1}{2}\lambda_i^2 g_i^T H_i g_i - \lambda_i g_i^T H_i g_i = \frac{1}{2}\lambda_i(\lambda_i - 2) g_i^T H_i g_i, \tag{270}$$

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 8 dává (269). Zbytek tvrzení je zřejmý. □

Důsledek 14 *Uvažujme metodu s lokálně omezeným krokem, pro kterou platí $s_i = -\lambda_i H_i g_i$, (267)–(268), (T2) a (T3). Pak splňuje-li funkce $F : \mathcal{D} \to R$ podmínky (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 (260)).*

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 31 splněna podmínka (T1c). Jelikož také $\|B_i\| \leq 1/\underline{H}$ $\forall i \in N$, jsou splněny předpoklady věty 75, takže platí (260). □

Poznámka 175 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 definitnosti a podmíněnosti matic B_i, $i \in N$). Tuto úpravu však můžeme použít v případě 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 (270) pomocí matice H_i.

Jiná kombinobvaná metoda je založena na tom, že se jednoduchý výběr délky kroku (T2a) nahradí složitějším výběrem vyžadujícím splnění podmínky

$$\rho_i(\alpha_i s_i) > \underline{\rho}. \tag{271}$$

(obvykle $\underline{\rho} = 0$). Používá se přitom modifikovaný Armijův výběr délky kroku, kdy $\alpha_i > 0$ je první člen vyhovující podmínce (271) v posloupnosti α_i^j, $j \in N$, takové, že $\alpha_i^1 = 1$, a

$$\underline{\beta}\alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta}\alpha_i^j \quad \forall j \in N, \tag{272}$$

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 29 *Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$, $i \in N$, je kombinovanou metodou s lokálně omezeným krokem, jestliže směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že*

$$\|s_i\| \leq \overline{\delta}\Delta_i, \tag{T1a}$$

$$\|s_i\| < \underline{\delta}\Delta_i \;\Rightarrow\; \|\omega_i(s_i)\| \leq \overline{\omega}_i \leq \overline{\omega}, \tag{T1b}$$

$$-Q_i(s_i) \;\geq\; \underline{\sigma}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|), \tag{T1c}$$

$$-g_i^T s_i \;\geq\; \overline{\sigma}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|), \tag{T1d}$$

kde $0 < \underline{\delta} < 1 < \overline{\delta}$, $0 < \underline{\sigma} < 1$, $0 < \overline{\sigma} < 1$ a $0 \leq \overline{\omega} < 1$, délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq \underline{\rho} \;\Rightarrow\; \rho_i(\alpha_i s_i) > \underline{\rho}, \tag{273}$$

$$\rho_i(s_i) > \underline{\rho} \;\Rightarrow\; \alpha_i = 1, \tag{274}$$

kde $\underline{\rho} \geq 0$, a čísla $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se volí tak, že

$$\rho_i(s_i) \leq \underline{\rho} \;\Rightarrow\; \max(\alpha_i, \underline{\beta})\|s_i\| \leq \Delta_{i+1} \leq \overline{\beta}\|s_i\|, \tag{275}$$

$$\underline{\rho} < \rho_i(s_i) < \overline{\rho} \;\Rightarrow\; \underline{\beta}\|s_i\| \le \Delta_{i+1} \le \overline{\beta}\|s_i\|, \tag{276}$$

$$\rho_i(s_i) \ge \overline{\rho} \;\Rightarrow\; \Delta_i \le \Delta_{i+1} \le \min(\underline{\gamma}\Delta_i, \overline{\Delta}), \tag{277}$$

kde $0 < \underline{\beta} \le \overline{\beta} < 1 < \underline{\gamma}$, *a* $0 \le \underline{\rho} < \overline{\rho} < 1$, *přičemž* $\overline{\beta}\,\overline{\delta} < 1$.

Poznámka 176 Kombinovaná metoda s lokálně omezeným krokem se od standardní metody s lokálně omezeným krokem (definice 25) liší tím, že směrový vektor musí splňovat dodatečnou podmínku (T1d), že podmínka (T2a) je nahražena podmínkou (273) (Armijův výběr délky kroku) a že k podmínkám (276), (277) přibude podmínka (275), která se vztahuje k případu (273). 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 $\rho_i(s_i) \le \underline{\rho}$.

Poznámka 177 Definice 14 je korektní (podmínky (273)–(277) lze splnit). Předně z $\|s_i\| > 0$, $\|g_i\| > 0$ a (255) plyne, že $g_i^T s_i < 0$, takže

$$\lim_{\alpha\to 0} \rho_i(\alpha s_i) \ge \lim_{\alpha\to 0} \frac{\alpha g_i^T s_i - \alpha^2\overline{G}\|s_i\|^2/2}{\alpha g_i^T s_i + \alpha^2 s_i^T B_i s_i/2} = \lim_{\alpha\to 0} \frac{g_i^T s_i - \alpha\overline{G}\|s_i\|^2/2}{g_i^T s_i + \alpha s_i^T B_i s_i/2} = 1,$$

a jelikož $\underline{\rho} < 1$, existuje číslo $\overline{\alpha}_i > 0$ takové, že $\rho_i(\alpha_i s_i) > \underline{\rho}$, pokud $0 < \alpha_i \le \overline{\alpha}_i$. Dále z (272) plyne, že $\alpha_i \le \overline{\beta}$, pokud $\rho_i(s_i) \le \underline{\rho}$, takže lze splnit nerovnost v (275).

V dalších úvahách budeme používat indexové množiny N_1, N_2, N_3, které mají stejný význam jako v poznámce 166.

Věta 86 *Nechť* $x_i \in R^n$, $i \in N$, *je posloupnost generovaná kombinovanou metodou s lokálně omezeným krokem (definice 14) taková, že*

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty, \quad M_i = \max_{1\le j\le i} \|B_j\|.$$

Nechť funkce $F : \mathcal{D} \to R$ *splňuje podmínky (F1) a (F3). Pak platí (260).*

Důkaz V případech, kdy $i \in N_1$, $i \notin N_3$, $i = 1$, můžeme postupovat stejně jako v částech (a), (b), (c) důkazu lemmatu 28. V případě, kdy $i \in N_3$, můžeme, tak jako v části (d) zmíněného důkazu, využít toho, že podle (275) a (276) platí $\Delta_{i+1} \ge \underline{\beta}\Delta_i$, pokud $i \notin N_3$. Platí tedy tvrzení analogické lemmatu 28, takže je splněna nerovnost (236). Jelikož podle (275) a (276) platí $\Delta_{i+1} \le \overline{\beta}\|s_i\|$, pokud $i \notin N_3$, lze použít lemma 29. Zbytek důkazu je totožný s důkazem věty 75. □

Zbývá ukázat, jak lze nalézt směrový vektor vyhovující podmínkám (T1a)–(T1d). Metody, které budeme vyšetřovat splňují podmínky (T1a)–(T1c) (lemma 30, věta 80). Proto stačí ukázat, že platí i (T1d).

Lemma 32 *Směrový vektor* $s_i(\alpha^*) \in R^n$ *vyšetřovaný v lemmatu 30 vyhovuje podmínce (T1d) s* $\overline{\sigma} = 1$.

Důkaz Pokud $g_i^T B_i g_i \ge \|g_i\|^3/\Delta_i$, platí

$$-g_i^T s_i(\alpha^*) = \frac{g_i^T g_i}{g_i^T B_i g_i} g_i^T g_i \le \Delta_i\|g_i\|.$$

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 87 *Směrový vektor* $s_i^* \in R^n$ *určený řešením úlohy (251) vyhovuje podmínce (T1d) s* $\overline{\sigma} = 1/4$.

Důkaz (a) Podle věty 81 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

$$\begin{aligned} -g_i^T s_i &= g_i^T (B_i + \lambda_i^* I)^{-1} g_i \geq \frac{\|g_i\|^2}{\overline{\lambda}(B_i + \lambda_i^* I)} \geq \frac{\|g_i\|^2}{\|B_i\| + \lambda_i^*} \geq \frac{\|g_i\|^2}{2\|B_i\| + \|g_i\|/\Delta_i} \\ &\geq \frac{\|g_i\|^2}{4\max(\|B_i\|, \|g_i\|/\Delta_i)} = \frac{1}{4}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|). \end{aligned}$$

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 92, věta 93).

6 Výpočet lokálně omezeného kroku

6.1 Výpočet optimálního lokálně omezeného kroku

Tvrzení věty 81 tvoří základ algoritmu, založeného na hledání čísla $\lambda_i > 0$ takového, 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 88 *Nechť $\lambda \geq 0$, $B + \lambda I \succeq 0$ a $\|s\| \geq \underline{\delta}\Delta$, kde $(B + \lambda I)s + g = 0$. Pak je splněna podmínka (255) 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

$$\begin{aligned} Q(s + z) &= g^T(s + z) + \frac{1}{2}(s + z)^T B(s + z) = -s^T(B + \lambda I)(s + z) + \frac{1}{2}(s + z)^T B(s + z) \\ &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda(s + z)^T(s + z)\right) + \frac{1}{2}z^T(B + \lambda I)z. \end{aligned}$$

Nechť $s^* = s + z^*$. Pak $(s + z^*)^T(s + z^*) = (s^*)^T s^* \leq \Delta^2$ a $(z^*)^T(B + \lambda I)z^* \geq 0$, takže podle předchozí rovnosti dostaneme

$$\begin{aligned} Q(s^*) &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda(s + z^*)^T(s + z^*)\right) + \frac{1}{2}(z^*)^T(B + \lambda I)z^* \\ &\geq -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda\Delta^2\right), \end{aligned}$$

což po dosazení do úvodní nerovnosti dává dokazované tvrzení. □

Číslo $\lambda > 0$ vyhovující předpokladům věty 88 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 a konvexity) je pro tento účel rovnice $\phi(\lambda) = 0$, kde $\phi(\lambda) = 1/\Delta - 1/\|s(\lambda)\|$. Tato rovnice se řeší pomocí Newtonovy metody.

Lemma 33 *Nechť $\phi(\lambda) = 1/\Delta - 1/\|s(\lambda)\|$, kde $\lambda \geq 0$, $B + \lambda I \succ 0$ a $(B + \lambda I)s(\lambda) + g = 0$. Pak platí*

$$\phi'(\lambda) = -\frac{s(\lambda)^T(B + \lambda I)^{-1}s(\lambda)}{\|s(\lambda)\|^3}$$

a $\phi''(\lambda) \geq 0$.

Důkaz Derivováním rovnosti $(B + \lambda I)s(\lambda) + g = 0$ dostaneme $(B + \lambda I)s'(\lambda) + s(\lambda) = 0$, takže $s'(\lambda) = -(B + \lambda I)^{-1}s(\lambda)$, a podle definice funkce $\phi(\lambda)$ platí

$$\phi'(\lambda) = -\frac{s(\lambda)^T s'(\lambda)}{\|s(\lambda)\|^3} = -\frac{s(\lambda)^T(B + \lambda I)^{-1}s(\lambda)}{\|s(\lambda)\|^3}.$$

Dalším derivováním dostaneme $(B + \lambda I)s''(\lambda) + 2s'(\lambda) = 0$, takže $s''(\lambda) = -2(B + \lambda I)^{-1}s'(\lambda)$ a

$$\phi''(\lambda) = -\frac{s(\lambda)^T s''(\lambda) + s'(\lambda)^T s'(\lambda)}{\|s(\lambda)\|^3} - \frac{3(s(\lambda)^T s'(\lambda))^2}{\|s(\lambda)\|^5} = 3\frac{\|s(\lambda)\|^2\|s'(\lambda)\|^2 - (s(\lambda)^T s'(\lambda))^2}{\|s(\lambda)\|^5}$$

a podle Schwarzovy nerovnosti pak platí $\phi''(\lambda) \geq 0$. □

Důsledek 15 *Nechť jsou splněny předpoklady lemmatu 33. Nechť λ_i, $1 \le i \le n$, jsou vlastní čísla matice B (seřazená vzestupně) a v_i, $1 \le i \le n$, jim odpovídající ortonormální vlastní vektory. Nechť $g = \sum_{1=1}^{n} \gamma_i v_i$ (takže $\gamma_i = v_i^T g$). Pak platí*

$$\phi(\lambda) = \frac{1}{\Delta} - \frac{1}{\sqrt{\sum_{1=1}^{n} \frac{\gamma_i^2}{(\lambda_i+\lambda)^2}}}$$

a

$$\phi'(\lambda) = -\frac{\sum_{1=1}^{n} \frac{\gamma_i^2}{(\lambda_i+\lambda)^3}}{\left(\sqrt{\sum_{1=1}^{n} \frac{\gamma_i^2}{(\lambda_i+\lambda)^2}}\right)^3}.$$

Důkaz Jelikož matice B je symetrická, existuje rozklad $V^T B V = \Lambda$, kde $V = [v_1, \ldots, v_n]$ (takže $V^T V = I$) a $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. Podle lemmatu 33 tedy platí

$$\phi(\lambda) = \frac{1}{\Delta} - \frac{1}{\|s(\lambda)\|} = \frac{1}{\Delta} - \frac{1}{\sqrt{g^T V(\Lambda + \lambda I)^{-2} V^T g}}$$

a

$$\phi'(\lambda) = -\frac{g^T V(\Lambda + \lambda I)^{-3} V^T g}{\left(\sqrt{g^T V(\Lambda + \lambda I)^{-2} V^T g}\right)^3}.$$

Využijeme-li toho, že $g = V\tilde{g}$, kde $\tilde{g}_i = \gamma_i$, $1 \le i \le n$, a ortogonalitu matice V, dostaneme dokazované tvrzení. □

Poznámka 178 Jelikož požadujeme, aby matice $B + \lambda^* I$ byla positivně semidefinitní, musí platit $\lambda^* \ge -\lambda_1$, 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 $\lambda^* = -\lambda_1$. Pokud v regulárním případě platí $-\lambda_1 < \lambda < \lambda^*$ (takže $\|s(\lambda)\| > \Delta$ a $\phi(\lambda) > 0$) je krok Newtonovy metody

$$\lambda_+ = \lambda + \frac{s(\lambda)^T (B + \lambda I)^{-1} s(\lambda)}{\|s(\lambda)\|^3} \left(\frac{1}{\Delta} - \frac{1}{\|s(\lambda)\|}\right)$$

dobře definován a platí $\lambda < \lambda_+ < \lambda^*$ (plyne to z konvexity funkce $\phi(\lambda)$). Pro $\lambda^* < \lambda$ (kdy $\|s(\lambda)\| < \Delta$ a $\phi(\lambda) < 0$) platí $\lambda_+ < \lambda^*$ a je třeba zajistit aby byla splněna podmínka $-\lambda_1 < \lambda_+$. To lze provést použitím mezí $\underline{\lambda} < \lambda^* < \overline{\lambda}$ aktualizovaných v každém kroku algoritmu (poznámka 182). Singulární případ může nastat jedině tehdy, jestliže $\gamma_1 = v_1^T g = 0$, neboť pro $\lambda^* = -\lambda_1$ platí

$$v_1^T g = -v_1^T (B + \lambda^* I) s(\lambda^*) = -v_1^T (B - \lambda_1 I) s(\lambda_1) = 0.$$

Poznámka 179 Jestliže $\gamma_1 = v_1^T g \neq 0$, lze se snadno přesvědčit (ze vzorců uvedených v důsledku 15), že platí

$$\lim_{\lambda \to \lambda_1} \phi(\lambda) = \frac{1}{\Delta}, \qquad \lim_{\lambda \to \lambda_1} \phi'(\lambda) = -\frac{1}{|\gamma_1|}$$

a

$$\lim_{\lambda \to \infty} \phi(\lambda) = -\infty, \qquad \lim_{\lambda \to \infty} \phi'(\lambda) = -\frac{1}{\|g\|}.$$

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 180 V singulárním případě nelze použít Newtonovu metodu, neboť $\phi(\lambda) < 0\ \forall \lambda > \lambda^* = \lambda_1$. 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 $\|s(\lambda^*)\| = \|s + \alpha v_1\| = \Delta$. 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 $\phi(\lambda) < 0$, 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 $\|s + z\| = \Delta$, vyhovuje podmínce (255). Kvantitativní vztahy udává následující věta.

Věta 89 *Nechť $\lambda \geq 0$, $B + \lambda I \succeq 0$ a $\|s + z\| = \Delta$, kde $(B + \lambda I)s + g = 0$ a*

$$z^T(B + \lambda I)z \leq (1 - \underline{\delta}^2)\left(s^T(B + \lambda I)s + \lambda\Delta^2\right).$$

Pak vektor $s + z$ vyhovuje podmínce (255) s $\underline{\nu} = \underline{\delta}^2$.

Důkaz Tak jako v důkazu věty 88 platí

$$Q(s + z) = \frac{1}{2}z^T(B + \lambda I)z - \frac{1}{2}\left(s^T(B + \lambda I)s + \lambda(s + z)^T(s + z)\right)$$

a použijeme-li předpoklady věty, dostaneme

$$Q(s + z) \leq -\frac{1}{2}\underline{\delta}^2\left(s^T(B + \lambda I)s + \lambda\Delta^2\right).$$

Spojením této nerovnosti s poslední nerovností v důkazu věty 88, dostaneme dokazované tvrzení. □

Poznámka 181 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

$$\alpha = \frac{\sqrt{(v^T s)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} - v^T s}{\|v\|^2} = \frac{\Delta^2 - \|s\|^2}{\sqrt{(v^T s)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} + v^T s}.$$

První vzorec volíme pokud $v^T s \leq 0$ a druhý v opačném případě. Oba vzorce se zjednoduší, pokud $\|v\| = 1$. Tyto vzorce lze snadno získat řešením kvadratické rovnice vzniklé roznásobením vztahu $\|s + \alpha v\| = \Delta$.

Poznámka 182 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

$$\underline{\lambda}(B) + \lambda^* \leq \frac{\|g\|}{\Delta} \leq \overline{\lambda}(B) + \lambda^*.$$

Jelikož $-\|B\| \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \|B\|$, můžeme položit

$$\underline{\lambda} = \frac{\|g\|}{\Delta} - \|B\| \leq \lambda^* \leq \frac{\|g\|}{\Delta} + \|B\| = \overline{\lambda}$$

(místo $-\|B\|$ a $\|B\|$ lze použí i jiné odhady pro vlastní čísla, například Gerschgorinovy kruhy). Dolní mez $\underline{\lambda}$ je třeba ještě upravit tak, aby platilo $\underline{\lambda} \geq 0$.

Poznámka 183 V počátečních krocích Newtonovy metody se může stát, že matice $B+\lambda I$ není positivně 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 32). Z nulovosti matice E lze zjistit positivní definitnost matice $B+\lambda I$ a věta 100 dává odhad čísla, které je třeba přičíst k λ.

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 5 Data $0 < \underline{\delta} < 1 < \overline{\delta}$ (obvykle $\underline{\delta} = 0.9$ a $\overline{\delta} = 1.1$), $\Delta > 0$.

Krok 1 Nechť $\underline{\mu}$ je maximální diagonální prvek matice $-B$. Položíme $\underline{\lambda} := \max(0,\ \underline{\mu},\ \|g\|/\Delta - \|B\|)$, $\overline{\lambda} := \|g\|/\Delta + \|B\|$ a $\lambda := \underline{\lambda}$.

Krok 2 Jestliže $\lambda < \underline{\lambda}$ položíme $\lambda := \sqrt{\underline{\lambda}\overline{\lambda}}$.

Krok 3 Určíme Gillův-Murrayův rozklad $B+\lambda I+E = R^T R$. 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 100), položíme $\underline{\mu} := \lambda - v^T(B+\lambda I)v$, $\underline{\lambda} := \max(\underline{\mu}, \underline{\lambda})$ a přejdeme na krok 2.

Krok 4 Určíme vektor $s \in R^n$ řešením rovnice $R^T R s + g = 0$. Jestliže $\|s\| > \overline{\delta}\Delta$, položíme $\underline{\lambda} := \lambda$ a přejdeme na krok 6. Jestliže $\underline{\delta}\Delta \le \|s\| \le \overline{\delta}\Delta$ 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 $v^T s \ge 0$ (tento vektor lze určit z rozkladu $R^T R$ způsobem, který používají programy knihovny LAPACK). Určíme číslo $\alpha \ge 0$ tak, aby platilo $\|s+\alpha v\| = \Delta$ (poznámka 181). Jestliže $\alpha^2\|Rv\|^2 \le (1-\underline{\delta}^2)(\|Rs\|^2 + \lambda\Delta^2)$, položíme $s := s+\alpha v$ a ukončíme výpočet. V opačném případě položíme $\underline{\mu} := \lambda - \|Rv\|^2$, $\underline{\lambda} := \max(\underline{\mu}, \underline{\lambda})$ a přejdeme na krok 6.

Krok 6 Určíme vektor $v \in R^n$ řešením rovnice $R^T v = s$ a položíme

$$\lambda := \lambda + \frac{\|s\|^2}{\|v\|^2}\left(\frac{\|s\|-\Delta}{\Delta}\right).$$

Pokud $\lambda < \underline{\lambda}$ položíme $\lambda := \underline{\lambda}$. Pokud $\lambda > \overline{\lambda}$ položíme $\lambda := \overline{\lambda}$. Přejdeme na krok 2

6.2 Využití směru největšího spádu (metody psí nohy)

Nevýhodou metod s optimálním lokálně omezeným krokem je nutnost řešení úlohy (251), 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 (251) často nahražuje úlohou

$$s_i = s_i(\alpha^*, \beta^*) = \arg \min_{\|s(\alpha,\beta)\| \le \Delta_i} Q_i(s(\alpha,\beta)), \tag{278}$$

kde

$$s(\alpha,\beta) = -(\alpha g_i + \beta B_i^{-1} g_i).$$

Věta 90 *Směrový vektor $s_i \in R^n$ určený podle (278) vyhovuje podmínkám (T1a)–(T1c) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\sigma} = 1/2$.*

Důkaz (a) Podmínka (T1a) je přímo součástí podmínky (278). Předpokládejme, že $s_i(\alpha^*, \beta^*) \in R^n$ je řešením úlohy (278), 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

$$-Q_i(s_i(\alpha^*,\beta^*)) = g_i^T B_i^{-1} g_i - \frac{1}{2} g_i^T B_i^{-1} g_i = \frac{1}{2} g_i^T B_i^{-1} g_i \geq \frac{1}{2}\|g_i\|^2/\|B_i\|.$$

(b) Nechť $\|s_i(\alpha^*,\beta^*)\| = \Delta_i$. Podle (278) 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 (252), takže podle lemmatu 30 platí

$$-Q_i(s_i(\alpha^*,\beta^*)) \geq -Q_i(s_i(\alpha^*)) \geq \frac{1}{2}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|).$$

□

Úloha (278) 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 (278) vyhovuje podle věty 90 podmínkám (T1a)–(T1c) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$ a $\underline{\sigma} = 1/2$ 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 (278) 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 91 *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*

$$s_C^T s_C / s_C^T s_N \leq \tau \leq 1, \tag{279}$$

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 $g^T Bg > 0$ a $(s_N - s_C)^T s_C \geq 0$. Prostým dosazením dostaneme

$$(s_N - s_C)^T s_C = \frac{g^T g}{(g^T Bg)^2} \left(g^T Bg g^T B^{-1} g - (g^T g)^2 \right), \tag{280}$$

takže nerovnost $(s_N - s_C)^T s_C \geq 0$ je splněna právě tehdy, jestliže $g^T Bg g^T B^{-1} g - (g^T g)^2 \geq 0$ (je-li matice B positivně definitní plyne tato nerovnost ze Schwarzovy nerovnosti). Musí tedy platit $g^T Bg g^T B^{-1} g > 0$, neboli

$$\frac{s_C^T s_C}{s_N^T s_C} = \frac{(g^T g)^2}{(g^T B^{-1} g g^T Bg)} > 0,$$

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 $g^T Bg > 0$ a číslo τ je určeno podle (279), můžeme psát

$$\begin{aligned}
(\tau s_N - s_C)^T s_C &= \frac{g^T g}{(g^T Bg)^2} \left(\tau g^T Bg g^T B^{-1} g - (g^T g)^2 \right) \geq 0, \qquad (281) \\
(\tau s_N - s_C)^T B(\tau s_N - s_C) &= \tau^2 g^T B^{-1} g - 2\tau \frac{(g^T g)^2}{g^T Bg} + \frac{(g^T g)^2}{g^T Bg} \\
&= \frac{\tau}{g^T Bg} \left(\tau g^T Bg g^T B^{-1} g - (g^T g)^2 \right) + (1 - \tau) \frac{(g^T g)^2}{g^T Bg} \geq 0,
\end{aligned}$$

přičemž poslední nerovnost je rovností právě tehdy, když $(s_N - s_C)^T s_C = 0$ (kdy nutně $\tau = 1$). Dále platí

$$\begin{aligned}
(\tau s_N - s_C)^T (g + Bs_C) &= (\tau s_N - s_C)^T (\tau g + Bs_C) + (1 - \tau)(\tau s_N - s_C)^T g \\
&= (\tau s_N - s_C)^T B(\tau B^{-1} g + s_C) + (1 - \tau)(\tau s_N - s_C)^T g \\
&= -(\tau s_N - s_C)^T B(\tau s_N - s_C) - (1 - \tau) \frac{g^T Bg}{g^T g} (\tau s_N - s_C)^T s_C,
\end{aligned}$$

takže pro derivaci kvadratické funkce $Q(s_C + \alpha(\tau s_N - s_C))$ dostaneme

$$\begin{aligned}
(\tau s_N - s_C)^T (g + B(s_C + \alpha(\tau s_N - s_C)) = & -(\tau s_N - s_C)^T B(\tau s_N - s_C)(1 - \alpha) \\
& - (1 - \tau) \frac{g^T Bg}{g^T g} (\tau s_N - s_C)^T s_C.
\end{aligned}$$

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 $g^T Bg > 0$ a $(s_N - s_C)^T s_C < 0$. Ze vztahu (280) plyne, že $g^T Bg g^T B^{-1} g - (g^T g)^2 < 0$. Platí tedy

$$\begin{aligned}
(s_N - s_C)^T B(s_N - s_C) &= g^T B^{-1} g - 2 \frac{(g^T g)^2}{g^T Bg} + \frac{(g^T g)^2}{g^T Bg} \\
&= \frac{1}{g^T Bg} \left(g^T Bg g^T B^{-1} g - (g^T g)^2 \right) < 0, \\
(s_N - s_C)^T (g + Bs_C) &= -(s_N - s_C)^T B(s_N - s_C) > 0,
\end{aligned}$$

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 91 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 181). 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 91 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 6 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 181), 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 92 *Směrový vektor získaný algoritmem 6 vyhovuje podmínkám (T1a)–(T1d) s* $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} = 0$, $\underline{\sigma} = 1/2$ *a* $\overline{\sigma} = 1$.

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 (252). Podle lemmatu 30 je tedy splněna podmínka (T1c) s $\underline{\sigma} = 1/2$ a podle lemmatu 32 platí (T1d) s $\overline{\sigma} = 1$. 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 91 platí $Q(s) \leq Q(s_C)$ a s použitím (281) 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 $\underline{\sigma} = 1/2$ a $\overline{\sigma} = 1$.

(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 91 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 $\underline{\sigma} = 1/2$ a $\overline{\sigma} = 1$. □

Poznámka 184 V algoritmu 6 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

$$Q(s) = g^T s + \frac{1}{2} s^T Bs.$$

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 $1 \leq i \leq n$. Můžeme používat větu 36, větu 38 a důsledek 3.

Chceme-li určit lokálně omezený krok pomocí metody sdružených gradientů, zastavujeme iterační proces nejen tehdy, když $\|g_i\| \leq \omega \|g\|$ (kde $0 < \omega \leq \overline{\omega} < 1$), ale také tehdy, když $\|s_i\| < \Delta$ a buď $p_i^T Bp_i \leq 0$ nebo $\|s_{i+1}\| \geq \Delta$. Pokud $p_i^T Bp_i \leq 0$, můžeme použít větu 38, 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 $\|s_{i+1}\| \geq \Delta$, můžeme použít větu 36, podle které $Q(s_i + \alpha p_i) < Q(s_i)$ a $g^T(s_i + \alpha p_i) < g^T s_i$ pro $0 < \alpha \leq \alpha_i$. V obou případech určíme číslo $\alpha_i \geq 0$ tak, aby platilo $\|s_i + \alpha_i p_i\| = \Delta$ (poznámka 181) a položíme $s = s_i + \alpha_i p_i$.

Dosavadní úvahy tvoří základ jednoduchého algoritmu:

Algoritmus 7 Data $C \succ 0$, $0 < \omega \leq \overline{\omega} < 1$, $\Delta > 0$, $m \geq n$ (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 $\tau \leq 0$, určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha p\| = \Delta$, položíme $s := s + \alpha p$ a ukončíme výpočet.

Krok 3 Položíme $\alpha := \rho/\tau$. Jestliže $\|s + \alpha p\| \geq \Delta$, určíme číslo $\alpha \geq 0$ 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 $\sigma \leq \omega^2 \overline{\sigma}$ nebo $k \geq m$, ukončíme výpočet.

Krok 5 Položíme $\beta := \sigma/\rho$, $p := r + \beta p$, $k := k + 1$ a přejdeme na krok 2.

Věta 93 *Směrový vektor získaný algoritmem 7 vyhovuje podmínkám (T1a)–(T1d) s $\underline{\delta} = 1$, $\overline{\delta} = 1$, $\overline{\omega} < 1$, $\underline{\sigma} = 1/(2\kappa(C))$ a $\overline{\sigma} = 1/\kappa(C)$.*

Důkaz Jak již bylo zmíněno, z věty 36 a věty 38 plyne, že $Q(s) < Q(s_i)$ a $g^T s < g^T s_i$. Pokud $i > 1$, můžeme použít důsledek 3 podle kterého platí

$$Q(s) \leq Q(s_2) \leq -\frac{\|g\|^2}{2\kappa(C)\|B\|}, \qquad g^T s \leq g^T s_2 \leq -\frac{\|g\|^2}{\kappa(C)\|B\|}.$$

Pokud $i = 1$, můžeme použít lemma 30 a lemma 269, takže

$$Q(s) = Q(s(\alpha^*)) \leq -\frac{1}{2}\|g\|\|s\|, \qquad g^T s = g^T s(\alpha^*) \leq -\|g\|\|s\|.$$

□

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 94 *Nechť jsou splněny předpoklady věty 36 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 94 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 8 Data $0 < \Delta$, $m \ll n$.

Krok 1 Jako v algoritmu 7.

Krok 2 Jako v algoritmu 7.

Krok 3 Jako v algoritmu 7.

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 30 *Nechť $B \in R^{n\times n}$ je symetrická matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy*

$$q_0 = 0, \qquad \gamma_1 q_1 = g$$

a

$$\delta_i = q_i^T B q_i, \quad \gamma_{i+1} q_{i+1} = B q_i - \delta_i q_i - \gamma_i q_{i-1}$$

pro $1 \leq i \leq n$, kde koeficienty $\gamma_i \geq 0$, $1 \leq i \leq n$, 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 185 Nechť $\gamma_i \neq 0$, $1 \leq i \leq k$, pro nějaký index $1 \leq k \leq n$. Pak podle definice 30 platí $g = Q_k(\gamma_1 e_1)$ a

$$BQ_k = Q_k T_k + \gamma_{k+1} q_{k+1} e_k^T \tag{282}$$

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

$$T_k = \begin{bmatrix} \delta_1, & \gamma_2, & \ldots, & 0, & 0 \\ \gamma_2, & \delta_2, & \ldots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \delta_{k-1}, & \gamma_k \\ 0, & 0, & \ldots, & \gamma_k, & \delta_k \end{bmatrix}$$

(matice $T_k \in R^{k\times k}$ je tridiagonální). Můžeme se o tom snadno přesvědčit roznásobením a použitím rekurentních vztahů metody LS.

Věta 95 *Uvažujme symetrický Lanczosův proces určený symetrickou maticí $B \in R^{n\times n}$ a vektorem $g \in R^n$. Nechť $\gamma_i \neq 0$, $1 \leq i \leq k$, 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 $q_1 = g/\parallel g \parallel$. Předpokládejme, že tvrzení platí pro nějaký index $1 \leq k < n$ a že $\gamma_{k+1} \neq 0$. Podle indukčního předpokladu platí $Q_k^T Q_k = I$, takže $Q_k^T B Q_k = T_k + \gamma_{k+1} Q_k^T q_{k+1} e_k^T$. 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 30 platí $\gamma_{k+1} q_k^T q_{k+1} = q_k^T B q_k - \delta_k = \delta_k - \delta_k = 0$. Vektor q_{k+1} je tedy ortogonální k vektorům q_i, $1 \leq i \leq k$, a má jednotkovou normu. Podle definice 30 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 186 Jelikož $Q_k^T Q_k = I$ a $Q_k^T q_{k+1} = 0$ (důkaz věty 95), můžeme psát

$$Q_k^T B Q_k = T_k,$$

takže symetrický Lanczosův proces lze použít k tridiagonalizaci matice B.

Poznámka 187 Symetrický Lanczosův proces můžeme použít k řešení soustavy rovnic $Bs + g = 0$. Pokládáme $s_1 = 0$ a vektory s_{i+1}, $1 \leq i \leq k$, hledáme tak, aby platilo

$$s_{i+1} = \arg\min_{s \in \mathcal{K}_i} \left(\frac{1}{2} s^T B s + g^T s \right).$$

Jelikož $s \in \mathcal{K}_i$ právě tedy, jestliže $s = Q_i z$, kde $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg\min_{z \in R^i} \left(\frac{1}{2} z^T T_i z + \gamma_1 e_1^T z \right)$$

(plyne to ze vztahů $g = Q_i(\gamma_1 e_1)$ a $Q_i^T Q_i = I$). Pokud $\gamma_{k+1} = 0$, je vektor $s_{k+1} \in \mathcal{K}_k$ řešením soustavy rovnic $Bs + g = 0$. Podle (282) totiž platí $BQ_k = Q_k T_k$ a jelikož matice T_k je regulární, lze položit $z_k = -T_k^{-1}(\gamma_1 e_1)$, což dává $Bs_{k+1} = BQ_k z_k = -Q_k T_k T_k^{-1}(\gamma_1 e_1) = -Q_k(\gamma_1 e_1) = -g$.

Věta 96 *Vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 187 jsou shodné s vektory s_{i+1}, $1 \leq i \leq k$, generovanými metodou sdružených gradientů (definice 22 s $C = I$). Navíc platí $\delta_1 = 1/\alpha_1$, $\varepsilon_1 = 1$ a*

$$\delta_{i+1} = \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}}, \qquad \gamma_{i+1} = \frac{\sqrt{\beta_i}}{|\alpha_i|}, \qquad \varepsilon_{i+1} = -\varepsilon_i \mathrm{sgn}(\alpha_i)$$

a

$$q_i = \varepsilon_i \frac{g_i}{\|g_i\|}$$

pro $1 \leq i \leq k$.

Důkaz Z důkazu věty 22 plyne, že vektory s_{i+1}, $1 \leq i \leq k$, určené metodou CG, leží v Krylovových podprostorech $\mathcal{K}_i$, $1 \leq i \leq k$, a realizují tam minimum kvadratické funkce $Q(s) = (1/2)s^T Bs + g^T s$. To je však právě definice vektorů s_{i+1}, $1 \leq i \leq k$, v poznámce 187. 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

$$G_k = Q_k D_k,$$

kde $G_k = [g_1, \ldots, g_k]$ a $D_k = \mathrm{diag}\ (\varepsilon_1 \parallel g_1 \parallel, \ldots, \varepsilon_k \parallel g_k \parallel)$ (čísla ε_i, $1 \leq i \leq k$, mohou nabývat hodnot ± 1). Položme $P_k = [p_1, \ldots, p_k]$. Pak z rekurentních vztahů metody CG 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 22 plyne, že matice $P_k^T B P_k$ je diagonální. Použijeme-li rekurentní vztahy metody CG, dostaneme

$$P_k^T B P_k = \mathrm{diag}(\parallel g_1 \parallel^2 / \alpha_1, \ldots, \parallel g_k \parallel^2 / \alpha_k) = D_k \mathrm{diag}(1/\alpha_1, \ldots, 1/\alpha_k) D_k,$$

takže

$$\begin{aligned} T_k &= Q_k^T B Q_k = D_k^{-1} G_k^T B G_k D_k^{-1} = D_k^{-1} B_k^T P_k^T B P_k B_k D_k^{-1} = \\ &= D_k^{-1} B_k^T D_k \mathrm{diag}\ (1/\alpha_1, \ldots, 1/\alpha_k) D_k B_k D_k^{-1}. \end{aligned}$$

Ale

$$D_k B_k D_k^{-1} = \begin{bmatrix} -1, & \beta_1 \frac{\varepsilon_1 \|g_1\|}{\varepsilon_2 \|g_2\|}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix} = \begin{bmatrix} -1, & \varepsilon_1 \varepsilon_2 \sqrt{\beta_1}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}.$$

Dosadíme-li tento vztah do vyjádření pro matici T_k, můžeme psát

$$T_k = \begin{bmatrix} \frac{1}{\alpha_1}, & \frac{-\varepsilon_1 \varepsilon_2 \sqrt{\beta_1}}{\alpha_1}, & \ldots, & 0 \\ \frac{-\varepsilon_1 \varepsilon_2 \sqrt{\beta_1}}{\alpha_1}, & \frac{\beta_1}{\alpha_1} + \frac{1}{\alpha_2}, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \frac{\beta_{k-1}}{\alpha_{k-1}} + \frac{1}{\alpha_k} \end{bmatrix},$$

což porovnáním se (282) dává $\delta_1 = 1/\alpha_1$ a

$$\begin{aligned} \delta_{i+1} &= \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}} \\ \gamma_{i+1} &= -\frac{\varepsilon_i \varepsilon_{i+1} \sqrt{\beta_i}}{\alpha_i} \end{aligned}$$

pro $1 \leq i \leq k$. Jelikož $\gamma_{i+1} \geq 0$, musí platit $\varepsilon_i \varepsilon_{i+1} = -\mathrm{sgn}(\alpha_i)$ pro $1 \leq i \leq k$. Protože podle definice 30 platí $\gamma_1 q_1 = g = g_1$ a $\gamma_1 \geq 0$, dostaneme $\varepsilon_1 = 1$. □

Poznámka 188 Symetrický Lanczosův proces můžeme použít k přibližnému určení optimálního lokálně omezeného kroku. Pokládáme $s_1 = 0$ a vektory s_{i+1}, $1 \leq i \leq k$, hledáme tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i, \|s\| \leq \Delta} \left(\frac{1}{2} s^T B s + g^T s \right). \tag{283}$$

Jelikož $s \in \mathcal{K}_i$ právě tedy, jestliže $s = Q_i z$, kde $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i, \|s\| \leq \Delta} \left(\frac{1}{2} z^T T_i z + \gamma_1 e_1^T z \right) \tag{284}$$

(plyne to z úvah použitých v poznámce 187 a z toho, že ortogonalita matice Q_i implikuje $\|s\| = \|Q_i z\| = \|z\|$).

Je-li vektor z_i řešením úlohy (284), zajímá nás, jak dobře aproximuje vektor s_{i+1} řešení úlohy (251). Podle věty 81 je vektor z_i řešením úlohy (284) právě tehdy, existuje-li číslo $\lambda_i \geq 0$ takové že matice $T_i + \lambda_i I$ je positivně semidefinitní, $(T_i + \lambda_i I) z_i + \gamma_1 e_1 = 0$, $\|z_i\| \leq \Delta$ 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 81. Kriteriem aproximace tedy může být hodnota $\|(B + \lambda_i) s_{i+1} + g\|$ (norma rezidua).

Věta 97 *Nechť $s_{i+1} = Q_i z_i$, kde vektor z_i je řešením úlohy (284). Pak platí*

$$(B + \lambda_i) s_{i+1} + g = \gamma_{i+1} e_i^T z_i q_{i+1},$$

takže $\|(B + \lambda_i) s_{i+1} + g\| = \gamma_{i+1} |e_i^T z_i|$.

Důkaz Použijeme-li vztah (282) a podmínku $(T_i + \lambda_i I) z_i + \gamma_1 e_1 = 0$, dostaneme

$$\begin{aligned} (B + \lambda_i I) s_{i+1} + g_i &= (B + \lambda_i I) Q_i z_i + \gamma_1 Q_i e_1 \\ &= Q_i ((T_i + \lambda_i I) z_i + \gamma_1 e_1) + \gamma_{i+1} q_{i+1} e_i^T z_i \\ &= \gamma_{i+1} q_{i+1} e_i^T z_i. \end{aligned}$$

Zbytek tvrzení plyne z toho, že $\|q_{i+1}\| = 1$. □

Nyní si podrobněji všimneme vlastností úlohy (284).

Definice 31 *řekneme, že matice T_i (jejíž tvar je uveden v poznámce 185) je ireducibilní, jestliže $\gamma_j \neq 0$ $\forall 1 < j \leq i$.*

Věta 98 *Je-li matice T_i ireducibilní, nenastane v úloze (284) singulární případ (matice $T_i + \lambda_i I$ je positivně definitní). Je-li matice T_n ireducibilní, nenastane singulární případ ani v úloze (251). Nenastane-li singulární případ v úloze (251) a platí-li $\gamma_{i+1} = 0$, je vektor $s_{i+1} = Q_i z_i$ řešením úlohy (251).*

Důkaz (a) Je-li vektor z_i řešením úlohy (284), 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

$$z_i^T (T_i + \lambda_i I)v_i = -\gamma_1 e_1^T v_i = 0,$$

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 $\gamma_2 \neq 0$, je i druhá složka vektoru v_i nulová (matice T_i je tridiagonální). Takto lze pokračovat dále a jsou-li všechna čísla γ_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 (284) tedy nenastane singulární případ.

(b) Pro $i = n$ je úloha (251) ekvivalentní úloze (283) a tedy i úloze (284). Je-li matice T_n ireducibilní, nenastane singulární případ v úloze (284) a tedy ani v úloze (251).

(c) Platí-li $\gamma_{i+1} = 0$, můžeme podle (282) 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 slouci 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 (251) nenastane singulární případ takže podle poznámky 178 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 97, vidíme, že jsou splněny nutné a postačující podmínky pro to, aby vektor $s_{i+1} = Q_i z_i$ byl řešením úlohy (251). □

Poznámka 189 Symetrický Lanczosův proces můžeme předpodmínit tak že ho aplikujeme na kvadratickou funkci (82). 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

$$q_0 = 0, \quad \gamma_1 q_1 = g, \quad v_1 = C^{-1} q_1$$

a

$$\delta_i = v_i^T B v_i, \quad \gamma_{i+1} q_{i+1} = B v_i - \delta_i q_i - \gamma_i q_{i-1}, \quad v_{i+1} = C^{-1} q_{i+1}$$

pro $1 \leq i \leq n$, kde koeficienty $\gamma_i \geq 0$, $1 \leq i \leq n$ 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

$$B V_k = Q_k T_k + \gamma_{k+1} q_{k+1} e_k^T,$$

kde $T_k = V_k^T B V_k$ je symetrická tridiagonální matice. Poznamenejme, že je-li vektor z_i řešením problému (284), kde matice T_i byla získána předpodmíněným symetrickým Lanczosovým procesem, je třeba v (284) nahradit podmínku $\|s\| \leq \Delta$ podmínkou $s^T C s \leq \Delta^2$.

Nyní můžeme přistoupit k popisu algoritmu pro výpočet lokálně omezeného kroku pomocí symetrického Lanczosova procesu.

Poznámka 190 Shrneme základní myšlenky, které se používají v algoritmu založeném na použití symetrického Lanczosova procesu.

(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 γ_i, δ_i a vektory q_i (věta 96).

(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 (284).

(c) Výpočet ukončíme, platí-li $\|g_{i+1}\| \leq \omega \|g\|$ v případě, že $s_{i+1} = s_i + \alpha_i p_i$, nebo $\gamma_{i+1}|e_i^T z_i| \leq \omega\|g\|$ v případě, že $s_{i+1} = Q_i z_i$. Přitom ω je předepsaná přesnost.

Dosavadní úvahy můžeme shrnout ve formě algoritmu. 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 (284).

Algoritmus 9 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 $u_k := Bp_k$ 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 δ_k podle věty 96, tedy $\delta_k := 1/\alpha_k$, pokud $k = 1$, nebo $\delta_k := 1/\alpha_k + \beta_{k-1}/\alpha_{k-1}$, pokud $k > 1$. Je-li $\alpha_k \leq 0$ nebo $\|s_k + \alpha_k p_k\| \geq 0$, 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 γ_{k+1} podle věty 96, tedy $\gamma_{k+1} := \sqrt{\beta_k}/|\alpha_k|$ a přejdeme na krok 6.

Krok 5 Položíme $M := 0$, $\delta_k := q_k^T B q_k$ a vypočteme číslo γ_{k+1} a vektor $v_{k+1} := \gamma_{k+1} q_{k+1}$ podle definice 30, tedy $\gamma_{k+1} := \|v_{k+1}\|$, kde $v_{k+1} := Bq_k - \delta_k q_k$, pokud $k = 1$, nebo $v_{k+1} := Bq_k - \delta_k q_k - \gamma_k q_{k-1}$, 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 (284) a pokud $\gamma_{k+1}|e_k^T z_k| \leq \omega\|g\|$ nebo $k > m$, ukončíme výpočet.

Krok 7 Jestliže $L = 0$, položíme $q_{k+1} := v_{k+1}/\gamma_{k+1}$. 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 81 a směrový vektor $s = s(\lambda)$ budeme hledat řešením úlohy

$$s(\lambda) = \arg \min_{\|s\| \leq \Delta} Q_\lambda(s), \qquad Q_\lambda(s) = \frac{1}{2} s^T (B + \lambda I) s + g^T s. \tag{285}$$

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 (251) má řešení takové, že $\|s_i^*\| < \Delta$. To je zaručeno, pokud je splněna nerovnost $\lambda \leq \lambda^*$, kterou nyní dokážeme. Budeme přitom používat označení

$$\mathcal{K}_k(\lambda) = \text{span}\{g, (B + \lambda I)g, \ldots, (B + \lambda I)^{k-1}g\}$$

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 $\mathcal{K}_k = \mathcal{K}_k(0)$ (pokud $\lambda = 0$ budeme argument λ vynechávat).

Věta 99 *Nechť pro daný index $1 \leq k \leq n$, je vektor s_k řešením úlohy*

$$s_k = \arg \min_{s \in \mathcal{K}_k, \|s\| \leq \Delta} Q(s), \qquad Q(s) = \frac{1}{2} s^T B s + g^T s \tag{286}$$

s odpovídajícím Lagrangeovým multiplikátorem λ_k. Jestliže $1 \leq i \leq j \leq n$, pak $\lambda_i \leq \lambda_j$. Speciálně $\lambda_k \leq \lambda^*$ pro libovolný index $1 \leq k \leq n$.

Důkaz (a) Nechť vektor s_k je řešením nepodmíněné úlohy

$$s_k = \arg \min_{s \in \mathcal{K}_k} Q(s).$$

Jestliže $1 \leq i \leq j \leq n$, pak podle věty 36 platí $\|s_i\| \leq \|s_j\|$. Speciálně $\|s_k\| \leq \|s_n\| = \|s^*\|$, kde $\|s^*\|$ je nepodmíněným minimem funkce $Q(s)$ na R^n.

(b) Indukcí dokážeme, že pro libovolné číslo $\lambda \in R$ platí $\mathcal{K}_k(\lambda) = \mathcal{K}_k$. Pro $k = 1$ je to zřejmé, neboť $\mathcal{K}_k(\lambda) = \text{span}\{g\} = \mathcal{K}_k$. Předpokládejme, že tvrzení 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.

(c) 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

$$\begin{aligned} B_1 - B_2 \succeq 0 \quad &\Longleftrightarrow \quad B_2^{-1} - B_1^{-1} \succeq 0, \\ B_1 - B_2 \succ 0 \quad &\Longleftrightarrow \quad B_2^{-1} - B_1^{-1} \succ 0. \end{aligned}$$

(d) Ukážeme, že vektor $s_k(\lambda)$, který minimalizuje $Q_\lambda(s)$ na $\mathcal{K}_k$ lze vyjádřit ve tvaru

$$s_k(\lambda) = -Z_k (Z_k^T (B + \lambda I) Z_k)^{-1} Z_k^T g,$$

kde $Z_k \in R^{n \times k}$ je matice, jejíž sloupce tvoří ortonormální bázi v $\mathcal{K}_k$. Jestliže $s \in \mathcal{K}_k$, můžeme psát $s = Z_k \tilde{s}$, kde $\tilde{s} \in R^k$. Pak

$$Q_\lambda(s) = \frac{1}{2} s^T (B + \lambda I) s + g^T s = \frac{1}{2} \tilde{s}^T Z_k^T (B + \lambda I) Z_k \tilde{s} + g^T Z_k \tilde{s} \stackrel{\Delta}{=} \tilde{Q}_\lambda(\tilde{s})$$

a minimum $\tilde{s}_k(\lambda)$ funkce $\tilde{Q}_\lambda(\tilde{s})$ na R_k lze vyjádřit ve tvaru $\tilde{s}_k(\lambda) = -(Z_k^T (B + \lambda I) Z_k)^{-1} Z_k^T g$, což po dosazení do $s_k = Z_k \tilde{s}_k$ dává hledaný výsledek.

(e) 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 (d). Ukážeme, že

$$\lambda_2 \leq \lambda_1 \quad \Longleftrightarrow \quad \|s_k(\lambda_2)\| \geq \|s_k(\lambda_1)\|.$$

Použijeme-li (d), dostaneme

$$\|s_k(\lambda)\|^2 = g^T Z_k (Z_k^T (B + \lambda I) Z_k)^{-2} Z_k^T g = g^T Z_k (Z_k^T B Z_k + \lambda I)^{-2} Z_k^T g.$$

Platí tedy

$$\|s_k(\lambda_2)\|^2 - \|s_k(\lambda_1)\|^2 = g^T Z_k \left[(Z_k^T B Z_k + \lambda_2 I)^{-2} - (Z_k^T B Z_k + \lambda_1 I)^{-2} \right] Z_k^T g.$$

Označíme-li $\tilde{B}_2 = (Z_k^T B Z_k + \lambda_2 I)$ a předpokláme-li, že $\lambda_2 \leq \lambda_1$, můžeme psát

$$(Z_k^T B Z_k + \lambda_1 I)^2 - (Z_k^T B Z_k + \lambda_2 I)^2 = (\tilde{B}_2 + (\lambda_1 - \lambda_2) I)^2 - \tilde{B}_2^2 = 2(\lambda_1 - \lambda_2)\tilde{B}_2 + (\lambda_1 - \lambda_2)^2 I \succeq 0,$$

což spolu s první ekvivalencí v (c) dává

$$(Z_k^T B Z_k + \lambda_2 I)^{-2} - (Z_k^T B Z_k + \lambda_1 I)^{-2} \succeq 0,$$

neboli $\|s_k(\lambda_2)\|^2 - \|s_k(\lambda_1)\|^2 \geq 0$. Použijeme-li druhou ekvivalenci v (c), dostaneme stejným postupem $\lambda_2 < \lambda_1 \Rightarrow \|s_k(\lambda_2)\|^2 > \|s_k(\lambda_1)\|^2$. Protože nezáleží na pořadí, můžeme psát $\lambda_1 < \lambda_2 \Rightarrow \|s_k(\lambda_1)\|^2 > \|s_k(\lambda_2)\|^2$, což dává $\|s_k(\lambda_2)\| \geq \|d_k(\lambda_1)\| \Rightarrow \lambda_2 \leq \lambda_1$.

(f) Nyní již můžeme přistoupit k důkazu samotné věty. Vektor s_k je řešením úlohy (286) právě tehdy, jestliže $\|s_k\| = \|Z_k \tilde{s}_k\| \leq \Delta$, kde $Z_k^T (B + \lambda_k I) Z_k \tilde{s}_k = -Z_k^T g$, $Z_k^T (B + \lambda_k I) Z_k \succeq 0$, $\lambda_k \geq 0$ a $\lambda_k(\Delta - \|s_k\|) = 0$ (věta 81). Toto řešení je nepodmíněným minimem (stejné řešení dostaneme i po odstranění omezení $s_k \leq \Delta$) právě tehdy, jestliže $\lambda_k = 0$. Jestliže $\lambda_j = 0$ (což znamená, že $\|s_j\|$ je nepodmíněným minimem) a $i \leq j$, pak podle (a) platí $\|s_i\| \leq \|s_j\| \leq \Delta$ pro nepodmíněné minimum $\|s_i\|$, takže $\lambda_i = 0$. Jestliže $\lambda_j > 0$ a $\lambda_i = 0$, není co dokazovat. Nechť $\lambda_j > 0$ a $\lambda_i > 0$, což znamená, že $\|s_j\| = \|s_i\| = \Delta$. Předpokládejme nejprve, že matice $Z_i^T (B + \lambda_i I) Z_i$ je singulární a $\lambda_j < \lambda_i$. Pak existuje vektor $v \in \mathcal{K}_i$ takový, že $v^T (B + \lambda_j I) v < 0$ a protože $\mathcal{K}_i \subset \mathcal{K}_j$, vztah $Z_j^T (B + \lambda_j I) Z_j \succeq 0$ nemůže platit. Tento spor dokazuje, že $\lambda_j \geq \lambda_i$. Předpokládejme nyní, 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 (b) platí $\mathcal{K}_i(\lambda_i) = \mathcal{K}_i$, je vektor s_i řešením nepodmíněné úlohy

$$s_i = \arg \min_{s \in \mathcal{K}_i} Q_{\lambda_i}(s).$$

Předpokládejme, že $\lambda_i > \lambda_j$, což implikuje, že $Z_j^T (B + \lambda_j I) Z_j \succ 0$. Nechť

$$s_j(\lambda_i) = \arg \min_{s \in \mathcal{K}_j} Q_{\lambda_i}(s).$$

Pak z (a) plyne, že $\|s_j(\lambda_i)\| \geq \|s_k\| = \Delta$. Protože

$$s_j = \arg \min_{s \in \mathcal{K}_j} Q_{\lambda_j}(s)$$

a $\|s_j\| = \Delta \leq \|d_j(\lambda_i)\|$, z (e) plyne, že $\lambda_i \leq \lambda_j$, což je spor. Musí tedy platit $\lambda_i \leq \lambda_j$. Předpokládejme nakonec, že matice $Z_j^T (B + \lambda_j I) Z_j$ je singulární. V tomto případě platí $\|d_j(\lambda_j + \varepsilon)\| \leq \Delta$ 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 z (a) plyne, že $\|s_i(\lambda_j + \varepsilon)\| \leq \|s_j(\lambda_j + \varepsilon)\| \leq \Delta$. Protože $\|s_i\| = \Delta$, z (e) plyne, že $\lambda_i \leq \lambda_j + \varepsilon$ a jelikož číslo ε je libovolné, platí $\lambda_i \leq \lambda_j$. □

Nyní se vrátíme k problému (285). Položíme-li $\lambda = \lambda_k$ pro nějaký index $k \leq n$, věta 99 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 36, 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 10 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

$$z_k = \arg \min_{z \in R^k, \|s\| \leq \Delta} \left(\frac{1}{2} z^T T_k z + \gamma_1 e_1^T z \right)$$

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 (285) 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.

Následující tabulka ukazuje srovnání efektivity několika metod pro výpočet lokálně omezeného kroku (A5 - metoda s optimálním lokálně omezeným krokem (algoritmus 5), A6 - metoda psí nohy (algoritmus 6), A7 - nepřesná metoda s lokálně omezeným krokem (algoritmus 7 s $C = I$), PA7 - předpodmíněná nepřesná metoda s lokálně omezeným krokem (algoritmus 7 s $C \neq I$), A8 - vícekroková metoda psí nohy (algoritmus 8 s $m = 5$), A9 - metoda založená na použití symetrické Lanczosovy metody (algoritmus 9), PSA7 - předpodmíněná posunutá nepřesná metoda s lokálně omezeným krokem popsaná v tomto oddílu) při řešení 22 testovacích problémů s 1000 a 5000 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV, gradientů NFG, vnitřních iterací NCG a celkový čas výpočtu). Výsledky uvedené v této tabulce byly získány diferenční verzí Newtonovy metody popsané v oddílu 8.3 (realizované jako metody s lokálně omezeným krokem určovaným pomocí uvedených algoritmů).

N	Metoda	NIT	NFV	NFG	NCG	Čas
1000	Alg.5	1918	1955	8797	-	4.65
	Alg.6	2515	2716	11859	-	4.42
	Alg.8	2292	2456	10673	12203	4.61
	Alg.7	3329	3784	16456	53573	8.20
	Alg.9	3107	3444	15306	55632	8.53
	Alg.7P	2631	2823	13019	910	5.14
	Alg.10P	1999	2046	9201	1161	4.25
5000	Alg.5	8391	8566	35824	-	122.44
	Alg.6	9657	10133	42425	-	115.77
	Alg.8	8938	9276	39032	47236	122.84
	Alg.7	16894	19163	83933	358111	364:42
	Alg.9	14679	16383	71483	366695	401.45
	Alg.7P	10600	11271	50365	3767	145.42
	Alg.10P	8347	8454	35939	4329	108.87

6.6 Maticové rozklady pro symetrické indefinitní matice

Definice 32 *Gillův-Murrayův rozklad matice B má tvar*

$$R^T R = B + E,$$

kde R je regulární horní trojúhelníková matice a E je pozitivně semidefinitní diagonální matice (může být $E = 0$).

Gillův-Murrayův rozklad se provádí tak, že na začátku i-tého eliminačního kroku máme matici

$$\begin{bmatrix} R_{(i-1),(i-1)}, & R_{(i-1),i}, & R_{(i-1),(n-i)} \\ *, & B_{ii}^{(i-1)}, & B_{i,(n-i)}^{(i-1)} \\ *, & *, & B_{(n-i),(n-i)}^{(i-1)} \end{bmatrix},$$

kde horní index v závorce značí počet již provedených eliminačních kroků a dolní indexy v závorkách značí submatice s $(i-1)$ řádky nebo $(n-i)$ sloupci. Eliminační krok vypadá takto:

$$\gamma_i = \max_{i<j\leq n}(|B_{i,j}^{(i-1)}|),$$

$$\rho_i^2 = \max\left(|B_{ii}^{(i-1)}|, \frac{\gamma_i^2}{\beta^2}, \delta^2\right),$$

$$R_{ii} = \rho_i,$$

$$R_{i,(n-i)} = B_{i,(n-i)}^{(i-1)}/R_{ii},$$

$$B_{(n-i),(n-i)}^{(i)} = B_{(n-i),(n-i)}^{(i-1)} - R_{i,(n-i)}^T R_{i,(n-i)},$$

kde δ je malé číslo a $\beta > \sqrt{\|B\|}$). Tento proces se od Choleského rozkladu liší pouze tím že může platit $\rho_i^2 \neq B_{ii}^{(i-1)}$. Bližším rozborem uvedených vztahů se dá dokázat že pro prvky matice E platí

$$E_{ii} = \rho_i^2 - B_{ii}^{(i-1)} = \rho_i^2 + R_{i,(n-i)}R_{i,(n-i)}^T - B_{ii},$$

kde B_{ii} je prvek původní matice.

Věta 100 *Nechť $R^TR = B + E$ je Gillův-Murrayův rozklad s $\delta = 0$ a $\beta > \sqrt{\|B\|}$. Nechť*

$$B_{kk}^{(k-1)} = \min_{1\leq i\leq n} B_{ii}^{(i-1)}$$

a nechť $v \in R^n$ je vektor určený řešením rovnice $Rv = e_k$ (e_k ke k-tý sloupec jednotkové matice). Není-li matice B pozitivně semidefinitní, platí

$$v^TBv = \frac{B_{kk}^{(k-1)}}{\rho_k^2} < 0.$$

Důkaz Z rovnice $Rv = e_k$ plyne, že $v_k = 1/\rho_k$. Platí tedy

$$\begin{aligned} v^TBv &= v^T(B+E)v - v^TEv \leq v^TR^TRv - v_k^2E_{kk} = \\ &= e_k^Te_k - E_{kk}/\rho_k^2 = \frac{\rho_k^2 - E_{kk}}{\rho_k^2} = \frac{B_{kk}^{(k-1)}}{\rho_k^2}. \end{aligned}$$

Není-li matice B pozitivně semidefinitní, musí existovat index $1 \leq i \leq n$ tak, že $E_{ii} \neq 0$, neboli $\rho_i^2 \neq B_{ii}^{(i-1)}$. Mohou nastat dva případy. Buď $\rho_i^2 = |B_{ii}^{(i-1)}| \neq B_{ii}^{(i-1)}$, takže $B_{ii}^{(i-1)} < 0$ a tedy i $B_{kk}^{(k-1)} < 0$, nebo $\rho_i^2 = \gamma_i^2/\beta^2$. Ve druhém případě musí existovat index $i < j \leq n$ tak, že $\gamma_i = |B_{ij}^{(i-1)}|$, takže

$$|R_{ij}| = \frac{|B_{ij}^{(i-1)}|}{\rho_i} = \frac{\gamma_i}{\gamma_i/\beta} = \beta,$$

což dává

$$B_{ii}^{(i-1)} = \rho_i^2 - E_{ii} = B_{ii} - R_{i,(n-i)} R_{i,(n-1)}^T \leq B_{ii} - \beta^2 < \|B\| - \|B\| = 0.$$

□

Definice 33 *Bunchův-Parlettův rozklad matice B má tvar*

$$LDL^T = PBP^T,$$

kde

$$L = \begin{bmatrix} I, & 0, & \dots, & 0 \\ L_{21}, & I, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ L_{n1}, & L_{n2}, & \dots, & I \end{bmatrix}, \quad D = \begin{bmatrix} D_{11}, & 0, & \dots, & 0 \\ 0, & D_{22}, & \dots, & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0, & 0, & \dots, & D_{nn} \end{bmatrix}.$$

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).

Bunchův-Parlettův rozklad se provádí tak, že na začátku i-tého eliminačního kroku máme matici

$$\begin{bmatrix} D_{11}, & L_{12}, & \dots, & L_{1,i-1}, & L_{1,(m-i+1)} \\ *, & D_{22}, & \dots, & L_{2,i-1}, & L_{2,(m-i+1)} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ *, & *, & \dots, & D_{i-1,i-1}, & L_{i-1,(m-i+1)} \\ *, & *, & \dots, & *, & B^{(i-1)} \end{bmatrix}.$$

Eliminační krok má tvar:

$$\beta_i = \max_k |B_{kk}^{(i-1)}|,$$

$$\gamma_i = \max_{k,l} |B_{kl}^{(i-1)}|,$$

$$\alpha_i = \beta_i / \gamma_i.$$

Jestliže $\alpha_i \geq \left(\sqrt{17}+1\right)/8$ volíme v i-tém kroku blok 1×1, jinak volíme blok 2×2. Je třeba provádět permutace (pivotový blok s indexy k a l se přenese do levého horního rohu matice $B^{(i-1)}$).Pak se provede transformace

$$B^{(i-1)} \to \begin{bmatrix} D_{ii}, & L_{i,(m-i)} \\ *, & B^{(i)} \end{bmatrix},$$

kde

$$D_{ii} = B_{ii}^{(i-1)},$$

$$L_{i,(m-i)} = D_{ii}^{-1} B_{i,(m-i)}^{(i-1)},$$

$$B^{(i)} = B_{(m-i),(m-i)}^{(i-1)} - L_{i,(m-i)}^T B_{i,(m-i)}^{(i-1)}.$$

Věta 101 *Nechť $LDL^T = PBP^T$ je Bunchův-Parlettův rozklad. Nechť $u_i = 0$, pokud $\underline{\lambda}(D_{ii}) \geq 0$, a nechť u_i je normalizovaný vlastní vektor příslušný $\underline{\lambda}(D_{ii})$, pokud $\underline{\lambda}(D_{ii}) < 0$. Nechť $L^T P v = u$, kde $u^T = [u_1, \dots, u_m]$. Není-li matice B pozitivně semidefinitní, platí*

$$v^T B v = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) < 0.$$

Důkaz Z rovnice $L^T Pv = u$ dostaneme

$$v^T Bv = v^T P^T LDL^T Pv = u^T Du = \sum_{i=1}^{m} u_i^T D_{ii} u_i = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}).$$

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

$$v^T Bv = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) \leq \underline{\lambda}(D_{kk}) < 0.$$

□

7 Metody pro minimalizaci součtu čtverců

V tomto oddílu budeme předpokládat, že minimalizovaná funkce má tvar

$$F(x) = \frac{1}{2} f^T(x) f(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x),$$

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

$$f(x) = \begin{bmatrix} f_1(x) \\ \vdots \\ f_m(x) \end{bmatrix}, \qquad J(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \dots, & \frac{\partial f_1(x)}{\partial x_n} \\ \vdots & \vdots & \vdots \\ \frac{\partial f_m(x)}{\partial x_1}, & \dots, & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix}.$$

pro zobrazení f a jeho Jacobiovu matici. Je li zobrazení $f : \mathcal{D}_F \to R^m$ spojitě diferencovatelné na nějaké otevřené množině $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$, pak platí

$$g(x) = J^T(x) f(x) = \sum_{k=1}^{m} f_k(x) g_k(x) \tag{287}$$

$\forall x \in \mathcal{D}$. Je-li zobrazení f dvakrát spojitě diferencovatelné na $\mathcal{D}$, pak

$$G(x) = J^T(x) J(x) + C(x) = \sum_{k=1}^{m} g_k(x) g_k^T(x) + \sum_{k=1}^{m} f_k(x) G_k(x) \tag{288}$$

$\forall x \in \mathcal{D}$. Při vyšetřování konvergence metod pro minimalizaci součtu čtverců budeme často používat předpoklad (F1) (který je splněn automaticky, neboť $F(x) \geq 0 \ \forall x \in \mathcal{D}_F$), předpoklad (F2) a tyto předpoklady.

Předpoklad 10 *Zobrazení $f : \mathcal{D}_F \to R^m$ je omezené na $\mathcal{D}$, takže existuje konstanta $\overline{f}$ taková, že*

$$\|f(x)\| \leq \overline{f} \quad \forall x \in \mathcal{D}. \tag{J1}$$

Předpoklad 11 *Množina $\mathcal{D}$ splňuje podmínku*

$$\mathcal{D} \subset \mathcal{D}_F(\overline{F}) + B(0, \overline{D}) = \bigcup_{y \in \mathcal{D}_F(\overline{F})} B(y, \overline{D}), \tag{J2}$$

takže ke každému bodu $x \in \mathcal{D}$ existuje bod $y \in \mathcal{D}_F(\overline{F})$ takový, že $\|x - y\| \leq \overline{D}$.

Poznámka 191 Jsou-li splněny podmínky (F2) a (J2), je množina $\overline{\mathcal{D}}$ (uzávěr) kompaktní.

Předpoklad 12 *Zobrazení $f : \mathcal{D}_F \to R^m$ je lipschitzovské na $\mathcal{D}$, takže existuje konstanta $\overline{J} > 0$ taková, že*

$$\|f(x_2) - f(x_1)\| \leq \overline{J} \|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{J3}$$

Předpoklad 13 *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*

$$\|J(x) d\| \leq \overline{J} \|d\| \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{J4}$$

Podmínka (J4) je ekvivalentní podmínce $\|J(x)\| \leq \overline{J} \ \forall x \in \mathcal{D}$.

Předpoklad 14 *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*

$$\|J(x)d\| \geq \underline{J}\|d\| \quad \forall x \in \mathcal{D} \quad \forall d \in R^n. \tag{J5}$$

Podmínka (J5) je ekvivalentní podmínce $\|J^{-1}(x)\|^{-1} \geq \underline{J}\ \forall x \in \mathcal{D}$.

Předpoklad 15 *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*

$$\|J(x_2) - J(x_1)\| \leq \overline{G}\|x_2 - x_1\| \quad \forall x_1, x_2 \in \mathcal{D}. \tag{J6}$$

Předpoklad 16 *Zobrazení $f \in \mathcal{C}^2 : \mathcal{D} \to R^n$ má omezené druhé derivace na $\mathcal{D}$, takže existuje konstanta $\overline{G} > 0$ taková, že*

$$\|G_k(x)\| \leq \overline{G} \quad \forall 1 \leq k \leq m, \quad \forall x \in \mathcal{D}. \tag{J7}$$

Abychom ukázali význam uvedených předpokladů uvedeme analogii tvrzení 3 o střední hodnotě.

Tvrzení 5 *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 (J3) nebo tvrzení 5 a (J4), dostaneme

$$\|f(x+d) - f(x)\| \leq \overline{J}\|d\|, \tag{289}$$

$$d^T(f(x+fd) - g(x)) \leq \overline{J}\|d\|^2. \tag{290}$$

Použijeme-li tvrzení 5 a (J5), dostaneme

$$\|f(x+d) - f(x)\| \geq \underline{J}\|d\|, \tag{291}$$

$$d^T(f(x+d) - f(x)) \geq \underline{J}\|d\|^2. \tag{292}$$

Důkaz posledních dvou nerovností:

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

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

Poznámka 192 Použijeme-li tvrzení 5 a (J4), můžeme psát

$$\|f(x_2) - f(x_1)\| = \left\| \int_0^1 J(x+\lambda(x_2 - x_1))(x_2 - x_1)d\mathrm{d}\lambda \right\| \leq \overline{J}\|x_2 - x_1\|,$$

takže (J4) implikuje (J3).

Poznámka 193 Použijeme-li tvrzení 3 a (J7), můžeme pro $1 \le k \le m$ psát

$$\|g_k(x_2) - g_k(x_1)\| = \left\| \int_0^1 G(x + \lambda(x_2 - x_1))(x_2 - x_1) d\mathrm{d}\lambda \right\| \le \overline{G} \|x_2 - x_1\|,$$

neboli

$$\|J(x_2) - J(x_1)\| \le \|J(x_2) - J(x_1)\|_F = \sqrt{\sum_{k=1}^{m} \|g_k(x_2) - g_k(x_1)\|^2} \le \sqrt{m}\, \overline{G}\, \|x_2 - x_1\|,$$

takže (J7) implikuje (J6) (s konstantou $\sqrt{m}\, \overline{G}$ místo $\overline{G}$).

Poznámka 194 Je-li splněn předpoklad (J2) pak (J3) implikuje (J1), neboť ke každému bodu $x \in \mathcal{D}$ existuje bod $y \in \mathcal{D}_F(\overline{F})$ tak, že

$$\|f(x)\| \le \|f(y)\| + \|f(x) - f(y)\| \le \sqrt{2\overline{F}} + \overline{J}\|x - y\| \le \sqrt{2\overline{F}} + \overline{J}\, \overline{D} \triangleq \overline{f}.$$

Ukážeme, že předpoklady (J1), (J4), (J6) nebo (J1), (J4), (J7), kladené na zobnrazení f, implikují předpoklady (F3) nebo (F4), kladené na funkci F.

Věta 102 *Nechť jsou splněny predpoklady (J1), (J4), (J6). Pak je-li množina $\mathcal{D}$ konvexní, splňuje funkce $F = (1/2) f^T f$ předpoklad (F3). Nechť jsou splněny predpoklady (J1), (J4), (J7). Pak funkce F splňuje předpoklad (F4).*

Důkaz (a) Nechť jsou splněny predpoklady (J1), (J4), (J6) a nechť $x_1 \in \mathcal{D}$, $x_2 \in \mathcal{D}$. Pak podle (287) platí

$$\begin{aligned}
\|g(x_2) - g(x_1)\| &= \|J^T(x_2) f(x_2) - J^T(x_1) f(x_1)\| \\
&\le \|J^T(x_2)(f(x_2) - f(x_1))\| + \|(J^T(x_2) - J^T(x_1)) f(x_1)\| \\
&\le \overline{J}\|(f(x_2) - f(x_1)\| + \overline{G}\|x_2 - x_1\|\|f(x_1)\| \\
&= \overline{J} \left\| \int_0^1 J(x_1 + \tau(x_2 - x_1))(x_2 - x_1) \mathrm{d}\tau \right\| + \overline{G}\|x_2 - x_1\|\|f(x_1)\| \\
&\le \left(\overline{J}^2 + \overline{G}\, \overline{f} \right) \|x_2 - x_1\|. \qquad (293)
\end{aligned}$$

(b) Nechť jsou splněny predpoklady (J1), (J4), (J7) a nechť $x \in \mathcal{D}$. Pak podle (288) platí

$$\|G(x)\| \le \|J(x)^T J(x)\| + \sum_{k=1}^{m} |f_k(x)| \|G_k(x)\| \le \overline{J}^2 + m \overline{f}\, \overline{G}.$$

□

Poznámka 195 Při vyšetřování metod pro minimalizaci součtu čtverců budeme používat i předpoklady (F3) nebo (F4). Musíme však mít na paměti, že to znamená předpokládat (J1), (J4), (J6) nebo (J1), (J4), (J7). Poznamenejme, že je-li množina $\mathcal{D}$ omezená a je-li zobrazení f spojitě diferencovatelné na $\overline{\mathcal{D}}$ (uzávěr), platí (F3). Je-li navíc f dvakrát spojitě diferencovatelné na $\overline{\mathcal{D}}$, platí (F4).

Poznámka 196 Podmínky (J3)–(J7) jsou často zbytečně silné. Studujeme-li chování iteračního procesu v okolí limitního bodu $x^* \in R^n$, stačí předpokládat, že sobrazení $f : \mathcal{D}_F \to R^m$ je spojitě diferencovatelné v nějakém okolí bodu x^* a platí ($\overline{J3}$)–($\overline{J7}$) (místo $\mathcal{D}$ používáme $\mathcal{B}(x^*, \varepsilon)$).

Předpoklad 17 *Zobrazení $f : \mathcal{D}_F \to R^m$ je spojité v okolí bodu $x^* \in \mathcal{D}_F$ a existuje konstanta $\overline{J}$ a číslo $\varepsilon > 0$ tak, že platí*

$$\|f(x) - f(x^*)\| \leq \overline{J}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{$\overline{\text{J3}}$}$$

Předpoklad 18 *Zobrazení $f : \mathcal{D}_F \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}_F$. Pak pro libovolnou konstantu $\overline{J} > \|J(x^*)\|$ existuje číslo $\varepsilon > 0$ takové, že*

$$\|J(x)d\| \leq \overline{J}\|d\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{$\overline{\text{J4}}$}$$

Předpoklad 19 *Zobrazení $f : \mathcal{D}_F \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}_F$ a Jacobiova matice $J(x^*)$ je regulární. Pak pro libovolnou konstantu $0 < \underline{J} < \|J(x^*)\|$ existuje číslo $\varepsilon > 0$ takové, že*

$$\|J(x)d\| \geq \underline{J}\|d\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{$\overline{\text{J5}}$}$$

Předpoklad 20 *Zobrazení $f : \mathcal{D}_F \to R^m$ je spojitě diferencovatelné v okolí bodu $x^* \in \mathcal{D}_F$ a existuje konstanta $\overline{G}$ a číslo $\varepsilon > 0$ tak, že platí*

$$\|J(x) - J(x^*)\| \leq \overline{G}\|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{$\overline{\text{J6}}$}$$

Předpoklad 21 *Zobrazení $f : \mathcal{D}_\mathcal{F} \to R^m$ je dvakrát spojitě diferencovatelné v okolí bodu $x^* \in R^n$. Pak pro libovolnou konstantu $\overline{G} > \max_{1 \leq k \leq m}(\|G_k(x^*)\|)$ existuje číslo $\varepsilon > 0$ takové, že*

$$\|G_k(x)\| \leq \overline{G} \quad \forall 1 \leq k \leq m, \quad \forall x \in \mathcal{B}(x^*, \varepsilon). \tag{$\overline{\text{J7}}$}$$

Věta 103 *Jsou-li splněny předpoklady ($\overline{J4}$) a ($\overline{J7}$), je splněna podmínka ($\overline{F4}$).*

Důkaz Ze spojité diferencovatelnosti zobrazení f 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 byly splněny předpoklady ($\overline{\text{J4}}$) a ($\overline{\text{J7}}$). 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 104 *Jsou-li splněny předpoklady ($\overline{J5}$), ($\overline{J7}$) a platí-li $f(x^*) = 0$, je splněna podmínka ($\overline{F5}$).*

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 předpoklady ($\overline{\text{J5}}$) a ($\overline{\text{J7}}$). Pak v $\mathcal{B}(x^*, \varepsilon)$ platí

$$\begin{aligned} d^T G(x)d &= d^T \left(J^T(x)J(x) + \sum_{k=1}^{n} f_k(x)G_k(x) \right) d \geq \|J(x)d\|^2 - \sum_{k=1}^{n} |f_k(x)|\|G_k(x)\|\|d\|^2 \\ &\geq \left(\underline{J}^2 - \frac{\underline{J}^2}{2m\overline{G}} m\overline{G} \right) \|d\|^2 = \frac{1}{2}\underline{J}^2\|d\|^2. \end{aligned}$$

□

Poznámka 197 Jestliže $x_i \to x^*$, existuje k danému číslu $\varepsilon > 0$ index $k \in N$ takový, že $x_i \in \mathcal{B}(x^*, \varepsilon)$ pokud $i \geq k$. Pak, omezíme-li se na $\mathcal{D} = \mathcal{B}(x^*, \varepsilon)$, podmínky ($\overline{\text{J3}}$)–($\overline{\text{J7}}$) implikují (J3)–(J7). Abychom nemuseli stále ověřovat zda pro daný index $i \in N$ již platí $x_i \in \mathcal{B}(x^*, \varepsilon)$, budeme často používat podmínky (J3)–(J7) místo ($\overline{\text{J3}}$)–($\overline{\text{J7}}$). Tím se formálně zjednoduší a zpřehlední většina důkazů aniž by došlo k újmě na obecnosti.

7.1 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 $Q_i(s)$ zavedené v poznámce 162).

Poznámka 198 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 $f_k(x_i) \to 0$ $\forall 1 \leq k \leq m$. Je-li splněna podmínka (J6), pak i

$$\|C(x_i)\| = \|\sum_{k=1}^{m} f_k(x_i) G_k(x_i)\| \leq \overline{G} \sum_{k=1}^{m} |f_k(x_i)| \to 0$$

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 105 *Nechť jsou splněny podmínky (J1), (J4), (J6). Pak Gaussova–Newtonova metoda realizovaná jako metoda s lokálně omezeným krokem je globálně konvergentní. Jsou-li splněny podmínky* $(\overline{J4})$, $(\overline{J5})$, $(\overline{J7})$ *a platí-li* $x_i \to x^*$, $F(x^*) = 0$ *a* $\omega_i(s_i) \to 0$, *je rychlost konvergence Q-superlineární.*

Důkaz Jsou-li splněny podmínky (J1), (J4), (J6), platí (F1), (F3) (věta 102) a

$$\|B_i\| = \|J(x_i) J(x_i)^T\| \leq \overline{J}^2,$$

takže Gaussova–Newtonova metoda je podle věty 75 globálně konvergentní. Jsou-li splněny podmínky $(\overline{J4})$, $(\overline{J5})$, $(\overline{J7})$ a platí-li $F(x^*) = 0$ (neboli $f(x^*) = 0$) je podle věty 103 splněn předpoklad $(\overline{F4})$ a podle věty 104 předpoklad $(\overline{F5})$. Jak již bylo ukázáno (poznámka 198) 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 $(\overline{F4})$, $(\overline{F5})$ implikuje Q-superlineární konvergenci (věta 79, kde jsme pro zjednodušení použili podmínky (F4), (F5) místo $(\overline{F4})$, $(\overline{F5})$). □

Poznámka 199 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 7.3. 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

$$\begin{bmatrix} W^{-1}, & J_i \\ J_i^T, & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0,$$

takže některé váhové koeficienty mohou být i nekonečné. To je praktické v případě úloh s omezeními, neboť čtverce omezení (s nekonečnými váhovými koeficienty) pak můžeme přidat k minimalizované funkci.

7.2 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 106 *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 200 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 $(F_i - F_{i+1})/F_i \leq \underline{\vartheta}$, 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

$$C_+ = C + \frac{z z^T}{d^T z} - \frac{Cd(Cd)^T}{d^T C d} + \frac{\beta}{d^T C d} \left(\frac{d^T C d}{d^T z} z - Cd \right) \left(\frac{d^T C d}{d^T z} z - Cd \right)^T. \tag{294}$$

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

$$C_+ = C + \frac{(z - Cd)(z - Cd)^T}{d^T (z - Cd)}. \tag{295}$$

(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 (133), kde matice B je nahražena maticí $\bar{B}$. Protože $y - \bar{B}d = z - Cd$, je výhodné použít vzorec (176). Pak

$$\begin{aligned} C_+ &= C + \frac{(y-\bar{B}d)v^T + v(y-\bar{B}d)^T}{d^T v} - \frac{(y-\bar{B}d)^T d}{d^T v}\frac{vv^T}{d^T v} \\ &= C + \frac{(z-Cd)v^T + v(z-Cd)^T}{d^T v} - \frac{(z-Cd)^T d}{d^T v}\frac{vv^T}{d^T v}, \end{aligned} \tag{296}$$

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 (295)).

Poznámka 201 Vektory y a z mohou být definovány různým způsobem, musí ale platit $z = y - J_+^T J_+ d$. Standardní volba

$$z = J_+^T f_+ - J^T f - J_+^T J_+ d$$

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 $B_k^+ s = g_k^+ - g_k$, $1 \le k \le m$. Pak můžeme psát

$$z = \sum_{k=1}^m f_k^+ B_k^+ s = \sum_{k=1}^m f_k^+ (g_k^+ - g_k) = (J_+ - J)^T f_+.$$

Metody s proměnnou metrikou pro součet čtverců lze realizovat v součinovém tvaru (věta 52). 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 111), 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

$$(J_+ + L_+)^T z = y, \qquad (J_+ + L_+)d = z, \qquad z^T z = d^T y, \tag{297}$$

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 107 *Nechť T je symetrická pozitivně definitní 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 (297) 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

$$\begin{aligned} L &= \frac{1}{2}\left\|T^{-1/2}(L_+ - L)\right\|_F^2 + u^T\left((J_+ + L_+)^T z - y\right) \\ &= \sum_{i=1}^m \left[\frac{1}{2}\left(l_i^+ - l_i\right)^T T^{-1}\left(l_i^+ - l_i\right) + u_i z^T l_i^+\right] + u^T(J_+^T z - y), \end{aligned}$$

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

$$\frac{\partial L}{\partial l_i^+} = T^{-1}\left(l_i^+ - l_i\right) + u_i z.$$

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 (297), 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 $Tz = z - \bar{A}d$. 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

$$A_+ d = \bar{A}d + (z - \bar{A}d)\frac{d^T(y - \bar{A}^T z)}{z^T(z - \bar{A}d)} = \bar{A}d + (z - \bar{A}d) = z,$$

takže je splněna i druhá podmínka z (297). □

Poznámka 202 Metodu BFGS dostaneme, zvolíme-li vektor z tak, aby byl rovnoběžný s vektorem $\bar{A}d$, tedy $z = \lambda \bar{A}d$ a $Tz = (\lambda - 1)\bar{A}d$. Z poslední podmínky v (297) 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 107 dává

$$L_+ = L + \frac{\bar{A}d}{d^T \bar{A}^T \bar{A}d}\left(\sqrt{\frac{d^T \bar{A}^T \bar{A}d}{d^T y}}\, y - \bar{A}^T \bar{A}d\right)^T. \tag{298}$$

Pokud $J_+ = 0$, takže $\bar{A} = A$, přejde tento výraz v (469).

Jistá nevýhoda aktualizace (298) 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 43). 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 108 *Frobeniova norma $\|P(L_+ - L)\|_F$ je minimální na množině všech matic vyhovujících kvazinewtonovské podmínce (297) a omezení $L_+^T f_+ = 0$ právě tehdy, platí-li*

$$L_+ = PL + \frac{\tilde{A}d}{d^T\tilde{A}^T\tilde{A}d}\left(\sqrt{\frac{d^T\tilde{A}^T\tilde{A}d}{d^Ty}}\tilde{y} - \tilde{A}^T\tilde{A}d\right)^T. \tag{299}$$

kde

$$\tilde{A} = P(J_+ + L), \qquad \tilde{y} = y - \frac{J_+f_+(J_+f_+)^Td}{f_+^Tf_+}.$$

Důkaz (a) Nejprve ukážeme, že pokud $(J_+ + L_+)d = z$ a $L_+^Tf_+ = 0$, je podmínka $(J_+ + L_+)^Tz = y$ ekvivalentní podmínce $(J_+ + L_+)^TPz = \tilde{y}$. Z $(J_+ + L_+)d = z$ a $L_+^Tf_+ = 0$ totiž plyne $f_+^TJ_+d = f_+^Tz$, takže

$$\begin{aligned}(J_+ + L_+)^TPz - \tilde{y} &= J_+^Tz - \frac{J_+^Tf_+f_+^TJ_+d}{f_+^Tf_+} + L_+^TPz - y + \frac{J_+^Tf_+f_+^TJ_+d}{f_+^Tf_+}\\ &= J_+^Tz + L_+^TPz - y = (J_+ + L_+)^Tz - y.\end{aligned}$$

Poznamenejme, že z rovností $(J_+ + L_+)d = z$ a $(J_+ + L_+)^TPz = \tilde{y}$ plyne vztah $z^TPz = d^T\tilde{y}$.
(b) Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned}L &= \frac{1}{2}\|P(L_+ - L)\|_F^2 + u^T\left((J_+ + L_+)^TPz - \tilde{y}\right)\\ &= \sum_{i=1}^m\left[\frac{1}{2}\left(l_i^+ - l_i\right)^TP\left(l_i^+ - l_i\right) + u_iz^TPl_i^+\right] + u^T(J_+^TPz - \tilde{y}),\end{aligned}$$

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

$$\frac{\partial L}{\partial l_i^+} = P\left(l_i^+ - l_i\right) + u_iPz.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $P(l_i^+ - l_i) + u_iPz = 0$, $1 \le i \le m$, neboli

$$P(L_+ - L) = -Pzu^T.$$

Z rovnosti $(J_+ + L_+)^TPz = \tilde{y}$ dostaneme $(L_+ - L)^TPz = -z^TPzu = \tilde{y} - \tilde{A}^Tz$, takže

$$u = -\frac{\tilde{y} - \tilde{A}^Tz}{z^TPz},$$

což po dosazení do předchozí rovnosti dává

$$P(L_+ - L) = \frac{Pz(\tilde{y} - \tilde{A}^TPz)^T}{z^TPz} \tag{300}$$

(neboť $P^2 = P$ implikuje $P\tilde{A} = \tilde{A}$). Použijeme-li druhou podmínku z (297), dostaneme $P(L_+ - L)d = Pz - \tilde{A}d$, takže lze psát

$$\frac{Pz(\tilde{y} - \tilde{A}^TPz)^Td}{z^TPz} = Pz - \bar{A}d.$$

Z posledního vyjádření je zřejmé, že vektor Pz je rovnoběžný s vektorem $\tilde{A}d$, neboli $Pz = \lambda\tilde{A}d$. Použijeme-li vztah $z^TPz = d^T\tilde{y}$ dokázaný v (a), můžeme psát

$$\lambda^2d^T\tilde{A}^T\tilde{A}d = z^TPz = d^T\tilde{y} \quad \Rightarrow \quad \lambda = \pm\sqrt{\frac{d^T\tilde{y}}{d^T\tilde{A}^T\tilde{A}d}},$$

což po dosazení do $Pz = \lambda\tilde{A}d$ a potom do (300) dokazuje tvrzení věty. □

Poznámka 203 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 (294) nebo (296) 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 200: Nechť $C_1 = 0$. Jestliže $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme $C_{i+1} = 0$. Jestliže $(F_i - F_{i+1})/F_i \leq \underline{\vartheta}$, aktualizujeme matici C_i pomocí (294) nebo (296). 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 (298) nebo (299).

Poznámka 204 Velmi zajímavou možnost automatického škálování matice C nabízejí totálně strukturované metody s proměnnou metrikou pocházející od Huschense. 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 (176) na matici $\bar{B} = J_+^T J_+ + \|f\|T$. Nakonec položíme $B_+ = J_+^T J_+ + \|f_+\|T_+$. Užitím vztahu (176) dostaneme

$$\begin{aligned} T_+ &= T + \frac{1}{\|f\|} \left(\frac{(y - \bar{B}d)v^T + v(y - \bar{B}d)^T}{d^T v} - \frac{(y - \bar{B}d)^T d}{d^T v} \frac{vv^T}{d^T v} \right) \\ &= T + \frac{(\tilde{z} - Td)v^T + v(\tilde{z} - Td)^T}{d^T v} - \frac{(\tilde{z} - Td)^T d}{d^T v} \frac{vv^T}{d^T v}, \end{aligned} \tag{301}$$

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

$$T_+ = T + \frac{(\tilde{z} - Ts)(\tilde{z} - Td)^T}{d^T(\tilde{z} - Td)}.$$

Následující tabulka ukazuje srovnání několika metod pro minimalizaci součtu čtverců, které jsou realizovány buď jako metody spádových směrů (první část tabulky) nebo jako metody s lokálně omezeným krokem (druhá část tabulky). Bylo řešeno 82 testovacích problémů většinou se 100 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG, jakož i počet selhání a celkový čas výpočtu).

metody spádových směrů	NIT	NFV	NFG	Čas	selhání
metoda BFGS podle (118)	9343	10853	10853	1.67	1
Gaussova–Newtonova metoda	8615	16302	24914	19.02	8
hybridní metoda (poznámka 200)	3809	6080	9884	8.89	2
strukturovaná metoda BFGS podle (296)	3158	5897	9054	7.34	2
strukturovaná metoda BFGS podle (301)	3262	6085	9345	6.97	1
metody s lokálně omezeným krokem	NIT	NFV	NFG	Čas	selhání
metoda BFGS podle (136)	10684	11860	10764	3.39	1
Gaussova–Newtonova metoda	4321	4694	4402	12.84	1
hybridní metoda (poznámka 200)	3450	4013	3531	9.67	-
strukturovaná metoda BFGS podle (296)	2766	3130	2847	7.61	-
strukturovaná metoda BFGS podle (301)	2771	3239	2849	7.66	-

Poznámka 205 Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Metody s proměnou metrikou mají menší režii, neboť není třeba řešit soustavy lineárních rovnic. Výpočetní čas je tedy obvykle nižší než u specializovaných metod pro součet čtverců. Metody s proměnou metrikou není vhodné realizovat jako metody s lokálně omezeným krokem.
- 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 a to buď pomocí jednoduché hybridní strategie (poznámka 200) nebo pomocí strukturovaných aktualizací (296) a (301). 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 (souvisí to s dobrými konvergenčími vlastnostmi kombinovaných metod). Jejich vyšší režijní nároky mohou být vykompenzovány rychlejší konvergencí v případech, kdy je výpočet hodnoty (a gradientu) minimalizované funkce velmi náročný.

7.3 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 109 *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 2 a z vyjádření (287). Postačitelnost plyne z toho, že matice J^TJ je pozitivně definitní, takže kvadraticá funkce $\|Js+f\|^2$ je konvexní a má tedy jediný stacionární bod, který je jejím globálním minimem. Podle poznámky 102 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 24 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 206 V případě, že matice J má plnou hodnost, lze řešení lineární úlohy nejmenších čtverců získat řešením normální soustavy rovnic $J^TJs = -f$ 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 34 *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

$$Q^T Q = \begin{bmatrix} c, & s \\ -s, & c \end{bmatrix} \begin{bmatrix} c, & -s \\ s, & c \end{bmatrix} = \begin{bmatrix} c^2 + s^2, & 0 \\ 0, & s^2 + c^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Poznámka 207 Nechť $\|x\| = 1$, takže $x_1 = \cos\alpha$ 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 110 *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 208 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 209 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 210 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

$$\begin{bmatrix} \boxed{\begin{matrix} * & * & * & * \\ * & * & * & * \end{matrix}} \\ \begin{matrix} 0 & * & * & * \\ 0 & 0 & * & * \end{matrix} \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ \boxed{\begin{matrix} 0 & * & * & * \\ 0 & * & * & * \end{matrix}} \\ \begin{matrix} 0 & 0 & * & * \end{matrix} \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ \boxed{\begin{matrix} 0 & 0 & * & * \\ 0 & 0 & * & * \end{matrix}} \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & * & * \\ 0 & 0 & 0 & * \end{bmatrix}$$

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 35 *Řekneme, že matice $Q \in R^{n\times n}$ je Householderovou maticí elementární reflexe, jestliže platí*

$$Q = I - 2\frac{vv^T}{v^T v},$$

kde $v \in R^n$. Tato matice je ortogonální, neboť

$$Q^T Q = \left(I - 2\frac{vv^T}{v^T v}\right)\left(I - 2\frac{vv^T}{v^T v}\right) = I - 4\frac{vv^T}{v^T v} + 4\frac{vv^T}{v^T v} = I.$$

Poznámka 211 Nechť $x \in R^n$ a $v \in R^n$. Nechť vektor y je ortogonální projekcí vektoru x do směru vektoru v, takže

$$y = \frac{vv^T}{v^T v}x = \frac{v^T x}{v^T v}v.$$

Vynásobíme-li vektor x maticí Q, dostaneme

$$Qx = \left(I - 2\frac{vv^T}{v^T v}\right)x = x - 2\frac{v^T x}{v^T v}v = x - 2y,$$

takže vektor Qx vznikne zrcadlením vektoru x podle nadroviny kolmé k vektoru v.

Věta 111 *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 35. Platí

$$\begin{aligned} Qx &= x - 2(x-y)\frac{(x-y)^T x}{(x-y)^T(x-y)} = x - 2(x-y)\frac{\|x\|^2 - y^T x}{\|x\|^2 + \|y\|^2 - 2y^T x} \\ &= x - 2(x-y)\frac{\|x\|^2 - y^T x}{2(\|x\|^2 - y^T x)} = x - (x-y) = y. \end{aligned}$$

Vztah $x = Qy$ plyne z toho, že matice Q je symetrická a ortogonální. □

Poznámka 212 Householderovu matici můžeme použít k vynulování posledních $n-1$ prvků vektoru $x \in R^n$. Položíme-li ve větě 111 $y = -\sigma\|x\|e_1$, 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

$$Qz = z - \frac{v^Tz}{v^Tv}v = z - \frac{z^Tx + \sigma\|x\|z_1}{\|x\|(\|x\| + |x_1|)(x + \sigma\|x\|e_1)}.$$

Položíme-li $\tilde{x} = x/\|x\|$, můžeme psát $v = \tilde{x} + \sigma e_1$, $v^Tv = 2(1 + |\tilde{x}_1|)$ a pro libovolný vektor $z \in R^n$ platí

$$Qz = z - \frac{z^T\tilde{x} + \sigma z_1}{1 + |\tilde{x}_1|}(\tilde{x} + \sigma e_1).$$

Definice 36 *Ortogonálním rozkladem matice $J \in R^{m\times n}$, $m \geq n$, nazveme vyjádření*

$$J = Q \begin{bmatrix} R \\ 0 \end{bmatrix}, \tag{302}$$

kde Q je ortogonální matice řádu m a R je horní trojúhelníková matice řádu n.

Poznámka 213 K nalezení ortogonálního rozkladu (302) 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

$$Q_n \ldots Q_2Q_1J = \begin{bmatrix} R \\ 0 \end{bmatrix}$$

(matice Q_i vynuluje $m-i$ prvků i-tého sloupce průběžně upravované matice). Ortogonální rozklad se provádí podle schematu

$$\left[\begin{array}{|ccc|}\hline * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ * & * & * \\ \hline\end{array}\right] \to \begin{bmatrix} * & * & * \\ 0 & \boxed{\begin{matrix} * & * \\ * & * \\ * & * \\ * & * \end{matrix}} \\ 0 & \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & \boxed{\begin{matrix} * \\ * \\ * \end{matrix}} \\ 0 & 0 & \end{bmatrix} \to \begin{bmatrix} * & * & * \\ 0 & * & * \\ 0 & 0 & * \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

Pak platí (302), kde $Q = Q_1Q_2 \ldots Q_n$. Jelikož Householderovy matice Q_i, $1 \leq i \leq n$, jsou symetrické a ortogonální, platí $Q^TQ = Q_n \ldots Q_2Q_1Q_1Q_2 \ldots Q_n = I$, takže matice Q je ortogonální.

Poznámka 214 Má-li matice J hodnost $l < n$, není trojúhelníková matice R regulární a rozklad (302) 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

$$J = Q \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \tilde{Q}^T, \tag{303}$$

kde $\tilde{Q} = \tilde{Q}_1\tilde{Q}_2\ldots\tilde{Q}_l$. Dodatečné ortogonální úpravy se provádí podle schematu

$$\begin{bmatrix} * & * & * & * \\ 0 & * & * & * \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \to \begin{bmatrix} * & * & * & * \\ 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \to \left[\begin{array}{cc|cc} * & * & 0 & 0 \\ 0 & * & 0 & 0 \\ \hline 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{array}\right]$$

Věta 112 *Nechť*

$$J = Q \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \tilde{Q}^T,$$

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

$$s = \tilde{Q} \begin{bmatrix} -R^{-1}u \\ 0 \end{bmatrix}, \qquad Q^T f = \begin{bmatrix} u \\ v \end{bmatrix}.$$

Důkaz Nechť

$$Q^T Js = Q^T J\tilde{Q} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} R, & 0 \\ 0, & 0 \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix}, \quad Q^T f = \begin{bmatrix} u \\ v \end{bmatrix}.$$

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 jsme zvolili

$$s = \tilde{Q} \begin{bmatrix} y \\ z \end{bmatrix},$$

platí $\|s\|^2 = \|y\|^2 + \|z\|^2$, takže $\|s\|$ je minimální, pokud $z = 0$. Použijeme-li získané vektory, dostaneme tvrzení věty. □

Poznámka 215 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.

8 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 často 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 verzi Newtonovy metody.

8.1 Metody s proměnnou metrikou s omezenou pamětí

Definice 37 *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 216 Aktualizace uvedené v definici 37 jsou stejné jako aktualizace použité v definici 23. Lze je tedy vyjádřit ve tvaru

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + \frac{\rho_j^i}{\gamma_j^i} \frac{1}{b_j} d_j d_j^T - \frac{1}{a_j^i} H_j^i y_j (H_j^i y_j)^T + \frac{\eta_j^i}{a_j^i} \left(\frac{a_j^i}{b_j} d_j - H_j^i y_j \right) \left(\frac{a_j^i}{b_j} d_j - H_j^i y_j \right)^T \right) \tag{304}$$

a invertovat tak, že platí

$$B_{j+1}^i = \frac{1}{\gamma_j^i} \left(B_j^i + \frac{\gamma_j^i}{\rho_j^i} \frac{1}{b_j} y_j y_j^T - \frac{1}{c_j^i} B_j^i d_j (B_j^i d_j)^T + \frac{\beta_j^i}{c_j^i} \left(\frac{c_j^i}{b_j} y_j - B_j^i d_j \right) \left(\frac{c_j^i}{b_j} y_j - B_j^i d_j \right)^T \right), \tag{305}$$

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.

Zvláště výhodná je metoda BFGS s omezenou pamětí s aktualizací

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + \left(\frac{a_j^i}{b_j} + \frac{\rho_j^i}{\gamma_j^i} \right) \frac{1}{b_j} d_j d_j^T - \frac{1}{b_j} \left(H_j^i y_j d_j^T + d_j (H_j^i y_j)^T \right) \right), \tag{306}$$

$i-m \le j \le i-1$, pro kterou platí tato věta.

Věta 113 *Nechť x_i, $i \in N$, je posloupnost generovaná $\bar{m}$-krokovou metodou BFGS s omezenou pamětí s přesným výběrem délky kroku (takže $s_i^T g_{i+1} = 0 \ \forall i \in N$) aplikovaná na ryze konvexní kvadratickou funkci*

$$Q(x) = \frac{1}{2}(x - x^*)^T G (x - x^*). \tag{307}$$

Pak platí

$$s_i = -\left(\prod_{k=i-m}^{i-1} \gamma_k^i\right)\left(Hg_i - \frac{y_{i-1}^T Hg_i}{y_{i-1}^T d_{i-1}} d_{i-1}\right) \tag{308}$$

pro $1 \leq i \leq n$ *(směrové vektory* s_i*,* $1 \leq i \leq n$*, jsou rovnoběžné se směrovými vektory generovanými předpodmíněnou 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 (308) (důsledek 5 a jeho důkaz). Důkaz pro $i > \bar{m}$ provedeme indukcí. Předpokládejme, že (308) platí pro $1 \leq i < k$, kde $\bar{m} < k \leq n$. Pak podle věty 22 a poznámky 50 lze pro $1 \leq j < k$ psát

$$d_j^T g_k = 0, \qquad g_j^T H g_k = 0. \tag{309}$$

Ukážeme nejprve sestupnou indukcí, že pro libovolný index $k - m \leq i \leq k - 1$ platí

$$H_k^k g_k = \left(\prod_{j=i}^{k-1} \gamma_j^k\right)\left(H_i^k g_k - \sum_{j=i}^{k-1} \frac{y_j^T H_i^k g_k}{y_j^T d_j} d_j\right). \tag{310}$$

Platí to zřejmě pro $i = k - 1$, neboť z (306) a z $d_{k-1}^T g_k = 0$ (přesný výběr délky kroku) plyne, že

$$H_k^k g_k = \gamma_k^k \left(H_{k-1}^k g_k - \frac{y_{k-1}^T H_{k-1}^k g_k}{y_{k-1}^T d_{k-1}} d_{k-1}\right).$$

Nyní snížíme i o jedničku. Použijeme-li (310), (306) a rovnost $d_{i-1}^T g_k = 0$, která plyne z (309), můžeme psát

$$H_k^k g_k = \left(\prod_{j=i}^{k-1} \gamma_j^k\right)\left(\gamma_{i-1}^k \left(H_{i-1}^k g_k - \frac{y_{i-1}^T H_{i-1}^k g_k}{y_{i-1}^T d_{i-1}} d_{i-1}\right) - \gamma_{i-1}^k \sum_{j=i}^{k-1} \frac{y_j^T H_{i-1}^k g_k}{y_j^T d_j} d_j\right) \tag{311}$$

neboť podle (309) pro $i \leq j \leq k - 1$ platí $y_j^T d_{i-1} = (g_{j+1} - g_j)^T d_{i-1} = 0$, takže

$$y_j^T H_i^k g_k = \gamma_{i-1}^k y_j^T \left(H_{i-1}^k g_k - \frac{y_{i-1}^T H_{i-1}^k g_k}{y_{i-1}^T d_{i-1}} d_{i-1}\right) = \gamma_{i-1}^k y_j^T H_{i-1}^k g_k.$$

Jelikož vztah (311) je ekvivalentní s (310), kde i je nahraženo $i - 1$, je sestupný indukční krok ukončen. Můžeme tedy psát

$$s_k = -H_k^k g_k = -\left(\prod_{j=k-m}^{k-1} \gamma_j^k\right)\left(Hg_k - \sum_{j=k-m}^{k-1} \frac{y_j^T Hg_k}{y_j^T d_j} d_j\right) \tag{312}$$

(neboť $H_{k-m}^k = H$). Podle (309) však pro $j < k - 1$ platí $y_j^T Hg_k = (g_{j+1} - g_j)^T Hg_k = 0$, takže (312) přejde na (308) s $i = k$, čímž je hlavní indukční krok dokončen. □

Důsledek 16 *(Kvadratické ukončení). Nechť jsou splněny předpoklady věty 113. Pak existuje index* $k \leq n$ *takový, že* $g_{k+1} = 0$ *a* $x_{k+1} = x^*$*.*

Důkaz Podle věty 113 jsou směrové vektory generované $\bar{m}$-krokovou metodou s proměnnou metrikou s omezenou pamětí rovnoběžné s vektory generovanými metodou sdružených gradientů. Podle věty 22 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 217 Věta 113 a důsledek 16 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 41).

Jednu z dalších výhod metody BFGS s omezenou pamětí ukazuje tato věta.

Věta 114 *Nechť jsou splněny předpoklady věty 113. Pak pro* $1 \leq i \leq n$ *a* $i - m \leq j \leq i - 1$ *platí*

$$H_i^i y_j = \left(\prod_{k=j}^{i-1} \gamma_k^i \right) \frac{\rho_j^i}{\gamma_j^i} d_j$$

(je splněno m kvazinewtonovských podmínek).

Důkaz Důkaz provedeme indukcí. Použijeme toho, že podle věty 113 jsou směrové vektory s_i, $1 \leq i \leq n$, G–ortogonální, takže pro $1 \leq j < l$ lze psát

$$d_l^T y_j = 0, \qquad y_l^T d_j = 0. \tag{313}$$

Předpokládejme, že pro nělaký index $i - m \leq l < i$ a pro všechny indexy $i - m \leq j \leq l - 1$ platí

$$H_l^i y_j = \left(\prod_{k=j}^{l-1} \gamma_k^i \right) \frac{\rho_j^i}{\gamma_j^i} d_j, \triangleq \lambda_j^{l-1} d_j \tag{314}$$

(platí to pro $l = i-n$, neboť v tomto případě neexistuje žádný index j splňující nerovnost $i-m \leq j \leq l-1$). Použijeme-li (306), (313) a (314), dostaneme

$$\begin{aligned} H_{l+1}^i y_j &= \gamma_l^i \left(H_l^i y_j + \left(\frac{a_l^i}{b_l} + \frac{\rho_l^i}{\gamma_l^i} \right) \frac{1}{b_l} d_l d_l^T y_j - \frac{1}{b_l} \left(H_l^i y_l d_l^T y_j + d_l y_l^T H_l^i y_j \right) \right) \\ &= \gamma_l^i \left(\lambda_j^{l-1} d_j + \frac{1}{b_l} \lambda_j^{l-1} y_l^T d_j d_l \right) = \lambda_j^l d_j \end{aligned}$$

pro $i - m \leq j \leq l - 1$ a jelikož podle definice 37 platí $H_{l+1}^i y_l = \rho_l d_l = \lambda_l^l d_l$, můžeme psát $H_{l+1}^i y_j = \lambda_j^l d_l$ pro $i - m \leq j \leq l$. Platí tedy (314) pro index o jedničku vyšší. □

Nyní dokážeme globální konvergenci metod s proměnnou metrikou s omezenou pamětí.

Věta 115 *Uvažujme* $\bar{m}$*-krokovou metodu s proměnnou metrikou s omezenou pamětí (304) takovou, že* $\underline{\gamma} \leq \gamma_j^i \leq \overline{\gamma}$, $\underline{\rho} \leq \rho_j^i \leq \overline{\rho}$ *a* $0 \leq \eta_j^i \leq \overline{\eta}$. *Nechť funkce* $F : \mathcal{D} \to R$ *splňuje podmínky (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 $m \leq \bar{m}$ aktualizací (304), které jsou formálně shodné s aktualizacemi vyšetřovanými v oddílu 4.5. Můžeme tedy použít (305) a postupovat stejně jako v důkazu lemmatu 24. Použijeme-li (195) pro $i - m \leq j \leq i - 1$, můžeme psát

$$\|B_{j+1}^i\| \leq \mathrm{Tr} B_{j+1}^i \leq \overline{C}^{\bar{m}} \triangleq C.$$

Podobně podle (142) a (196) dostaneme

$$\frac{\det B_{j+1}^i}{\det B_j^i} \geq \left(\frac{1}{\overline{\gamma}} \right)^n \frac{\underline{\gamma}}{\overline{\rho}} \frac{1}{\overline{\eta}} \frac{y_j^T d_j}{d_j^T B_j^i d_j} \geq \underline{K} \frac{1}{\|B_j^i\|} \frac{y_j^T d_j}{d_j^T d_j} \geq \frac{\underline{K}\,\underline{G}}{C},$$

takže

$$\det B_i^i \geq \left(\frac{\underline{K}\,\underline{G}}{C}\right)^{\overline{m}} \det B \triangleq K,$$

kde $B = H^{-1}$. Můžeme tedy psát

$$\kappa(B_i^i) = \|B_i^i\|\|(B_i^i)\|^{-1} = \frac{\overline{\lambda}(B_i^i)}{\underline{\lambda}(B_i^i)} \leq \frac{\|B_i^i\|^n}{\det B_i^i} \leq \frac{C^n}{K} \triangleq \overline{\kappa},$$

což podle poznámky 26 dává $\lim_{i\to\infty} \|g_i\| = 0$. □

Poznámka 218 Numerické testy ukazují, že je výhodné použít pouze počáteční škálování a to zahrnout do výběru matice H_{i-m}^i (obvykle pokládáme $H_{i-m}^i = (y_{i-1}^T d_{i-1}/y_{i-1}^T y_{i-1})I$). Proto budeme předpokládat, že $\gamma_j^i = 1$, $i-m \leq j \leq i-1$. Dále budeme předpokládat, že $\rho_j^i = \rho_j$, $i-m \leq j \leq i-1$.

Metody s proměnnou metrikou s omezenou pamětí lze realizovat přirozeným způsoben tak, že kromě vektorů d_j, y_j, $i-m \leq j \leq i-1$ počítáme a ukládáme vektory $H_j^i y_j$, $i-m \leq j \leq i-1$. 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. Tato reprezentace je založená na pseudosoučinovém tvaru (121), podle kterého pro metodu BFGS platí

$$H_{j+1}^i = V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T, \qquad V_j = I - \frac{1}{b_j} y_j d_j^T, \tag{315}$$

kde $y_j = g_{j+1} - g_j$, $d_j = x_{j+1} - x_j$ a $b_j = y_j^T d_i$ pro $i-m \leq j \leq i-1$.

Věta 116 *Nechť H_{j+1}^i je matice získaná v j-tém kroku metody BFGS. Pak platí*

$$H_{j+1}^i = \left(\prod_{k=i-m}^{j} V_k\right)^T H_{i-m}^i \left(\prod_{k=i-m}^{j} V_k\right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k\right).$$

Důkaz (Indukcí) Pro $j = i-m$ to bezprostředně plyne z (315). Indukční krok vypadá takto

$$\begin{aligned} H_{j+1}^i &= V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T = V_j^T \left(\prod_{k=i-m}^{j-1} V_k\right)^T H_{i-m}^i \left(\prod_{k=i-m}^{j-1} V_k\right) V_j + \\ &\quad + \sum_{l=i-m}^{j-1} \frac{\rho_l}{b_l} V_j^T \left(\prod_{k=l+1}^{j-1} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j-1} V_k\right) V_j + \frac{\rho_j}{b_j} d_j d_j^T \\ &= \left(\prod_{k=1}^{i} V_k\right)^T H_1 \left(\prod_{k=i-m}^{j} V_k\right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k\right). \end{aligned}$$

□

Důsledek 17 *Nechť jsou splněny předpoklady věty 116. Pak vektor $s_i = -H_i^i g_i$ lze získat pomocí dvou rekurentních vztahů (Strangovy rekurence). Nejprve se položí $u_i = -g_i$ a zpětnou rekurencí*

$$\sigma_j = \frac{d_j^T u_{j+1}}{b_j}, \qquad u_j = u_{j+1} - \sigma_j y_j \tag{316}$$

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

$$v_{j+1} = v_j + \left(\rho_j \sigma_j - \frac{y_j^T v_j}{b_j} \right) d_j \tag{317}$$

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 (316). 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

$$v_{j+1} = V_j^T v_j + \frac{\rho_j}{b_j} d_j d_j^T u_{j+1} = \left(I - \frac{1}{b_j} d_j y_j^T \right) v_j + \rho_j \sigma_j^i d_j = v_j + (\rho_j \sigma_j - y_j^T v_j / b_j) d_j,$$

což je právě rekurence (317). □

Poznámka 219 Tvrzení věty 116 ukazuje, že matici H_i^i můžeme určit z matice H^i_{i-m} (která je řídká) pomocí vektorů d_j, y_j, $i-m \leq j \leq i-1$. Matici H_i^i nemusíme konstruovat explicitně. Podle důsledku 17 stačí počítat vektor $s_i = -H_i^i g_i$, pomocí rekurentních vztahů (316)–(317). V těchto vztazích je třeba uchovávat čísla σ_j, $i-m \leq j \leq i-1$. Vektory u_j, v_j, $i-m \leq j \leq i-1$ 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, $i-\bar{m} \leq j \leq i-1$, 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 11 Data $\bar{m} < n$, $\varepsilon > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

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 $\|g_i\| \leq \varepsilon$, 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 (316)–(317).

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 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.

Strangovy rekurence lze použít pouze pro metodu BFGS, kdy se v (315) nevyskytují vektory $H_j^i y_j$, $i-m \le j \le i-1$. Aby bylo možné realizovat libovolnou metodu z Broydenovy třídy, je třeba tyto vektory nějak aproximovat, například nahradit je vektory $H_j^j y_j$, $i-m \le j \le i-1$, které se používají v předchozích iteračních krocích. Touto modifikací se ztratí některé výhodné teoretické vlastnosti (neplatí věta 113 ani její důsledek 16). Nicméně některé z těchto metod dávají dobré praktické výsledky. Při jejich popisu budeme používat označení $H_i = H_i^i$, $i \in N$.

Teoreticky by bylo možné použít obecný pseudosoučinový tvar (121). Metody získané tímto způsovem nejsou příliš vhodné, neboť je třeba ukládat m vektorů navíc. Proto je účelné vyjádřit Broydenovu třídu ve tvaru formálně shodném s metodou BFGS, neboli nalézt vektor $\hat{d} = d + \lambda Hy$ tak, aby platilo

$$\frac{1}{\gamma} H_+ = H + \hat{U}\hat{M}\hat{U}^T = H + [d + \lambda Hy, Hy] \begin{bmatrix} \hat{m}_1, & \hat{m}_2 \\ \hat{m}_2, & 0 \end{bmatrix} [d + \lambda Hy, Hy]^T. \tag{318}$$

Lemma 34 *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

$$\begin{aligned} \det(\hat{U}^T\hat{U}) &= (d+\lambda Hy)^T(d+\lambda Hy)(Hy)^THy - \left((d+Hy)^THy\right)^2 \\ &= \left(d^Td + 2\lambda d^THy + \lambda^2(Hy)^THy\right)(Hy)^THy - \left(d^THy + \lambda(Hy)^THy\right)^2 \\ &= d^Td(Hy)^THy - (d^THy)^2 = \det(U^TU). \end{aligned}$$

Podle důsledku 7 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 117 *Aktualizaci z Broydenovy třídy (vzorec (116)) můžeme vyjádřit ve tvaru*

$$\frac{1}{\gamma} H_+ = H + \left(\frac{a}{\hat{b}} + \frac{\hat{\rho}}{\gamma}\right)\frac{1}{\hat{b}}\hat{d}\hat{d}^T - \frac{1}{\hat{b}}\left(Hy\hat{d}^T + \hat{d}(Hy)^T\right) = \hat{V}^TH\hat{V} + \frac{\hat{\rho}}{\gamma\hat{b}}\hat{d}\hat{d}^T, \quad \hat{V} = I - \frac{1}{\hat{b}}y\hat{d}^T, \tag{319}$$

kde

$$\hat{d} = d + \lambda Hy, \quad \lambda = \frac{m_2 + \sqrt{\mu}}{m_1}, \quad \hat{b} = \frac{1}{\sqrt{\mu}}, \quad \frac{\hat{\rho}}{\hat{b}} = \eta\frac{\rho}{b}, \tag{320}$$

přičemž m_1, m_2 jsou odpovídající prvky matice M a $\mu = -\det M$ je výraz určený vzorcem (115)), neboli

$$m_1 = \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right), \quad m_2 = -\frac{\eta}{b}, \quad \mu = \frac{1}{ab}\left(\eta\frac{a}{b} + (1-\eta)\frac{\rho}{\gamma}\right).$$

Důkaz Podle lemmatu 34 platí $\hat{m}_2^2 = -\det \hat{M} = -\det M = \mu$, kde μ je výraz určený vztahem (115). Protože je vhodné, aby platilo $\hat{b} = -1/\hat{m}_2 > 0$, volíme $\hat{m}_2 = -\sqrt{\mu}$. Porovnáme-li koeficiemty u dd^T a $Hyd^T + d(Hy)^T$ ve výrazech (318) a (103), 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

$$\lambda = \frac{m_2 - \hat{m}_2}{\hat{m}_1} = \frac{m_2 + \sqrt{\mu}}{m_1}.$$

Jelikož $\hat{m}_1 = m_1$ a $1/\hat{b}^2 = \hat{m}_2^2 = \mu$, můžeme s použitím (115) psát

$$\hat{m}_1 = \frac{1}{\hat{b}}\left(\frac{a}{\hat{b}} + \frac{\hat{\rho}}{\gamma}\right) = a\mu + \frac{\hat{\rho}}{\gamma\hat{b}} = \frac{1}{b}\left(\eta\frac{a}{b} + (1-\eta)\frac{\rho}{\gamma}\right) + \frac{\hat{\rho}}{\gamma\hat{b}} = m_1 = \frac{1}{b}\left(\eta\frac{a}{b} + \frac{\rho}{\gamma}\right),$$

takže $\hat{\rho}/\hat{b} = \eta\rho/b$. □

Pseudosoučinový tvar uvedený v (319) lze použít k realizaci libovolné modifikované metody s proměnnou metrikou s omezenou pamětí z Broydenovy třídy pomocí Strangových rekurencí. Kromě směrových vektorů

$s_i = -H_i g_i$, $i \in N$, je však třeba počítat vektory $H_i y_i$, $i \in N$, což zvyšuje počet operací. K odstranění této nevýhody by bylo vhodné počítat směrový vektor podle některého ze vzorců (218) nebo (224) a Strangovy rekurence použít pouze pro výpočet vektoru $H_i y_i$. Problém je v tom, že pro takto získané směrové vektory platí $s_i = -H_i g_i = -H_i^i g_i$ 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 $H_{i-1} = H_{i-1}^{i-1}$, použité pro výpočet směrového vektoru s_{i-1}.

Poznámka 220 Kdybychom použili vzorec (218) v případě, že $s \neq -Hg$, mohlo by se stát, že by získaný směrový vektor nebyl spádový. Proto je výhodnější použít vzorec

$$s_+ = -\frac{d^T g_+}{b} d - \frac{b + \eta\tau}{b + \tau} V^T p, \qquad V = I - \frac{1}{b} y d^T, \tag{321}$$

kde $p = HVg_+$ a $\tau = \max(\alpha p^T y, b)$, který je modifikací vzorce (224) a který zaručuje, že $g_+^T s_+ < 0$ (poznámka 150). Vektor $p = HVg_+$ můžeme (tak jako v důkazu věty 70) použít k výpočtu vektoru Hy. Položíme

$$Hy = \frac{b}{c}(d + \alpha p), \qquad a = y^T Hy = \frac{b}{c}(b + \alpha y^T p), \qquad b = d^T y, \qquad c = -\alpha d^T g. \tag{322}$$

Poznámka 221 K výpočtu vektoru $p_i = H_i V_i g_i$ lze použít modifikované Strangovy rekurence (odvozené z aktualizace (319)), kde vystupují veličiny se stříškou. Nejprve položíme

$$u_i = V_i g_i = g_i - \frac{d_{i-1}^T g_i}{y_{i-1}^T d_{i-1}} y_{i-1}$$

a zpětnou rekurencí

$$\sigma_j = \frac{\hat{d}_j^T u_{j+1}}{\hat{b}_j}, \qquad u_j = u_{j+1} - \sigma_j^i y_j \tag{323}$$

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

$$v_{j+1} = v_j + \left(\hat{\rho}_j \sigma_j - \frac{y_j^T v_j}{\hat{b}_j} \right) \hat{d}_j, \tag{324}$$

spočítáme vektory v_{j+1}, $i - m \leq j \leq i - 1$. Nakonec položíme $p_i = v_i$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 12 Data $\bar{m} < n$, $\varepsilon > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

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 $\|g_i\| \leq \varepsilon$, 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 podle vzorce (321).

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 vektor p_i pomocí rekurentních vztahů (323)–(324). Určíme vektor $H_i y_i$ a číslo a_i podle (322). Zvolíme hodnotu $\eta_i \geq 0$, určíme vektor $\hat{d}_i$ a číslo $\hat{b}_i$ podle (320) a položíme $\hat{\rho}_i := \eta_i \hat{b}_i / b_i$.

Krok 5 Uložíme vektory $\hat{d}_i$, y_i a čísla $\hat{b}_i$, $\hat{\rho}_i$ do pracovního pole. Pokud $m = \bar{m}$, odstraníme vektory $\hat{d}_{i-m}$, y_{i-m} a čísla $\hat{b}_{i-m}$, $\hat{\rho}_{i-m}$ z pracovního pole. Zvětšíme i o 1 a přejdeme na krok 2.

Popíšeme ještě jednu modifikaci metody BFGS, která využívá vektory z předchozího iteračího kroku a která jako metoda s omezenou pamětí je velmi efektivní.

Věta 118 *Nechť H je symetrická pozitivně definitní matce a $Hy_- = d_-$. Nechť $\bar{d} = d - \lambda d_-$, $\bar{y} = y - \lambda y_-$ a $\bar{b} = \bar{d}^T y \neq 0$. Položme*

$$\bar{H}_+ = \frac{\bar{\rho}}{\bar{b}} \bar{d}\bar{d}^T + \bar{V}^T H \bar{V}, \qquad \bar{V} = I - \frac{1}{\bar{b}} \bar{y}\bar{d}^T. \tag{325}$$

Pak pokud

$$\bar{\rho} = \left(1 - \lambda^2 \frac{b_-}{b}\right) \frac{b}{\bar{b}}, \tag{326}$$

platí $\bar{H}_+ y = d$ a jestliže $\lambda^2 < b/b_-$, je matice $\bar{H}_+$ pozitivně definitní.

Důkaz (a) Protože $Hy_- = d_-$, platí $H\bar{V}y = H(y - \bar{y}) = \lambda H y_- = \lambda d_-$. Použijeme-li (325), můžeme psát

$$\bar{H}_+ y = \left(\frac{\bar{\rho}}{\bar{b}} \bar{d}\bar{d}^T + \bar{V}^T H \bar{V}\right) y = \bar{\rho}\bar{d} + \lambda \bar{V}^T d_- = \left(\bar{\rho} - \lambda \frac{\bar{y}^T d_-}{\bar{b}}\right) \bar{d} + \lambda d_-$$

a je-li výraz v závorce roven 1, platí $\bar{H}_+ y = \bar{d} + \lambda d_- = d$. Ale

$$\bar{\rho} - \lambda \frac{\bar{y}^T d_-}{\bar{b}} = \bar{\rho} - \lambda \frac{y^T d_-}{\bar{b}} + \lambda^2 \frac{b_-}{\bar{b}} = \bar{\rho} - \frac{y^T(d - \bar{d})}{\bar{b}} + \lambda^2 \frac{b_-}{\bar{b}} = \bar{\rho} - \frac{b}{\bar{b}} + 1 + \lambda^2 \frac{b_-}{\bar{b}},$$

takže tento výraz je roven 1, pokud platí (326).

(b) Z vyjádření (325) plyne, že matice $\bar{H}_+$ je pozitivně definitní, pokud $\bar{\rho}/\bar{b} > 0$. Pak z $v^T \bar{d} \neq 0$ plyne $v^T \bar{H}_+ v \geq (v^T \bar{d})^2 \bar{\rho}\bar{b} > 0$ a z $v^T \bar{d} = 0$ plyne $Vv = v$, takže $v^T \bar{H}_+ v = v^T H v > 0$ (pokud $\|v\| > 0$), neboť matice H je pozitivně definitní. Jestliže $\bar{\rho} = (1 - \lambda^2 b_-/b) b/b_-$, je podmínka $\bar{\rho}/\bar{b} > 0$ splněna, pokud $1 - \lambda^2 b_-/b > 0$, neboli $\lambda^2 < b/b_-$. □

Poznámka 222 Položíme-li $\lambda = 0$, dostaneme standardní metodu BFGS. Položíme-li $\lambda = d^T y_-/d_-^T y_-$, platí $\bar{d}^T y_- = d^T y_- - \lambda d_-^T y_- = 0$, takže $\bar{d}^T \bar{y} = \bar{d}^T y = \bar{b}$. Dostaneme tak metodu BFGS, kde vystupují veličiny s pruhem.

Věta 119 *Nechť jsou splněny předpoklady věty 118, přičemž $y = Gd$, $y_- = Gd_-$, kde G je symetrická pozitivně definitní matice (platí to, minimalizujeme-li ryze konvexní kvadratickou funkci (307)). Předpokládejme, že vektory d, d_- jsou lineárně nezávislé a položme $\lambda = d^T y_-/d_-^T y_-$. Pak platí $\lambda^2 < b/b_-$, $\bar{b} > 0$ a $\bar{H}_+ y_- = d_-$ (jsou splněny dvě kvazinewtonovské podmínky).*

Důkaz Podle Schwarzovy nerovnosti platí

$$\frac{b}{b_-} - \lambda^2 = \frac{d^T y\, d_-^T y_- - (d^T y_-)^2}{(d_-^T y_-)^2} = \frac{d^T G d\, d_-^T G d_- - (d^T G d_-)^2}{(d_-^T G g_-)^2} > 0,$$

$$\bar{b} = d^T y - \lambda d_-^T y = \frac{d^T y\, d_-^T y_- - d^T y_-\, d_-^T y}{d_-^T y_-} = \frac{d^T G d\, d_-^T G d_- - (d^T G d_-)^2}{d_-^T G g_-^2} > 0,$$

neboť předpokláme, že vektory d, d_- jsou lineárně nezávislé a matice G je pozitivně definitní. Dále dostaneme

$$\bar{H}_+ y_- = \bar{V}^T H \bar{V} y_- = \bar{V}^T H \left(I - \frac{1}{\bar{b}} \bar{y}\bar{d}^T\right) y_- = \bar{V}^T H y_- = \bar{V}^T d_- = \left(I - \frac{1}{\bar{b}} \bar{d}\bar{y}^T\right) d_- = d_-,$$

neboť podle poznámky 222 platí $\bar{d}^T y_- = 0$, a podle předpokladu také $\bar{y}^T d_- = \bar{d}^T G d_- = \bar{d}^T y_- = 0$. □

Poznámka 223 V dalším textu budeme předpokládat, že $\lambda = \sigma\sqrt{b/b_-}$, takže $\bar\rho/\bar b > 0$, pokud $|\sigma| < 1$. V tomto případě platí

$$\bar d = d - \sigma\sqrt{\frac{b}{b_-}}d_-, \qquad \bar y = y - \sigma\sqrt{\frac{b}{b_-}}y_-, \qquad 0 < \sigma < 1. \tag{327}$$

Zbývá ukázat, jak se určí hodnota parametru σ. Numerické testy ukazují, že hodnoty $|\sigma| \approx 1$ značně zhoršují konvergenci metody a hodnoty $|\sigma| \le 1/2$ (opatřené vhodným znaménkem) dávají dobré výsledky. Budeme tedy předpokládat, že $|\sigma| \le \overline\sigma < 1$, kde číslo $\overline\sigma$ není o mnoho větší než $1/2$. Podle poznámky 222 a věty 119 je výhodné volit znaménko parametru σ podle znaménka čísla $d^T y_-$, nebo čísla $d_-^T y$, pokud $|d_-^T y| \ge 20|d^T y_-|$ (hodnota 20 byla získána experimentálně). K důkazu globální konvergence budeme potřebovat, aby platilo $\bar b \ge (1-\overline\lambda)b$, kde $0 < \overline\lambda < 1$, neboli $\sigma d_-^T y \le \overline\lambda\sqrt{bb_-}$. Není-li tato podmínka splněna (což nastane pouze tehdy, když $d_-^T y > 0$), položíme $\sigma = \overline\lambda\sqrt{bb_-}/d_-^T y$. Číslo σ se tím nezvětší.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 13 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 (316)–(317), kde místo veličin ρ, b, d, y používáme veličiny s pruhem.

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 Zvolíme číslo σ_i podle poznámky 223. Určíme vektory $\bar d_i$, $\bar y_i$ podle (327) a číslo $\bar\rho_i$ podle (326).

Krok 5 Uložíme vektory $\bar d_i$, $\bar y_i$ a čísla $\bar b_i$, $\bar\rho_i$ do pracovního pole. Pokud $m = \bar m$, odstraníme vektory $\bar d_{i-m}$, $\bar y_{i-m}$ a čísla $\bar b_{i-m}$, $\bar\rho_{i-m}$ z pracovního pole. Zvětšíme i o 1 a přejdeme na krok 2.

Nyní se budeme zabývat globální konvergencí algoritmu 13. Tak jako v oddílu 4.5 budeme předpokládat, že funkce $F : R^n \to R$ vyhovuje podmínkám (F1), (F4), (F5). Pak (podobně jako v důkazu lemmatu 24) dostaneme

$$\underline G \le \frac{y^T y}{y^T d} = \frac{\|y\|^2}{b} \le \overline G, \qquad \frac{1}{\overline G} \le \frac{d^T d}{y^T d} = \frac{\|d\|^2}{b} \le \frac{1}{\underline G}. \tag{328}$$

Dále je zřejmé, že metoda s proměnnou metrikou s omezenou pamětí, realizovaná algoritmem 13, je metodou spádových směrů (s výběrem délky kroku splňujícím slabou Wolfeho podmínku).

Lemma 35 *Nechť H je pozitivně definitní matice, $u \in R^n$, $v \in R^n$, a $\vartheta > 0$. Pak matice*

$$H_+ = \tau^2\vartheta uu^T + (I - \tau uv^T)H(I - \tau vu^T) \tag{329}$$

je pozitivně definitní a platí

$$\mathrm{Tr}(H_+) \le \tau^2\vartheta\|u\|^2 + \mathrm{Tr}(H)(1 + |\tau|\|u\|\|v\|)^2, \tag{330}$$

$$\mathrm{Tr}(H_+^{-1}) \le \mathrm{Tr}(H^{-1}) + \frac{1}{\vartheta}\|v\|^2. \tag{331}$$

Důkaz (a) Z vyjádření (329) plyne, že matice H_+ je pozitivně definitní, pokud $\vartheta > 0$. Pak pro $\tau^2 > 0$ a $w^T u \ne 0$ platí $w^T H_+ w \ge \tau^2\vartheta(w^T u)^2 > 0$, 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 (329) můžeme zapsat ve tvaru

$$H_+ = H + \tau^2(\vartheta + v^T H v)uu^T - \tau(Hvu^T + uv^T H). \tag{332}$$

Použijeme-li (332), dostaneme

$$\begin{aligned}
\mathrm{Tr}(H_+) &= \tau^2(\vartheta + v^T H v)u^T u - 2\tau u^T H v \\
&\leq \tau^2(\vartheta + \mathrm{Tr}(H)\|v\|^2)\|u\|^2 + 2\,|\tau|\sqrt{u^T H u v^T H v} \\
&\leq \tau^2(\vartheta + \mathrm{Tr}(H)\|v\|^2)\|u\|^2 + 2\,|\tau|\,\mathrm{Tr}(H)\|u\|\|v\| \\
&\leq \tau^2\vartheta\|u\|^2 + \mathrm{Tr}(H)(1 + |\tau|\|u\|\|v\|)^2.
\end{aligned}$$

(c) Ukážeme, že pro dva vektory $\bar{u} \in R^n$, $\bar{v} \in R^n$ platí

$$\left(I + (\bar{u} - \bar{v})(\bar{u} - \bar{v})^T - \bar{v}\bar{v}^T\right)^{-1} = I + \frac{\bar{v}\bar{v}^T}{1 - \|\bar{v}\|^2} - \frac{(\bar{u} - \theta\bar{v})(\bar{u} - \theta\bar{v})^T}{\|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2)}, \qquad \theta = \frac{1 - \bar{u}^T\bar{v}}{1 - \|\bar{v}\|^2} \tag{333}$$

(pokud jsou oba jmenovatele nenulové). Označme $W = I - \bar{v}\bar{v}^T$ a $w = \bar{u} - \bar{v}$. Pak podle lemmatu 10 (d) platí

$$W^{-1} = I + \frac{\bar{v}\bar{v}^T}{1 - \bar{v}\bar{v}^T}, \qquad (W + ww^T)^{-1} = W^{-1} - \frac{W^{-1}ww^TW^{-1}}{1 + w^TW^{-1}w}. \tag{334}$$

První rovnost v (334) odpovídá prvnímu členu v (333). 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

$$\begin{aligned}
1 + w^TW^{-1}w &= w^T\bar{u} + 1 - \theta w^T\bar{v} = 1 + \|\bar{u}\|^2 - \bar{u}^T\bar{v} - \theta(\bar{u} - \bar{v})^T\bar{v} \\
&= \|\bar{u}\|^2 + \theta(1 - \|\bar{v}\|^2 - \bar{u}^T\bar{v} + \|\bar{v}\|^2) = \|\bar{u}\|^2 + \theta(1 - \bar{u}^T\bar{v}) = \|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2).
\end{aligned}$$

Dosadíme-li oba tyto vztahy do druhé rovnosti v (334), dostaneme druhý člen v (333).

(d) Pro $\tau = 0$ je nerovnost (331) triviální. Nechť $\tau \neq 0$ a $\omega = \tau(\vartheta + v^T H v)$ (takže $\tau\omega > 0$). Pak vzorec (329) můžeme zapsat ve tvaru

$$\begin{aligned}
H_+ &= H + \frac{\tau}{\omega}\left((\omega u - Hv)(\omega u - Hv)^T - Hvv^TH\right) \\
&= H^{1/2}\left(I + (\bar{u} - \bar{v})(\bar{u} - \bar{v})^T - \bar{v}\bar{v}^T\right)H^{1/2},
\end{aligned} \tag{335}$$

kde $\bar{u} = \sqrt{\tau\omega}H^{-1/2}u$ a $\bar{v} = \sqrt{\tau/\omega}H^{1/2}v$. Protože

$$1 - \|\bar{v}\|^2 = 1 - \frac{\tau}{\omega}v^THv = 1 - \frac{\tau}{\omega}\left(\frac{\omega}{\tau} - \vartheta\right) = \frac{\tau}{\omega}\vartheta > 0, \tag{336}$$

můžeme podle (335)–(336) psát

$$\mathrm{Tr}(H_+^{-1}) = \mathrm{Tr}(H^{-1}) + \frac{\bar{v}^TH^{-1}\bar{v}}{1 - \|\bar{v}\|^2} - \frac{(\bar{u} - \theta\bar{v})^TH^{-1}(\bar{u} - \theta\bar{v})}{\|\bar{u}\|^2 + \theta^2(1 - \|\bar{v}\|^2)} \leq \mathrm{Tr}(H^{-1}) + \frac{\bar{v}^TH^{-1}\bar{v}}{1 - \|\bar{v}\|^2} = \mathrm{Tr}(H^{-1}) + \frac{v^Tv}{\vartheta}.$$

□

Věta 120 *Uvažujme metodu s proměnnou metrikou s omezenou pamětí, realizovanou algoritmem 13, s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Nechť funkce F splňuje podmínky (F1), (F4), (F5). Pak platí*

$$\lim_{i\to\infty}\|g_i\| = 0. \tag{337}$$

Důkaz (a) Podle poznámky 223 pro $i-m \leq j \leq i-1$ platí $|\sigma_j| < 1$ a $\bar{b}_j \geq (1-\overline{\lambda})b_j$, což spolu s (326) dává $\bar{\rho}_j \leq 1/(1-\overline{\lambda})$. Použijeme-li (328), dostaneme

$$\begin{aligned}
\frac{\|\bar{y}_j\|^2}{b_j} &= \frac{1}{b_j}\left\|y_j - \sigma_j\sqrt{b_j/b_{j-1}}y_{j-1}\right\|^2 \leq 2\left(\frac{\|y_j\|^2}{b_j} + \sigma_j^2\frac{\|y_{j-1}\|^2}{b_{j-1}}\right) \leq 4\overline{G}, \\
\frac{\|\bar{d}_j\|^2}{b_j} &= \frac{1}{b_j}\left\|d_j - \sigma_j\sqrt{b_j/b_{j-1}}d_{j-1}\right\|^2 \leq 2\left(\frac{\|d_j\|^2}{b_j} + \sigma_j^2\frac{\|d_{j-1}\|^2}{b_{j-1}}\right) \leq \frac{4}{\underline{G}}
\end{aligned}$$

a podle lemmatu 35, kde $H = H_j^i$, $u = \bar{d}_j$, $v = \bar{y}_j$, $\tau = 1/\bar{b}_j$ a $\vartheta = \bar{\rho}_j\bar{b}_j = b_j/(1-\sigma_j^2)$, platí

$$\begin{aligned}
\mathrm{Tr}(H_{j+1}^i) &\leq \frac{\bar{\rho}_j}{\bar{b}_j}\|\bar{d}_j\|^2 + \mathrm{Tr}(H_j^i)\left(1 + \frac{\|\bar{d}_j\|\|\bar{y}_j\|}{\bar{b}_j}\right)^2 \\
&\leq \frac{4}{(1-\overline{\lambda})^2\underline{G}} + \mathrm{Tr}(H_j^i)\left(1 + \frac{4}{1-\overline{\lambda}}\sqrt{\frac{\overline{G}}{\underline{G}}}\right)^2 \triangleq \overline{C}_1 + \mathrm{Tr}(H_j^i)\,\overline{C}_2, \\
\mathrm{Tr}(B_{j+1}^i) &\leq \mathrm{Tr}(B_j^i) + \frac{\|\bar{y}_j\|^2}{b_j(1-\sigma_j^2)} \leq \mathrm{Tr}(B_j^i) + \frac{4\overline{G}}{1-\overline{\sigma}^2} \triangleq \mathrm{Tr}(B_j^i) + \overline{K}_1
\end{aligned}$$

(zde $B_j^i = (H_j^i)^{-1}$, $i-m \leq j \leq i$).

(b) Použijeme-li nerovnosti (328) spolu s definicí matice H_{i-m}^i (poznámka 218), dostaneme

$$\|H_{i-m}^i\| = \frac{y_{i-1}^T d_{i-1}}{y_{i-1}^T y_{i-1}} \leq \frac{1}{\underline{G}} \leq \overline{C}_1, \qquad \|B_{i-m}^i\| = \frac{y_{i-1}^T y_{i-1}}{y_{i-1}^T d_{i-1}} \leq \overline{G} \leq \overline{K}_1,$$

což spolu s (a) dává

$$\mathrm{Tr}(H_i^i) \leq \overline{C}_1(1 + \overline{C}_2 + \overline{C}_2^2 + \ldots + \overline{C}_2^{\overline{m}}) \triangleq \overline{C}, \qquad \mathrm{Tr}(B_i^i) \leq (\overline{m}+1)\overline{K}_1 \triangleq \overline{K}.$$

(c) Podle (b) můžeme psát

$$\cos^2\theta_i = \frac{(s_i^T g_i)^2}{s_i^T s_i\, g_i^T g_i} = \frac{s_i^T B_i^i s_i}{s_i^T s_i}\,\frac{s_i^T B_i^i s_i}{s_i^T (B_i^i)^2 s_i} \geq \frac{1}{\|H_i^i\|\,\|B_i^i\|} \geq \frac{1}{\mathrm{Tr}(H_i^i)\,\mathrm{Tr}(B_i^i)} \geq \frac{1}{\overline{C}\,\overline{K}},$$

takže podle poznámky 25 platí (337). □

Strangovy rekurence (316)–(317) jsou nejstarší a nejjednodušší realizací metody BFGS s omezenou pamětí. Pro některé aplikace jsou výhodnější maticové reprezentace, které nyní odvodíme. Abychom se při popisu těchto realizací vyhnuli dvojímu indexování, budeme předpokládat, že $i \leq \bar{m}$. Pak matice H_j^i, $1 \leq j \leq i$, nezávisejí na horním indexu, který můžeme vynechat. Případ, kdy $i > \bar{m}$, je vyšetřen v poznámce 226.

Lemma 36 *Nechť $N = -M^{-1}$, kde M je matice vystupující ve větě 44 s $\gamma = 1$ a $\rho = 1$. Pak platí*

$$N = \begin{bmatrix} \dfrac{\eta ab}{\eta a + (1-\eta)b} - b, & \dfrac{\eta ab}{\eta a + (1-\eta)b} \\ \dfrac{\eta ab}{\eta a + (1-\eta)b}, & \dfrac{\eta ab}{\eta a + (1-\eta)b} + a \end{bmatrix}. \tag{338}$$

Důkaz Z vyjádření matice M (věta 44) plyne

$$N = -M^{-1} = -\frac{1}{\det M} \left[\begin{array}{ll} \frac{\eta-1}{a}, & \frac{\eta}{b} \\ \frac{\eta}{b}, & \frac{1}{b}\left(\eta\frac{a}{b}+1\right) \end{array} \right].$$

Dosadíme-li za $-\det M$ vztah μ definovaný v poznámce 85 (s $\gamma = 1$ a $\rho = 1$), dostaneme po úpravě tvrzení lemmatu. □

Poznámka 224 Pro metodu DFP je $\eta = 0$, takže

$$N = \left[\begin{array}{ll} -d^T y, & 0 \\ 0, & y^T H y \end{array} \right]. \tag{339}$$

Pro metodu BFGS je $\eta = 1$, takže

$$N = \left[\begin{array}{ll} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{array} \right]. \tag{340}$$

Lemma 37 *Nechť B a $\beta - b^T B^{-1} b$ jsou čtvercové regulární matice. Pak platí*

$$[A, a] \left[\begin{array}{ll} B, & b \\ b^T, & \beta \end{array} \right]^{-1} [A, a]^T = AB^{-1}A^T + (a - AB^{-1}b)(\beta - b^T B^{-1} b)^{-1}(a - AB^{-1}b)^T.$$

Důkaz Vynásobením se snadno přesvědčíme, že platí

$$\left[\begin{array}{ll} B, & b \\ b^T, & \beta \end{array} \right]^{-1} = \left[\begin{array}{ll} B^{-1} + B^{-1}b(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & -B^{-1}b(\beta - b^T B^{-1} b)^{-1} \\ -(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & (\beta - b^T B^{-1} b)^{-1} \end{array} \right].$$

Zbytek tvrzení snadno ověříme dosazením tohoto vyjádření do výchozího vzorce a následným roznásobením. □

V dalším textu budeme předpokládat, že H_1 je symetrická pozitivně definitní matice a že pro libovolný index $1 \leq i \leq m$ platí

$$H_{i+1} = H_i - [d_i,\ H_i y_i]\, N_i^{-1}\, [d_i,\ H_i y_i]^T, \tag{341}$$

kde N_i je matice specifikující konkrétní metodu s proměnnou metrikou. Budeme se snažit nalézt vyjádření

$$H_{i+1} = H_1 - [D_i,\ H_1 Y_i]\, \bar{N}_i^{-1}\, [D_i,\ H_1 Y_i]^T, \tag{342}$$

kde $D_i = [d_1, \ldots, d_i]$, $Y_i = [y_1, \ldots, y_i]$ a kde $\bar{N}_i$ je symetrická matice řádu $2k$. Budeme přitom používat označení R_i pro horní trojúhelníkovou matici řádu i takovou, že $(R_i)_{kl} = d_k^T y_l$, $k \leq l$, a $(R_i)_{kl} = 0$, $k > l$, a C_i pro diagonální matici řádu i takovou, že $(C_i)_{kk} = d_k^T y_k$. 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 $D = D_{i-1}$, $Y = Y_{i-1}$, $R = R_{i-1}$, $C = C_{i-1}$, takže $D_i = [D, d]$, $Y_i = [Y, y]$ a

$$R_i = \left[\begin{array}{ll} R, & D^T y \\ 0, & d^T y \end{array} \right], \qquad R_i - C_i = \left[\begin{array}{ll} R - C, & D^T y \\ 0, & 0 \end{array} \right].$$

Poznamenejme, že dosadíme-li do (341) matici H_i vyjádřenou pomocí (342), dostaneme

$$
\begin{aligned}
H_+ &= H - \Big[d,\ H_1 y - [D,\ H_1 Y]\,\bar{N}^{-1}\,[D,\ H_1 Y]^T\,y\Big]\cdot \\
&\quad N^{-1}\Big[d,\ H_1 y - [D,\ H_1 Y]\,\bar{N}^{-1}\,[D,\ H_1 Y]^T\,y\Big]^T \\
&= H - \left([d,\ H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)\cdot \\
&\quad N^{-1}\left([d,\ H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^T, \qquad (343)
\end{aligned}
$$

kde $\bar{N} = \bar{N}_{i-1}$ a $N = N_i$.

Následující věty uvádějí maticové reprezentace nejznámějších metod s proměnnou metrikou s omezenou pamětí.

Věta 121 *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le i \le m$ platí (341), kde matice N_i je určená vztahem (339) (metoda DFP). Pak lze psát*

$$
H_{i+1} = H_1 - [D_i,\ H_1 Y_i]\begin{bmatrix} -C_i, & R_i - C_i \\ (R_i - C_i)^T, & Y_i^T H_1 Y_i \end{bmatrix}^{-1}[D_i,\ H_1 Y_i]^T. \qquad (344)
$$

Důkaz Pro $i = 1$ je (344) ekvivalentní s (117) (s (341) kde matice N je určena pomocí (339)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (344) platí pro všechny indexy menší než i. Pro index i můžeme (344) zapsat (po permutaci) ve tvaru

$$
H_+ = H_1 - [D,\ H_1 Y,\ d,\ H_1 y]\begin{bmatrix} -C, & R - C, & 0, & D^T y \\ (R-C)^T, & Y^T H_1 Y, & 0, & Y^T H_1 y \\ 0, & 0, & -d^T y, & 0 \\ y^T D, & y^T H_1 Y, & 0, & y^T H_1 y \end{bmatrix}^{-1}\begin{bmatrix} D^T \\ Y^T H_1 \\ d^T \\ y^T H_1 \end{bmatrix}.
$$

Použijeme-li lemma 37 a označíme-li

$$
\bar{N} = \begin{bmatrix} -C, & R - C \\ (R-C)^T, & Y^T H_1 Y \end{bmatrix},
$$

dostaneme

$$
\begin{aligned}
H_+ &= H_1 - [D,\ H_1 Y]\,\bar{N}^{-1}\,[D,\ H_1 Y]^T - \\
&\quad \left([d,\ H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)\cdot \\
&\quad \left(\begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix}\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^{-1}\cdot \\
&\quad \left([d,\ H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^T \\
&= H - \left([d, H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)\cdot \\
&\quad \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}^{-1}\left([d, H_1 y] - [D,\ H_1 Y]\,\bar{N}^{-1}\begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^T,
\end{aligned}
$$

což je právě vztah (343) s maticí N určenou pomocí (339). □

Věta 122 *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le i \le m$ platí (341), kde matice N_i je určená vztahem (340) (metoda BFGS). Pak lze psát*

$$H_{i+1} = H_1 - [D_i,\ H_1Y_i] \begin{bmatrix} 0, & R_i \\ R_i^T, & C_i + Y_i^T H_1 Y_i \end{bmatrix}^{-1} [D_i,\ H_1Y_i]^T . \tag{345}$$

Důkaz Pro $i = 1$ je (345) ekvivalentní s (118) (s (341) kde matice N je určena pomocí (340)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (345) platí pro všechny indexy menší než i. Pro index i můžeme (345) 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 37 a označíme-li

$$\bar{N} = \begin{bmatrix} 0, & R \\ R^T, & C + Y^TH_1Y \end{bmatrix},$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D,\ H_1Y]\,\bar{N}^{-1}\,[D,\ H_1Y]^T - \\ &\quad \left([d,\ H_1y] - [D, H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \left(\begin{bmatrix} 0, & d^Ty \\ d^Ty, & d^Ty + y^TH_1y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^TD, & y^TH_1Y \end{bmatrix} \bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^{-1} \cdot \\ &\quad \left([d,\ H_1y] - [D, H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T \\ &= H - \left([d, H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \begin{bmatrix} 0, & d^Ty \\ d^Ty, & d^Ty + y^THy \end{bmatrix}^{-1} \left([d, H_1y] - [D,\ H_1Y]\,\bar{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T, \end{aligned}$$

což je právě vztah (343) s maticí N určenou pomocí (340). □

Věta 123 *Nechť H_1 je symetrická pozitivně definitní matice a nechť pro libovolný index $1 \le i \le m$ platí*

$$H_{i+1} = H_i + (d_i - H_iy_i)(d_i^Ty_i - y_i^TH_iy_i)^{-1}(d_i - H_iy_i)^T \tag{346}$$

(metoda hodnosti 1). Pak lze psát

$$H_{i+1} = H_1 + (D_i - H_1Y_i)(R_i + R_i^T - C_i - Y_i^TH_1Y_i)^{-1}(D_i - H_1Y_i)^T. \tag{347}$$

Důkaz Vztah (346) je pro $i = 1$ ekvivalentní se vztahem (347). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (347) platí pro všechny indexy menší než i. Pro index i můžeme (347) zapsat ve tvaru

$$H_+ = H_1 + [D - H_1Y,\ d - H_1y] \begin{bmatrix} R + R^T - C - Y^TH_1Y, & D^Ty - Y^TH_1y \\ y^TD - y^TH_1Y, & d^Ty - y^TH_1y \end{bmatrix}^{-1} \begin{bmatrix} D^T - Y^TH_1 \\ d^T - y^TH_1 \end{bmatrix}.$$

Použijeme-li lemma 37 a označíme-li

$$\bar{N} = R + R^T - C - Y^T H_1 Y,$$

dostaneme

$$\begin{aligned} H_+ &= H_1 + (D - H_1 Y)\,\bar{N}^{-1}\,(D - H_1 Y)^T + \\ &\quad \Big((D - H_1 Y)\,\bar{N}^{-1}\,(D - H_1 Y)^T\, y - (d - H_1 y)\Big)\cdot \\ &\quad \Big(d^T y - y^T H_1 y - y^T\,(D - H_1 Y)\,\bar{N}^{-1}\,(D - H_1 Y)^T\, y\Big)^{-1}\cdot \\ &\quad \Big((D - H_1 Y)\,\bar{N}^{-1}\,(D - H_1 Y)^T\, y - (d - H_1 y)\Big)^T \\ &= H + (d - Hy)\left(d^T y - y^T H y\right)^{-1}(d - Hy)^T, \end{aligned}$$

což je právě vztah (346). □

Poznámka 225 Podobná maticová vyjádření můžeme odvodit pro matici $B = H^{-1}$. Lze k tomu použít dualitu (poznámka 94). 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

$$B_{i+1} = B_1 - [Y_i, B_1 D_i] \begin{bmatrix} 0, & L_i^T \\ L_i, & C_i + D_i^T B_1 D_i \end{bmatrix}^{-1} [Y_i, B_1 D_i]^T. \tag{348}$$

Pro metodu BFGS dostaneme

$$B_{i+1} = B_1 - [Y_i, B_1 D_i] \begin{bmatrix} -C_i, & (L_i - C_i)^T \\ L_i - C_i, & D_i^T B_1 D_i \end{bmatrix}^{-1} [Y_i, B_1 D_i]^T. \tag{349}$$

Pro metodu hodnosti 1 dostaneme

$$B_{i+1} = B_1 + (Y_i - B_1 D_i)\left(L_i + L_i^T - C_i - D_i^T B_1 D_i\right)^{-1}(Y_i - B_1 D_i)^T. \tag{350}$$

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 (345) lze po dosazení $H_1 = \gamma_i I$, kde $\gamma_i = d_i^T y_i / y_i^T y_i$, zapsat ve tvaru

$$H_{i+1} = \gamma_i I + [D_i, \gamma_i Y_i] \begin{bmatrix} (R_i^{-1})^T (C_i + \gamma_i Y_i^T Y_i) R_i^{-1}, & -(R_i^{-1})^T \\ -R_i^{-1}, & 0 \end{bmatrix} [D_i, \gamma_i Y_i]^T. \tag{351}$$

Lze se o tom přesvědčit vynásobením použité matice maticí $\bar{N}_i$ z (345) (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 trojúhelníkovou maticí R_i řádu m). Matici (349) 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

$$\overline{L}_i\overline{L}_i^T = (L_i - C_i)^T C_i^{-1}(L_i - C_i) + \frac{1}{\gamma_i} D_i^T D_i \tag{352}$$

(lze se o tom přesvědčit prostým vynásobením). Použijeme-li vzorec pro inverzi blokově trojúhelníkové matice

$$\begin{bmatrix} A & 0 \\ B & C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1}, & 0 \\ -C^{-1}BA^{-1}, & C^{-1} \end{bmatrix},$$

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

$$B_{i+1} = \frac{1}{\gamma_i} I - \left[Y_i, \frac{1}{\gamma_i} D_i\right] \begin{bmatrix} -C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix}^T \begin{bmatrix} C_i^{-\frac{1}{2}}, & 0 \\ \overline{L}_i^{-1}(L_i - C_i)C_i^{-1}, & \overline{L}_i^{-1} \end{bmatrix} \left[Y_i, \frac{1}{\gamma_i} D_i\right]^T \tag{353}$$

(opět se řeší pouze soustavy rovnic s trojúhelníkovou maticí $\overline{L}_i$ řádu m). Matice $\overline{L}_i$ se získává Choleského rozkladem matice (352). Poznamenejme, že metoda BFGS založená na maticovém vyjádření (351) není numericky efektivnější než metoda BFGS používající Strangovy rekurence. Vzorce (349) a (353) 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 (351), 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 $D_i^T Y_i$.

Poznámka 226 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 227 Metoda používající vzorec (351) 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 (351)). 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 $d_j^T y_i = d_j^T g_{i+1} - d_j^T g_i$, $y_j^T y_i = y_j^T g_{i+1} - y_j^T g_i$, $i-m+1 \le j \le i-1$, 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} podle (351), takže odpadne $2mn$ operací násobení a sčítání.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 14 Data $\bar{m} < n$, $\varepsilon > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

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 $\|g_i\| \le \varepsilon$, 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 (351) (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} podle poznámky 226 (staré sloupce ubíráme pouze tehdy, pokud $i > \bar{m}$). Zvětšíme i o 1 a přejdeme na krok 2.

Vzorec (351) 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, že $i \leq \bar{m}$). Výhodou tohoto vzorce je, že matice $\tilde{M}_i$ 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 (344), (345) a (347)). Pro obecnou metodu z Broydenovy třídy takové explicitní vyjádření neznáme, matici $\tilde{M}_i$ však lze spočítat rekurentně, přičemž počet potřebných aritmetických operací je řádově stejný jako u metod DFP, BFGS a R1.

Věta 124 *Nechť*

$$H_+ = H + UMU^T, \tag{354}$$

kde $U = [d,\ Hy]$ *a*

$$H = H_1 + \bar{U}\bar{M}\bar{U}^T. \tag{355}$$

Pak

$$H_+ = H_1 + \bar{U}_+ \bar{M}_+ \bar{U}_+^T,$$

kde $\bar{U}_+ = [\bar{U},\ d,\ H_1 y]$ *a*

$$\bar{M}_+ = \begin{bmatrix} \bar{M} + m_3\, zz^T, & m_2\, z, & m_3\, z \\ m_2\, z^T, & m_1, & m_2 \\ m_3\, z^T, & m_2, & m_3 \end{bmatrix}. \tag{356}$$

Přitom m_1, m_2, m_3 *jsou prvky matice* M *(určené podle (114)) a*

$$z = \bar{M}\bar{r}, \qquad \bar{r} = \bar{U}^T y.$$

Důkaz Dosadíme-li (355) do (354), dostaneme

$$\begin{aligned} H_+ &= H_1 + \bar{U}\,\bar{M}\,\bar{U}^T + \left[d,\ H_1 y + \bar{U}\,\bar{M}\,\bar{U}^T y\right]\ M\ \left[d,\ H_1 y + \bar{U}\,\bar{M}\,\bar{U}^T y\right]^T \\ &= H_1 + \bar{U}\,\bar{M}\,\bar{U}^T + m_1\, dd^T \\ &\quad + m_2\left(d(H_1 y)^T + H_1 y d^T\right) + m_2\left(d\,(\bar{U}z)^T + \bar{U}z\,d^T\right) \\ &\quad + m_3\, H_1 y (H_1 y)^T + m_3\left(H_1 y\,(\bar{U}z)^T + \bar{U}z\,(H_1 y)^T\right) + m_3\,\bar{U}zz^T\bar{U}^T \\ &= H_1 + \left[\bar{U},\ d,\ H_1 y\right]\ \bar{M}_+\ \left[\bar{U},\ d,\ H_1 y\right]^T, \end{aligned}$$

kde $\bar{M}_+$ je matice určená vzorcem (356). □

Poznámka 228 Číslo $a = y^T H y$ potřebné k určení hodnot m_1 a m_3 počítáme podle vzorce

$$a = y^T H y = y^T(H_1 y + \bar{U}\,\bar{M}\,\bar{U}^T y) = y^T H_1 y + \bar{r}^T z.$$

Ukážeme jak se konstruuje matice H_{i+1} v obecném případě, kdy může platit $i > \bar{m}$. Budeme předpokládat (tak jako v (351)), že $H_1 = \gamma_i I$, a používat obdélníkovou matici $\bar{V}_i = [d_{i-m+1}, y_{i-m+1}, \ldots, d_i, y_i]$ (takže $\bar{U}_i$ vznikne z $\bar{V}_i$ vynásobením každého sudého sloupce číslem γ_i) a horní horní blokově trojúhelníkovou matici

$$\bar{R}_i = \begin{bmatrix} d_{i-m+1}^T y_{i-m+1}, & \ldots & d_{i-m+1}^T y_i \\ y_{i-m+1}^T y_{i-m+1}, & \ldots & y_{i-m+1}^T y_i \\ \ldots\ldots\ldots & \ldots & \ldots\ldots\ldots \\ 0, & \ldots & d_i^T y_i \\ 0, & \ldots & y_i^T y_i \end{bmatrix},$$

jejíž každý blok obsahuje dva řádky a jeden sloupec.

Poznámka 229 Matice $\bar{V}_i$ vznikne z matice $\bar{V}_{i-1}$ přidáním dvou nových sloupců d_i, y_i, a pokud $i > \bar{m}$, ubráním dvou starých sloupců d_{i-m}, y_{i-m}. Podobně jednoduše získáme matici $\bar{R}_i$ z matice $\bar{R}_{i-1}$. Počítá se přitom pouze poslední sloupec $\bar{V}_i^T y_i$ této matice. Matici $\bar{M}_i = \bar{M}_i^i$ získáme rekurentně tak, že položíme

$$\bar{M}_{i-m+1}^i = \begin{bmatrix} m_{i-m+1}^1, & m_{i-m+1}^2 \\ m_{i-m+1}^2, & m_{i-m+1}^3 \end{bmatrix}$$

(indexy 1, 2, 3 jsou umístěny nahoře) a pro $i - m + 1 \le j \le i - 1$ vypočteme vektor $z_j = \bar{M}_j^i \bar{r}_j$, kde $\bar{r}_j$ je $j - i + m$-tý sloupec matice $\bar{R}_i$, jehož každý sudý prvek je vynásoben číslem γ_i, a položíme

$$\bar{M}_{j+1}^i = \begin{bmatrix} \bar{M}_j^i + m_{j+1}^3 z_j z_j^T, & m_{j+1}^2 z_j, & m_{j+1}^3 z_j \\ m_{j+1}^2 z_j^T, & m_{j+1}^1, & m_{j+1}^2 \\ m_{j+1}^3 z_j^T, & m_{j+1}^2, & m_{j+1}^3 \end{bmatrix}.$$

Tím máme k dispozici všechny matice potřebné k výpočtu matice H_{i+1}.

Poznámka 230 Směrový vektor s_{i+1} se určí podle vzorce

$$s_{i+1} = -H_{i+1} g_{i+1} = -\gamma_i g_{i+1} - \bar{U}_i \bar{M}_i \bar{U}_i^T g_{i+1}, \tag{357}$$

kde matice $\bar{U}_i = [d_{i-m+1}, \gamma_i y_{i-m+1}, \ldots, d_i, \gamma_i y_i]$ vznikne z matice $\bar{V}_i$ vynásobením každého sudého sloupce číslem γ_i. Je tedy vidět, že v i-tém iteračním kroku spotřebujeme zhruba $6mn$ operací násobení a sčítání a použijeme-li postup uvedený v poznámce 227, lze ušetřit zhruba $2(m-1)n$ operací.

Algoritmus 15 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 podle vzorce (357) (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 $\bar{V}_i$, $\bar{R}_i$ z matic $\bar{V}_{i-1}$, $\bar{R}_{i-1}$ podle poznámky 229 (staré sloupce ubíráme pouze tehdy, pokud $i > \bar{m}$). Určíme rekurentně matici $\bar{M}_i$ podle poznámky 229. Zvětšíme i o 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.

Věta 125 *Nechť*

$$H_+ = H + UMU^T, \qquad u_+ = \left(I - \frac{uy^T}{y^T u}\right)(d - Hy), \tag{358}$$

kde $U = [u,\ d - Hy]$ *a*

$$H = H_1 + \bar{U}\bar{M}\bar{U}^T. \tag{359}$$

Pak

$$H_+ = H_1 + \bar{U}_+ \bar{M}_+ \bar{U}_+^T, \qquad u_+ = \left(I - \frac{uy^T}{y^T u}\right)(d - H_1 y - \bar{U} z)$$

kde $\bar{U}_+ = [\bar{U},\ u,\ d - H_1 y]$ *a*

$$\bar{M}_+ = \begin{bmatrix} \bar{M} + m_3\, zz^T, & -m_2\, z, & -m_3\, z \\ -m_2\, z^T, & m_1, & m_2 \\ -m_3\, z^T, & m_2, & m_3 \end{bmatrix}. \tag{360}$$

Přitom m_1, m_2, m_3 jsou prvky matice M

$$z = \bar{M}\bar{r}, \qquad \bar{r} = \bar{U}^T y.$$

Důkaz Dosadíme-li (359) do (358), dostaneme

$$\begin{aligned} H_+ &= H_1 + \bar{U}\,\bar{M}\,\bar{U}^T + \left[u,\ d - H_1 y - \bar{U}\,\bar{M}\,\bar{U}^T y\right]\ M\ \left[u,\ d - H_1 y - \bar{U}\,\bar{M}\,\bar{U}^T y\right]^T \\ &= H_1 + \bar{U}\,\bar{M}\,\bar{U}^T + m_1\, uu^T \\ &\quad + m_2\left(u(d - H_1 y)^T + (d - H_1 y)u^T\right) - m_2\left(u\,(\bar{U}z)^T + \bar{U}z\,u^T\right) \\ &\quad + m_3\,(d - H_1 y)(d - H_1 y)^T - m_3\left((d - H_1 y)\,(\bar{U}z)^T + \bar{U}z\,(d - H_1 y)^T\right) + m_3\,\bar{U}zz^T\bar{U}^T \\ &= H_1 + \left[\bar{U},\, u,\, d - H_1 y\right]\ \bar{M}_+\ \left[\bar{U},\, u,\, d - H_1 y\right]^T, \end{aligned}$$

kde $\bar{M}_+$ je matice určená vzorcem (360). □

8.2 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 jednotkovou matici pouze v prvním iteračním kroku. Směrový vektor s_i se počítá pomocí redukované matice $\tilde{H}_i$ řádu m, která se aktualizuje, a pomocí m vektorů z_i, $1 \le i \le m$, 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 126 *Uvažujme metodu s proměnnou metrikou (102)–(104), kde $H_1 = I$. Pak*

$$H_i z \in \mathcal{G}_i, \quad \forall z \in \mathcal{G}_i, \tag{361}$$

$$H_i w = \left(\prod_{k=1}^{i-1} \gamma_k\right) w, \quad \forall w \in \mathcal{G}_i^{\perp}, \tag{362}$$

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 $Hz = z \in \mathcal{G}_1$, $\forall z \in \mathcal{G}_1$, a $Hw = w$, $\forall w \in \mathcal{G}_1^{\perp}$. Předpokládejme, že (361)–(362) 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 $w \in \mathcal{G}_{i+1}^{\perp} \subset \mathcal{G}_i^{\perp}$. Pak podle (362) 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 (103) 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 (104) 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 (103) 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 18 *Nechť jsou splněny předpoklady věty 126 a nechť matice H_i, $i \in N$, jsou pozitivně definitní. Pak $\mathcal{S}_i = \mathcal{G}_i$, $i \in N$, kde $\mathcal{S}_i = \mathcal{L}(s_1, \ldots, s_i)$.*

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 126 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 (218) 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 127 *Nechť jsou splněny předpoklady věty 126 a nechť Z_i je matice, jejíž sloupce tvoří ortonormální bázi v $\mathcal{G}_i$. Pak platí*

$$H_i = Z_i \tilde{H}_i Z_i^T + \left(\prod_{k=1}^{i-1} \gamma_k \right) (I - Z_i Z_i^T), \qquad \tilde{H}_i = Z_i^T H_i Z_i. \tag{363}$$

Důkaz Podle věty 126 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 126 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 (363). □

Poznámka 231 Jelikož $g_i \in \mathcal{G}_i$, můžeme podle (363) 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

$$\tilde{g}_i = Z_i^T g_i, \qquad \tilde{s}_i = -\tilde{H}_i \tilde{g}_i, \qquad s_i = Z_i \tilde{s}_i.$$

Poznamenejme, že v těchto vzorcích se používá pouze redukovaná matice $\tilde{H}_i$ a matice Z_i jejíž sloupce tvoří bázi v $\mathcal{G}_i$.

Poznámka 232 Matice $Z_{i+1} = [Z_i, z_{i+1}]$ vznikne z matice Z_i přidáním sloupce z_{i+1}, který se získá ortogonalizací vektoru g_{i+1}. Nechť $P_i = I - Z_i Z_i^T$. Předpokládejme, že vektor g_{i+1} neleží v $\mathcal{G}_i$ (takže $P_i g_{i+1} \neq 0$). Pak můžeme položit $z_{i+1} = P_i g_{i+1} / \|P_i g_{i+1}\|$ a

$$\hat{H}_i = \begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \prod_{k=1}^{i-1} \gamma_k \end{bmatrix}.$$

Věta 128 *Nechť jsou splněny předpoklady věty 126 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*

$$H_i = Z_{i+1} \hat{H}_i Z_{i+1}^T + \left(\prod_{k=1}^{i-1} \gamma_k \right) (I - Z_{i+1} Z_{i+1}^T), \tag{364}$$

takže

$$\hat{g}_i = Z_{i+1}^T g_i, \qquad \hat{s}_i = -\hat{H}_i \hat{g}_i, \qquad s_i = Z_{i+1} \hat{s}_i.$$

Nechť

$$\tilde{H}_{i+1} = \gamma_i(\hat{H}_i + \hat{U}_i M_i \hat{U}_i^T), \qquad \hat{U}_i = [\hat{d}_i, \hat{H}_i\hat{y}_i], \tag{365}$$

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 (103)). 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 (363) 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 127 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

$$\hat{g}_i = Z_{i+1}^T g_i = \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix}, \quad \hat{s}_i = -\hat{H}_i \hat{g}_i = -\begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \omega_{i-1} \end{bmatrix} \begin{bmatrix} \tilde{g}_i \\ 0 \end{bmatrix} = \begin{bmatrix} \tilde{s}_i \\ 0 \end{bmatrix}, \quad \hat{d}_i = \alpha_i \hat{s}_i = \begin{bmatrix} \tilde{d}_i \\ 0 \end{bmatrix}. \tag{366}$$

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

$$\hat{y} = Z_{i+1}^T y_i = \begin{bmatrix} Z_i^T y_i \\ z_{i+1}^T y_i \end{bmatrix} = \begin{bmatrix} \tilde{y}_i \\ \|P_i g_{i+1}\| \end{bmatrix}. \tag{367}$$

Číslo $\|P_i g_{i+1}\|$ známe, neboť ho potřebujeme k určení vektoru z_{i+1} (poznámka 232).

(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 (364) $H_i y_i = Z_{i+1}\hat{H}_i\hat{y}_i$ 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 (363) a (365), 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 (103). □

Poznámka 233 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 (365) použít vektory $\hat{d}_i = \tilde{d}_i = \alpha_i \tilde{s}_i$ a $\hat{y}_i = \tilde{y}_i = Z_i^T y_i$. Pak $Z_{i+1}\hat{U}_i = Z_i\tilde{U}_i = U_i$, takže (365) je ekvivalentní s (103).

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 16 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 232, vektor $\hat{d}_i$ podle (366) a vektor $\hat{y}_i$ podle (367).

Krok 5 Zvolíme parametry ρ_i, γ_i a η_i (tak jako v Algoritmu 4) a určíme matici $\tilde{H}_{i+1}$ podle (365).

Krok 6 Zvětšíme i o 1 a přejdeme na krok 2.

Algoritmus 16 je teoreticky shodný s algoritmem 4 a měl by být správě uveden v oddílu 4.8 jako modifikace algoritmu 4. Výhoda algoritmu 16 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 4. Hlavní přínos metod redukovaných Hessiánů spočívá v tom, že lze jejich myšlenku použít pro rozsáhlé úlohy, omezíme-li řád redukovaných matic na $\bar{m} \ll n$.

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

$$H_i = Z_i \tilde{H}_i Z_i^T + \hat{\gamma}_i (I - Z_i Z_i^T), \qquad \tilde{H}_i = Z_i^T H_i Z_i, \tag{368}$$

kde Z_i je matice, jejíž sloupce tvoří bázi v $\mathcal{S}'_i$, přičemž $S'_i = Z_i R'_i$, kde R'_i je nějaká horní trojúhelníková matice ($\hat{\gamma}_i$ je hodnota specifikovaná v poznámce 237). Metody redukovaných Hessiánů s omezenou pamětí používají směrový vektor $s_i = -H_i g_i$, kde H_i je matice určená vztahem (368). 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 231. Stačí tedy uchovávat pouze redukovanou matici $\tilde{H}_i$ a matici Z_i, jejíž sloupce tvoří bázi v $\mathcal{S}'_i$.

Poznámka 234 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 sloupce v matici Z_{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 16 liší pouze tím, že matice Z_i a $S'_i = Z_i R'_i$ mohou mít nanejvýš $\bar{m}$ sloupců. Má-li matice Z_{i+1} (získaná v kroku 5 algoritmu 16) více než $\bar{m}$ sloupců, je třeba ji upravit tak, aby měla $\bar{m}$ sloupců a aby platilo $S'_{i+1} = Z_{i+1} R'_{i+1}$, kde R'_{i+1} je nějaká horní trojúhelníková matice řádu $\bar{m}$. Tato procedura vnáší do algoritmu metod redukovaných Hessiánů s omezenou pamětí nové problémy technického charakteru, které je třeba vyřešit. Předně je třeba ukládat horní trojúhelníkovou matici R'_i řádu nejvýše $\bar{m}$ takovou, že $S'_i = Z_i R'_i$. Tuto matici je třeba upravovat společně s maticí Z_i.

Lemma 38 *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 39 *Nechť $S_i = Z_i R_i$, kde Z_i, R_i jsou matice vystupující v lemmatu 38. 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 Zřejmě $Z_i R_i = S_i$ a $Z_i Z_i^T g_{i+1} + z_{i+1}\|P_i g_{i+1}\| = Z_i Z_i^T g_{i+1} + (I - Z_i Z_i^T) g_{i+1} = g_{i+1}$. □

Lemma 40 *Nechť matice Z_{i+1}^+, $R_{i+1}^+ = [t_{i+1}, T_{i+1}]$, vystupující v lemmatu 39, 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*

$$Q_{i+1}^T T_{i+1} = \begin{bmatrix} R'_{i+1} \\ 0 \end{bmatrix}, \tag{369}$$

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 39 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 (369). 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 235 Ortogonální matice Q_{i+1} je obvykle součinem Givensových matic elementárních rotací (poznámka 210). 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 208.

Poznámka 236 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 40. 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 237 Snížením dimenze redukované matice (poznámka 236) 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

$$\hat{H}_i = \begin{bmatrix} \tilde{H}_i, & 0 \\ 0, & \hat{\gamma}_i \end{bmatrix}, \tag{370}$$

kde buď $\hat{\gamma}_i = 1$ nebo $\hat{\gamma}_i = \gamma_i$. Pokud $\hat{\gamma}_i = 1$, použijeme v aktualizaci

$$\tilde{H}_{i+1}^+ = \gamma_i (\hat{H}_i + \hat{U}_i M_i \hat{U}_i^T), \qquad \hat{U}_i = [\hat{d}_i, \hat{H}_i \hat{y}_i], \tag{371}$$

škálovací koeficient γ_i. Pokud $\hat{\gamma}_i = \gamma_i$, položíme v (371) $\gamma_i = 1$.

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 17 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 $\|g_i\| \leq \varepsilon$, 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 38.

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 $Z_{i+1}^+ := Z_i$, $R_{i+1}^+ := R_i$, $\hat{H}_i := \tilde{H}_i$, $\hat{d}_i := \tilde{d}_i$, $\hat{y}_i := \tilde{y}_i$. Pokud $\|P_i g_{i+1}\| > \underline{\delta}\|g_{i+1}\|$, určíme matice Z_{i+1}^+, R_{i+1}^+ tak jako v lemmatu 39, matici $\hat{H}_i$ podle (370) a vektory $\hat{d}_i$, $\hat{y}_i$ podle (366), (367).

Krok 5 Zvolíme parametry ρ_i, γ_i a η_i a určíme matici $\tilde{H}_{i+1}^+$ podle (371).

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 40 a matici $\tilde{H}_{i+1}$ tak jako v poznámce 236.

Krok 7 Zvětšíme i o 1 a přejdeme na krok 2.

Abychom se v algoritmu 17 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 3 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 17 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 41 *Nechť H_i je matice určená vztahem (368) a*

$$B_i = Z_i \tilde{B}_i Z_i^T + \frac{1}{\hat{\gamma}_i}(I - Z_i Z_i^T), \qquad \tilde{B}_i = Z_i^T B_i Z_i. \tag{372}$$

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 238 Ukázali jsme, že pokud $H_i = B_i^{-1}$, platí $\tilde{H}_i = \tilde{B}_i^{-1}$ a (368) 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

$$\tilde{g}_i = Z_i^T g_i, \qquad \tilde{B}_i \tilde{s}_i = -\tilde{g}_i, \qquad s_i = Z_i \tilde{s}_i.$$

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 239 Používáme-li Choleského rozklad matice $\tilde{B}_i$, zkomplikuje se krok 6 algoritmu 17. 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 17 lze modifikovat tak, abychom ušetřili operace potřebné k úpravám matice Z_{i+1}^+ (jde o ortogonální transformace použité v lemmatu 40). 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 17 vypadá takto.

(a) Směrový vektor v kroku 3 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.

8.3 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 8.2). 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 (351)), 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 (368)), 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 odílu se budeme zabývat metodami, které používají směrový vektor $s_i = -H_i g_i$, kde

$$H_i = \zeta_i I + \bar{H}_i = \zeta_i I + \bar{S}_i \bar{S}_i^T, \tag{373}$$

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 $H_{i+1} y_i = \rho_i d_i$, neboli

$$\bar{H}_{i+1} y_i = \rho_i \bar{d}_i, \qquad \bar{d}_i = d_i - \zeta_{i+1} y_i \tag{374}$$

(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

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T - \frac{1}{\bar{a}} \bar{H}y(\bar{H}y)^T + \frac{\eta}{\bar{a}} \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right) \left(\frac{\bar{a}}{\bar{b}} \bar{d} - \bar{H}y \right)^T, \tag{375}$$

kde $\bar{a} = y^T \bar{H} y$ a $\bar{b} = y^T \bar{d}$, která vyhovuje kvazinewtonovské podmínce (374). Jelikož v prvních n iteračních krocích je matice $\bar{H}$ singulární, může nastat případ, že $\bar{H}y = 0$. V tomto případě vynecháme v (375) všechny členy obsahující $\bar{H}y$, takže

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}} \bar{d}\bar{d}^T. \tag{376}$$

Věta 129 *Nechť matice $\bar{H}$ je pozitivně semidefinitní, $\eta \geq 0$ a $\zeta_+ < b/y^T y$. Pak matice $\bar{H}_+$ určená vztahem (375) je pozitivně semidefinitní.*

Důkaz Pokud $\eta \geq 0$, můžeme použít pseudosoučinový vzorec (121), kam dosadíme hodnoty s pruhem. Podle tohoto vzorce je matice $\bar{H}_+$ pozitivně semidefinitní pokud $\bar{b} > 0$ (předpokládáme, že $\rho/\gamma > 0$). Stačí tedy ověřit podmínku $\bar{b} = b - \zeta_+ y^T y > 0$, což dává $\zeta_+ < b/y^T y$. Hodnotu $\bar{b} = 0$ vylučujeme, neboť $\bar{b}$ se vyskytuje ve jmenovateli pseudosoučinového vzorce. □

U posunutých metod s proměnnou metrikou velmi záleží na hodnotě parametru ζ_{i+1}. Abychom vyhověli předpokladům věty 129 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 130 *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 $y^T\bar{d} = y^T(d - \sigma(y^T d/y^T y)y) = (1-\sigma)y^T d$ a

$$\bar{V}^T v = v - \frac{y^T v}{y^T\bar{d}}\bar{d} = v - \frac{y^T d}{y^T\bar{d}}\bar{d} = v - \frac{1}{1-\sigma}\bar{d} = v - d - \frac{\sigma}{1-\sigma}\bar{d}.$$

Protože $\|\bar{d}\| = \|d - \sigma(y^T d/y^T y)y\| = \|d\|\sqrt{1-\tau}$, můžeme psát $|\,\|\bar{V}^T v\| - \omega\|d\|\,| \leq \|v-d\|$, takže

$$\left| \frac{\|\bar{V}^T v\|}{\|v\|} - \omega \right| \leq \left| \frac{\|\bar{V}^T v\|}{\|v\|} - \omega\frac{\|d\|}{\|v\|} \right| + \omega\frac{\|v-d\|}{\|v\|} \leq (1+\omega)\frac{\|v-d\|}{\|v\|}.$$

□

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 $y^T v = y^T d$ a číslo $\|v-d\|/\|v\|$ je malé. Pak podle věty 130 platí $\|\bar{V}^T v\|/\|v\| \approx \omega$. Jelikož vektor $\bar{V}^T v$ je podle (385) prvním sloupcem matice $\bar{S}_+$, je vhodné volit číslo σ tak, aby platilo $\omega \leq 1$, což dává $\sigma \leq 1/(1+\sqrt{1-\tau})$. Většinou je nutné tuto hodnotu ještě zmenšit. Ukázalo se, že je vhodné vynásobit ji číslem $\sqrt{1-\bar{a}/a}$, takže

$$\sigma = \frac{\sqrt{1-\bar{a}/a}}{1+\sqrt{1-b^2/(y^T y d^T d)}}. \tag{377}$$

Volba (377) vyhovuje předpokladům věty 134 a odůvodňuje ji také následující věta.

Věta 131 *Nechť $\bar{H}_+$ je matice určená podle vzorce (375), kde $\bar{H} = 0$. Pak matice $H_+ = \zeta_+ I + \bar{H}_+$, kde $\zeta_+ = \sigma\bar{b}/y^T y$ a číslo σ je určeno podle (377), je optimálně podmíněná.*

Důkaz Pokud $\bar{H} = 0$, platí $\bar{H}y = 0$ a $\bar{a} = 0$ a použitím (376) dostaneme $H_+ = \zeta_+ I + (\rho/\bar{b})\bar{d}\bar{d}^T$, takže matice $(1/\zeta_+)H_+$ má podle lemmatu 10 $n-1$ 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 $\zeta_+ = \sigma b/y^T y$, $\bar{b} = (1-\sigma)b$ a $\bar{d} = d - \sigma b y/y^T y$, dostaneme pro číslo podmíněnosti matice $(1/\zeta_+)H_+$ (a tedy i H_+) rovnost

$$\begin{aligned}\frac{\kappa_+ - 1}{\rho} &= \frac{1}{\zeta_+\bar{b}}\bar{d}^T\bar{d} = \frac{y^T y}{\sigma(1-\sigma)b^2}\left(d^T d - 2\sigma\frac{b^2}{y^T y} + \sigma^2\frac{b^2}{y^T y}\right) \\ &= \frac{1}{\sigma(1-\sigma)}\left(\frac{y^T y d^T d}{b^2} - 2\sigma + \sigma^2\right) = \frac{1}{1-\sigma}\left(\frac{1}{\sigma\tau} - 1\right) - 1,\end{aligned}$$

takže

$$\left(\frac{\kappa_+}{\rho}+1\right)' = \frac{1}{(1-\sigma)^2}\left(\frac{1}{\sigma\tau}-1\right) - \frac{1}{1-\sigma}\left(\frac{1}{\sigma^2\tau}\right) = -\frac{\sigma^2\tau-2\sigma+1}{\sigma^2(1-\sigma)^2\tau}.$$

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 (377). □

V kvazinewtonovské podmínce (374) používáme parametr $\rho > 0$. Tento parametr má poněkud jiný význam, než v případě standardních metod s proměnnou metrikou. Jeho význam ukazuje následující věta.

Věta 132 *Nechť $H_+ = \zeta_+ I + \bar{H}_+$, kde $\zeta_+ = \sigma y^T d/y^T y$ a $\bar{H}_+$ je matice získaná aktualizací (375) (kde $\bar{d} = d - \zeta_+ y$). Pak jestliže $\rho = \sigma/(1-\sigma)$, platí*

$$\frac{y^T \bar{H}_+ y}{y^T y} = \zeta_+.$$

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 $\rho = \sigma/(\sigma-1)$, 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 240 Z věty 132 plyne, že pokud $\rho = \sigma/(\sigma-1)$, jsou oba členy v (373) 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 podmínkám (F1), (F4), (F5). Pak (podobně jako v důkazu lemmatu 24) dostaneme

$$\underline{G} \le \frac{y^T y}{y^T d} \le \overline{G}, \qquad \frac{\underline{\sigma}}{\overline{G}} \le \frac{\sigma}{\overline{G}} \le \zeta_+ \le \frac{\sigma}{\underline{G}} \le \frac{\overline{\sigma}}{\underline{G}} \tag{378}$$

(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 vyšetříme případ kdy $0 \le \eta < 1$.

Lemma 42 *Uvažujme posunutou metodu s proměnnou metrikou s aktualizací (375), 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 podmínky (F1), (F4), (F5). Pak, existuje-li konstanta $\overline{C} > 0$ taková, že*

$$\mathrm{Tr}\bar{H}_{i+1} \le \mathrm{Tr}\bar{H}_i + \overline{C} \qquad \forall i \in M, \tag{379}$$

platí

$$\liminf_{i\to\infty} \|g_i\| = 0. \tag{380}$$

Důkaz Platí-li (379), můžeme s použitím (378) psát

$$\|H_{i+1}\| \le \zeta_{i+1} + \|\bar{H}_{i+1}\| \le \frac{\overline{\sigma}}{\underline{G}} + \mathrm{Tr}\bar{H}_i + \overline{C} \le \frac{\overline{\sigma}}{\underline{G}} + \mathrm{Tr}\bar{H}_1 + \overline{C}i \le \overline{C}(i+1)$$

(předpokládáme bez újmy na obecnosti, že konstanta $\overline{C}$ byla vybrána tak, že $\overline{\sigma}/\underline{G} + \mathrm{Tr}\bar{H}_1 \le \overline{C}$). 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 9 plyne (380). □

Věta 133 *Nechť jsou splněny předpoklady lemmatu 42. Pak, existuje-li konstanta $C > 0$ taková že $\eta \leq \bar{b}C/(\bar{a} + \bar{b}C)$, platí (380).*

Důkaz Jestliže $\eta < 1$, můžeme aktualizaci (375) zapsat ve tvaru (124), kde vystupují veličiny s pruhem a kde

$$\bar{\mu} = \frac{1}{\bar{a}\bar{b}} \left(\eta \frac{\bar{a}}{\bar{b}} + (1-\eta)\rho \right).$$

Jelikož pro libovolnou symetrickou matici $\bar{H}$ a pro libovolné vektory u, v platí $\mathrm{Tr}(\bar{H} + uu^T - vv^T) = \mathrm{Tr}\bar{H} + u^T u - v^T v \leq \mathrm{Tr}\bar{H} + u^T u$, dostaneme použitím (124) nerovnost

$$\mathrm{Tr}\bar{H}_+ \leq \mathrm{Tr}\bar{H} + \frac{\bar{\mu}\bar{a}\bar{b}}{1-\eta} \frac{\bar{d}\bar{d}}{\bar{y}^T\bar{d}} \leq \mathrm{Tr}\bar{H} + \frac{\bar{\mu}\bar{a}\bar{b}}{(1-\eta)\underline{G}} = \mathrm{Tr}\bar{H} + \left(\frac{\eta}{1-\eta} \frac{\bar{a}}{\bar{b}} + \rho \right) \frac{1}{\underline{G}}.$$

Jelikož funkce $\eta/(1-\eta)$ je pro $0 \leq \eta < 1$ rostoucí, plyne z $\eta \leq \bar{b}C/(\bar{a} + \bar{b}C)$ nerovnost $\eta/(1-\eta) \leq (\bar{b}/\bar{a})C$, neboli $\mathrm{Tr}\bar{H}_+ \leq \mathrm{Tr}\bar{H} + (C+\rho)/\underline{G}$. Jsou tedy splněny předpoklady lemmatu 42 s $\overline{C} = (C+\rho)/\underline{G}$, takže platí (380). □

Nyní vyšetříme případ, kdy $0 \leq \eta \leq 1$ a $\underline{\sigma} \leq \sqrt{1 - \bar{a}/a}$.

Lemma 43 *Uvažujme posunutou metodu s proměnnou metrikou s aktualizací (375), kde $0 \leq \eta \leq 1$. Pak platí*

$$\frac{\det H_+}{\det H} \leq \frac{\bar{d}^T B \bar{d}}{\bar{b}} \left(\rho + \zeta \frac{y^T y}{\bar{b}} \right) \left(1 + \frac{\zeta_+}{\zeta} \right)^n. \tag{381}$$

Důkaz (a) Z vyjádření (123) a z důsledku 6 (c) plyne, že nerovnost (381) stačí dokázat pro $\eta = 1$, tedy pro metodu BFGS. Aktualizaci metody BFGS lze zapsat ve tvaru (126) (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 7 (c), 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 (373) platí $\bar{H}y = Hy - \zeta y$, můžeme psát $\bar{d}^T B \bar{H} y = \bar{b} - \zeta \bar{d}^T B y$ a $y^T \bar{H} B \bar{H} y = \bar{a} - \zeta y^T y + \zeta^2 y^T B y$. Dosadíme-li tyto hodnoty do předchozí rovnosti a použijeme-li Schwarzovu nerovnost, dostaneme

$$\begin{aligned} \frac{\det(\zeta I + \bar{H}_+)}{\det H} &= \left(\rho + \frac{\bar{a}}{\bar{b}} \right) \frac{\bar{d}^T B \bar{d}}{\bar{b}} + \zeta^2 \frac{(\bar{d} B y)^2}{\bar{b}^2} - \frac{\bar{d}^T B \bar{d} y^T \bar{H} B \bar{H} y}{\bar{b}^2} \\ &= \left(\rho + \zeta \frac{y^T y}{\bar{b}} \right) \frac{\bar{d}^T B \bar{d}}{\bar{b}} + \zeta^2 \frac{(\bar{d}^T B y)^2 - \bar{d}^T B \bar{d} y^T B y}{\bar{b}^2} \leq \left(\rho + \zeta \frac{y^T y}{\bar{b}} \right) \frac{\bar{d}^T B \bar{d}}{\bar{b}}. \end{aligned}$$

(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. □

Lemma 44 *Uvažujme posunutou metodu s proměnnou metrikou s aktualizací (375), kde* $0 \le \eta \le 1$, $0 < \underline{\rho} \le \rho \le \overline{\rho}$ *a* $0 < \max(\underline{\sigma}, \sqrt{1-\bar{a}/a}) \le \sigma \le \min(\overline{\sigma}, \sqrt{1-\bar{a}/a}) < 1$ *(takže* $\sigma = \underline{\sigma}$ *pokud* $\sqrt{1-\bar{a}/a} \le \underline{\sigma}$*), s výběrem délky kroku splňujícím slabou Wolfeho podmínku. Nechť funkce* F *splňuje podmínky (F1), (F4), (F5). Pak existují konstanty* C_1, C_2 *takové, že*

$$\frac{\det H_+}{\det H} \le C_1 \frac{c}{b} + C_2 \underline{\sigma},$$

přičemž konstanta C_2 *nezávisí na* $\underline{\sigma}$.

Důkaz (a) Předpokládejme nejprve, že $\sigma \le \sqrt{1-\bar{a}/a}$. Pak, jelikož platí $ac - b^2 \ge 0$ (Schvarzova nerovnost), dostaneme $\sigma^2 \le 1 - \bar{a}/a = \zeta y^T y/a \le \zeta y^T y c/b^2$, 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 $\zeta_+^2 y^T B y \le \zeta c y^T B y/y^T y \le c$. 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

$$\bar{d}^T B \bar{d} = (d - \zeta_+ y)^T B (d - \zeta_+ y) \le (\sqrt{d^T B d} + \zeta_+ \sqrt{y^T B y})^2 \le 2(d^T B d + \zeta_+^2 y^T B y), \tag{382}$$

což spolu s $d^T B d = c$ a $\zeta_+^2 y^T B y \le c$ dává $\bar{d}^T B \bar{d} \le 4c$. Zbylé činitele ve výrazu (381) již odhadneme snadno. Podle (378) platí

$$\rho + \zeta \frac{y^T y}{\bar{b}} \le \overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}} \frac{\overline{G}}{\underline{G}} \tag{383}$$

(neboť $\bar{b} = (1-\sigma)b \ge (1-\overline{\sigma})b$) a z (374) plyne $1 + \zeta_+/\zeta \le 1 + \overline{\sigma}\,\overline{G}/(\underline{\sigma}\,\underline{G})$. Po dosazení dostaneme

$$\frac{\det H_+}{\det H} \le \frac{4}{1-\overline{\sigma}} \left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}} \frac{\overline{G}}{\underline{G}} \right) \left(1 + \frac{\overline{\sigma}\,\overline{G}}{\underline{\sigma}\,\underline{G}} \right)^n \frac{c}{b}.$$

(b) Nechť nyní $\sigma = \underline{\sigma} > \sqrt{1-\bar{a}/a}$. Pak podle (378) platí

$$\frac{\zeta_+}{\zeta} \le \frac{\sigma}{\underline{G}} \frac{\overline{G}}{\underline{\sigma}} = \frac{\overline{G}}{\underline{G}}, \tag{384}$$

což s použitím (382) dává

$$\bar{d}^T B \bar{d} \le 2(d^T B d + \zeta_+^2 y^T B y) \le 2(c + \frac{\zeta_+^2}{\zeta} y^T y) = 2c + 2\sigma b \frac{\overline{G}}{\underline{G}} = 2c + 2\underline{\sigma} b \frac{\overline{G}}{\underline{G}}$$

(neboť $y^T B y/y^T y \le \|B\| \le 1/\zeta$) a použijeme-li odhady (383), (384), můžeme psát

$$\frac{\det H_+}{\det H} \le \frac{2}{1-\overline{\sigma}} \left(\overline{\rho} + \frac{\overline{\sigma}}{1-\overline{\sigma}} \frac{\overline{G}}{\underline{G}} \right) \left(1 + \frac{\overline{G}}{\underline{G}} \right)^n \left(\frac{c}{b} + \underline{\sigma} \frac{\overline{G}}{\underline{G}} \right).$$

(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 134 *Nechť jsou splněny předpoklady lemmatu 44. Pak je-li číslo* $\underline{\sigma} > 0$ *dostatečně malé, platí (380).*

Důkaz (a) Abychom nemuseli vyšetřovat první iterační krok zvlášť, začneme od druhého kroku. Protože v každém iteračním kroku platí $\det H \geq \zeta^n \geq (\underline{\sigma}/\overline{G})^n$, můžeme pro $k \in N$ psát

$$C \triangleq \frac{\underline{\sigma}^n}{\overline{G}^n \det H_2} \leq \frac{\det H_{k+2}}{\det H_2} = \prod_{i=2}^{k+1} \frac{\det H_{i+1}}{\det H_i} \leq \left(\frac{1}{k} \sum_{i=2}^{k+1} \frac{\det H_{i+1}}{\det H_i} \right)^k$$

(používáme nerovnost (8)). Podle lemmatu 44 tedy platí

$$kC^{1/k} \leq \sum_{i=2}^{k+1} \frac{\det H_{i+1}}{\det H_i} \leq C_1 \sum_{i=2}^{k+1} \frac{c_i}{b_i} + kC_2\underline{\sigma}.$$

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 (9) a (378), 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 9 platí (380). □

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 (373) 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

$$\bar{S}_{i+1} = \left[\bar{V}_i^T \bar{S}_i, \sqrt{\frac{\rho_i}{\bar{b}_i}} \bar{d}_i \right]. \tag{385}$$

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 (121) (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 (122), 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

$$\bar{S}_+^T y = \tilde{z}, \qquad \bar{S}_+ \tilde{z} = \rho\bar{d}, \qquad \tilde{z}^T \tilde{z} = \rho\bar{b}, \tag{386}$$

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 58, dostaneme

$$\bar{S}_+ = \bar{S} - \frac{Ty}{y^T T y}\tilde{y}^T + \left(\rho\bar{d} - \bar{z} + \frac{y^T\bar{z}}{y^T T y}Ty\right)\frac{\tilde{z}^T}{\tilde{z}^T\tilde{z}}, \tag{387}$$

kde $\tilde{y} = \bar{S}^T y$ a $\bar{z} = \bar{S}\tilde{z}$ (viz (179)). Zvolíme-li matici T tak, aby platilo $Ty = \rho\bar{d} - \bar{z}$, výraz (387) se velmi zjednoduší. Po dosazení a úpravě dostaneme

$$\bar{S}_+ = \bar{S} - \frac{\rho\bar{d} - \bar{z}}{\rho - y^T\bar{z}}\left(\tilde{y} - \tilde{z}\right)^T. \tag{388}$$

Použití vzorců (387) a (388) 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 (375) (poznámka 244). Nemůžeme ani položit $\tilde{d} = \bar{S}^T\bar{B}\bar{d}$, kde $\bar{B} = \hat{H}^\dagger$ (což by odpovídalo volbě (155)), 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 241 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 (388), dostaneme

$$\bar{S}_+ = \bar{S} - \frac{\rho\bar{d} - \vartheta\bar{H}Bd}{\rho\bar{b} - \vartheta y^T\bar{H}Bd}\left(y - \vartheta Bd\right)^T\bar{S}. \tag{389}$$

Znaménko koeficientu ϑ je vhodné volit tak, aby jmenovatel v (389) byl co největší, tedy tak, aby platilo $\vartheta y^T\bar{H}Bd \le 0$.

Poznámka 242 Podle věty 59 dostaneme standardní metodu BFGS, dosadíme-li $Ty = d$ a $\tilde{z} = \vartheta S^T B d$ do 179 (ϑ se volí tak, aby byla splněna poslední rovnost v (178)). Analogicky můžeme dosadit $Ty = \bar{d}$ a $\tilde{z} = \vartheta\bar{S}^T B d$ do (387). V tomto případě platí

$$\bar{S}_+ = \bar{S} - \frac{1}{\bar{b}}\bar{d}y^T\bar{S} + \frac{1}{\bar{c}}\left[\left(\frac{\rho}{\vartheta} + \frac{y^T\bar{H}Bd}{\bar{b}}\right)\bar{d} - \bar{H}Bd\right]d^T B\bar{S} \tag{390}$$

(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 $\vartheta y^T\bar{H}Bd \le 0$.

Poznámka 243 Vztahy (389) a (390) definují jednoduché metody, které jsou v jistém smyslu zobecněním metody BFGS. Obecnější metody dostaneme, volíme-li

$$Ty = \frac{\sqrt{\eta}}{\bar{b}}\bar{d} + \frac{1-\sqrt{\eta}}{\bar{a}}\bar{H}y \qquad \tilde{z} = \vartheta\bar{S}^T\left(\frac{\sqrt{\eta}}{\bar{b}}Bd + \frac{1-\sqrt{\eta}}{\bar{a}}y\right), \tag{391}$$

kde parametr ϑ se vybírá tak, aby platilo $\tilde{z}^T\tilde{z} = \rho\bar{b}$. Položíme-li

$$\hat{a} = y^T\bar{H}y, \qquad \hat{b} = y^T\bar{H}Bd, \qquad \hat{c} = d^T B\bar{H}Bd$$

(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

$$\vartheta^2 = \frac{\rho\bar{a}\bar{b}^3}{\eta\left(\hat{a}\hat{c} - \hat{b}^2\right) + \left(\bar{b} + (\hat{b} - \bar{b})\sqrt{\eta}\right)^2}. \tag{392}$$

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 (391)–(392) dosazujeme do obecného výrazu (387) (platí $y^T T y = 1$ a $\tilde{z}^T\tilde{z} = \rho\bar{b}$).

Dosavadní úvahy můžeme shrnout ve formě algoritmu, který lze popsat zhruba takto:

Algoritmus 18 Data $\bar{m} < n$, $\underline{\varepsilon} > 0$, $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$.

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 (377). 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 (385). Pokud $i > \bar{m}$, určíme matici $\bar{S}_{i+1}$ podle vzorce (390 nebo podle vzorce (387) s volbou (391)–(392). Zvětšíme i o 1 a přejdeme na krok 2.

Nyní vyšetříme globální konvergenci metody s proměnnou metrikou s omezenou pamětí realizované algoritmem 18.

Lemma 45 *Uvažujme posunutou metodu s proměnnou metrikou s omezenou pamětí s aktualizací (387) a volbou (391)–(392). Pak platí*

$$\bar{H}_+ = \bar{H} + \frac{\rho}{\bar{b}}\bar{d}\bar{d}^T - \frac{1}{\bar{a}}\bar{H}y(\bar{H}y)^T + \frac{\eta}{\bar{a}}\left(\frac{\bar{a}}{\bar{b}}\bar{d} - \bar{H}y\right)\left(\frac{\bar{a}}{\bar{b}}\bar{d} - \bar{H}y\right)^T - uu^T, \tag{393}$$

kde $u = (I - Tyy^T/y^TTy)\bar{S}\tilde{z}/\|\tilde{z}\|$.

Důkaz Roznásobením se snadno ukáže, že vzorec (387) lze zapsat ve tvaru

$$\bar{S}_+ = \left(I - \frac{Tyy^T}{y^TTy}\right)\bar{S}\left(I - \frac{\tilde{z}\tilde{z}^T}{\tilde{z}^T\tilde{z}}\right) + \rho\frac{\bar{d}\tilde{z}^T}{\tilde{z}^T\tilde{z}}.$$

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

$$\begin{aligned}\bar{H}_+ = \bar{S}_+\bar{S}_+^T &= \left(I - \frac{Tyy^T}{y^TTy}\right)\bar{S}\left(I - \frac{\tilde{z}\tilde{z}^T}{\tilde{z}^T\tilde{z}}\right)\bar{S}^T\left(I - \frac{Tyy^T}{y^TTy}\right)^T + \rho\frac{\bar{d}\bar{d}^T}{\bar{b}} \\ &= \left(I - \frac{Tyy^T}{y^TTy}\right)\bar{H}\left(I - \frac{Tyy^T}{y^TTy}\right)^T + \rho\frac{\bar{d}\bar{d}^T}{\bar{b}} - uu^T,\end{aligned}$$

kde $u = (I - Tyy^T/y^TTy)\bar{S}\tilde{z}/\|\tilde{z}\|$. Použijeme-li první rovnost v (391), můžeme tento vzorec zapsat ve tvaru

$$\begin{aligned}\bar{H}_+ &= \left(I - \frac{Tyy^T}{y^TTy}\right)\bar{H}\left(I - \frac{Tyy^T}{y^TTy}\right)^T + \rho\frac{\bar{d}\bar{d}^T}{\bar{b}} - uu^T \\ &= \bar{H} + \frac{\rho}{\bar{b}}\bar{d}\bar{d}^T - \frac{1}{\bar{a}}\bar{H}y(\bar{H}y)^T + \frac{\eta}{\bar{a}}\left(\frac{\bar{a}}{\bar{b}}\bar{d} - \bar{H}y\right)\left(\frac{\bar{a}}{\bar{b}}\bar{d} - \bar{H}y\right)^T - uu^T\end{aligned}$$

(lze se o tom přesvědčit prostým dosazením a roznásobením závorek). □

Poznámka 244 Z vyjádření (393) vyplývá, že aktualizace (387) s volbou (391)–(392) je ekvivalentní aktualizaci (375) právě tehdy, když $u = 0$ v (393), neboli když $S\tilde{z} \parallel Ty$ v (391). 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 45 použijeme k důkazu důsledků vět 133 a 134.

Důsledek 19 *Nechť jsou splněny předpoklady věty 133, kde aktualizace (375) je nahražena aktualizací (387) s volbou (391)–(392). Pak platí (380).*

Důkaz Označíme-li $\hat{H}_+$ matici určenou vztahem (375) a použitou v lemmatu 43, můžeme psát $\bar{H}_+ = \hat{H}_+ - uu^T$, takže

$$\mathrm{Tr}\bar{H}_+ = \mathrm{Tr}\hat{H}_+ - u^T u \leq \mathrm{Tr}\hat{H}_+.$$

Můžeme tedy použít lemma 43, lemma 44 a větu 133. □

Důsledek 20 *Nechť jsou splněny předpoklady věty 134, kde aktualizace (375) je nahražena aktualizací (387) s volbou (391)–(392). Pak platí (380).*

Důkaz Označíme-li $\hat{H}_+$ matici určenou vztahem (375) a použitou v lemmatu 43, můžeme psát $\bar{H}_+ = \hat{H}_+ - uu^T$ a použijeme-li důsledek 8, 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 43, lemma 44 a větu 134. □

8.4 Diferenční verze Newtonovy metody pro husté úlohy

Diferenční verze nepřesné Newtonovy metody se obvykle realizují jako nepřesné metody spádových směrů (algoritmus 3) nebo nepřesné metody s lokálně omezeným krokem (algoritmus 6). V případě hustých úloh se nepoužívá matice $B = G(x)$ a násobení $q = Bp = G(x)p$ se nahražuje numerickým derivováním

$$G(x)p \approx \frac{g(x + \delta p) - g(x)}{\delta},$$

kde $\delta = \varepsilon/\|p\|$ je vhodná diference (obvykle $\varepsilon = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost). Jinak se algoritmy nemění. Jestliže výpočet gradientu vyžaduje $O(n)$ operací, je tento způsob úspornější než násobení matice vektorem (obecně $O(n^2)$ operací). Navíc není třeba počítat druhé derivace. Vliv diference $\delta = \varepsilon/\|p\|$ na přesnost Newtonovy metody udává tato věta.

Věta 135 *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

$$\begin{aligned} \|\tilde{q}-q\| &= \frac{1}{\delta}\left\|\int_0^1 (G(x+\tau\delta p)-G(x))\delta p \mathrm{d}\tau\right\| \leq \frac{1}{\delta}\int_0^1 \|G(x+\tau\delta p)-G(x))\|\|\delta p\|\mathrm{d}\tau \\ &\leq \frac{1}{\delta}\int_0^1 \overline{L}\|\delta p\|^2\tau\mathrm{d}\tau = \frac{1}{2}\overline{L}\delta\|p\|^2 = \frac{1}{2}\varepsilon\overline{L}\|p\| \end{aligned}$$

□

Nyní budeme předpokládat, že směrový vektor s se určuje lineární metodou sdružených gradientů popsanou v oddílu 3.6. 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 136 *Uvažujme metodu sdružených gradientů (definice 22) 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 věty 36 a věty 135 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

$$\|\tilde{g}_{m+1} - g_{m+1}\| \leq \overline{\vartheta}\|s_{m+1}\|, \qquad \overline{\vartheta} = \frac{m}{2}\varepsilon\overline{L}. \tag{394}$$

Důkaz Použijeme-li nerovnost uvedenou ve větě 135, dostaneme

$$\|\tilde{g}_{m+1} - g_{m+1}\| = \left\|\sum_{j=1}^m \alpha_j(\tilde{q}_j - q_j)\right\| \leq \sum_{j=1}^m \alpha_j\|\tilde{q}_j - q_j\| \leq \frac{1}{2}\varepsilon\overline{L}\sum_{j=1}^m \alpha_j\|p_j\|.$$

V části (c) důkazu věty 36 je ukázáno, že pro $1 \leq i \leq m$ platí $\alpha_j\|p_j\| \leq \|s_{j+1}\| \leq \|s_{m+1}\|$, takže

$$\sum_{i=j}^m \alpha_j\|p_j\| \leq m\|s_{m+1}\|,$$

což spolu s předchozí nerovností dokazuje tvrzení věty. □

Poznámka 245 Předpokládejme, že v m-tém kroku metody sdružených gradientů platí $\|\tilde{g}_{m+1}\| \leq \overline{\omega}\|g\|$, $0 \leq \overline{\omega} < 1$. Pak, položíme-li $s = s_{m+1}$ a $\tilde{g} = \tilde{g}_{m+1}$, můžeme podle předpokladu a podle věty 136 psát

$$\frac{\|\tilde{G}s+g\|}{\|g\|} \leq \overline{\omega}, \qquad \frac{\|(\tilde{G}-G)s\|}{\|s\|} \leq \overline{\vartheta},$$

kde $\tilde{G}$ je nějaká symetrická matice, pro kterou platí $\tilde{G}s + g = \tilde{g}$, a kde $\overline{\vartheta} = m\varepsilon\overline{L}/2$. Zvolíme-li číslo ε tak, že $\varepsilon < (1-\overline{\omega})\underline{G}/(m\overline{L})$, platí $\overline{\vartheta} < (1-\overline{\omega})\,\underline{G}/2$ a můžeme použít větu 17 a poznámku 36 k určení asymptotické rychlosti konvergence. Poznamenejme, že odhady použité ve větě 17 a ve větě 136 jsou značně nadhodnocené, takže diferenční verze nepřesné Newtonovy metody obvykle fungují velmi dobře i pro standardní volbu $\varepsilon = \sqrt{\varepsilon_M}$.

Nevýhodou metod studovaných v tomto oddílu 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 diferenčních verzí Newtonovy metody pro husté úlohy:

(1) Použití metod s proměnnou metrikou s omezenou pamětí.
(2) Použití pásových matic určených standardní metodou BFGS, která je ekvivalentní předpodmíněné metodě sdružených gradientů.
(3) Použití pásových matic určených numerickým derivováním.
(4) Použití tridiagonálních matic určených Lanczosovou metodou, která je ekvivalentní nepředpodmíněné metodě sdružených gradientů.

Použití 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 37. Vektory $C^{-1}g_j = H_i^i\, g_j$, sloužící k výpočtu vektorů p_j (definice 22), určujeme buď pomocí Strangových rekurencí (důsledek 17), kde místo vektoru g_i použijeme postupně vektory g_j, $1 \le j \le m$, nebo pomocí maticových reprezentací (věta 122, věta 124). 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 (81) generují stejnou posloupnost vektorů p_j, $1 \le \jmath \le m$, jako předpodmíněná metoda sdružených gradientů uvedená v definici 22 (věta 41 a důsledek 5). Pro metodu BFGS platí $B_j p_j + g_j = 0$, $1 \le \jmath \le m$, kde $B_1 = C$ a

$$B_{j+1} = B_j + \frac{y_j y_j^T}{d_j^T y_j} - \frac{B_j d_j (B_j d_j)^T}{d_j^T B_j d_j} = B_j + \frac{G p_j (G p_j)^T}{p_j^T G p_j} + \frac{g_j g_j^T}{p_j^T g_j},$$

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 $q_j = G p_j$ a g_j vektory $\tilde{q}_j$ (určené numerickým derivováním) a $\tilde{g}_j$, můžeme psát

$$B_{j+1} = B_j + \frac{\tilde{q}_j \tilde{q}_j^T}{p_j^T \tilde{q}_j} + \frac{\tilde{g}_j \tilde{g}_j^T}{p_j^T \tilde{g}_j}. \tag{395}$$

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 těchto aktualizacích 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. Zde se omezíme na diagonální, tridiagonální a pentadiagonální předpodmiňovače. Ukazuje se, že je důležité, aby tyto předpodmiňovače byly pozitivně definitní (poznámka 155).

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

Tvrzení 6 *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 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 provádět při určování trojúhelníkového rozkladu matice T. Výhodnější je však upravit matici T předem tak, aby byla pozitivně definitní. Jednou z možností je použít větu 137.

Lemma 46 *Uvažujme tridiagonální matici*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \dots & 0 & 0 \\ \beta_1 & \alpha_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & \dots & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{396}$$

a označme Δ_i hlavní subdeterminant i-tého řádu matice T (obsahující řádky a sloupce s indexy $1, 2, \dots, i$). Pak platí $\Delta_1 = \alpha_1$ a

$$\Delta_i = \alpha_i \Delta_{i-1} - \beta_{i-1}^2 \Delta_{i-2}, \qquad 1 < i \le n, \tag{397}$$

kde pokládáme $\Delta_0 = 1$.

Důkaz Pro $i = 2$ je tvrzení lemmatu zřejmé. Nechť $i > 2$. Rozvedeme-li subdeterminant Δ_i podle i-tého řádku, dostaneme $\Delta_i = \alpha_i \Delta_{i-1} - \beta_{i-1}^2 \Delta_{i-1}$, neboť i-tý řádek obsahuje pouze dva prvky. □

Věta 137 *Tridiagonální matice (396) je pozitivně definitní právě tehdy, když $\gamma_i > 0$ pro $1 \le i \le n$, kde $\gamma_1 = \alpha_1$ a*

$$\gamma_i = \alpha_i - \frac{\beta_{i-1}^2}{\gamma_{i-1}}, \qquad 1 < i \le n. \tag{398}$$

Důkaz Dokážeme indukcí, že $\Delta_i = \gamma_i \Delta_{i-1}$ pro $1 \le i \le n$, kde opět $\Delta_0 = 1$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme, že pro nějaký index $i > 1$ platí $\Delta_{i-1} = \gamma_{i-1} \Delta_{i-2}$. Použijeme-li (397) a (398), dostaneme

$$\begin{aligned} \Delta_i &= \alpha_i \Delta_{i-1} - \beta_{i-1}^2 \Delta_{i-2} = \alpha_i \Delta_{i-1} + \gamma_{i-1}(\gamma_i - \alpha_i)\Delta_{i-2} \\ &= (\Delta_{i-1} - \gamma_{i-1}\Delta_{i-2})\alpha_i + \gamma_{i-1}\gamma_i \Delta_{i-2} = \gamma_i \Delta_{i-1}, \end{aligned}$$

čímž je indukční krok dokončen. Jelikož $\Delta_i = \gamma_i \Delta_{i-1}$ pro $1 \le i \le n$, platí $\Delta_i > 0$ právě tehdy, když $\gamma_i > 0$ (pro $1 \le i \le n$). □

Větu 137 můžeme použít tak, že počítáme čísla γ_i, $1 < i \le n$, a pokud pro nějaký index platí $\gamma_i \le 0$, zmenšíme mimodiagonální prvek β_{i-1} tak, aby platilo $\beta_{i-1}^2 < \gamma_{i-1}\alpha_i$ (například položíme $\beta_{i-1}^2 = \lambda_{i-1}\gamma_{i-1}\alpha_i$, kde $0 < \lambda_{i-1} < 1$). Pak je nové číslo γ_i kladné. Potíž je v tom, že zde není žádná rezerva a zvolíme-li λ_{i-1} nevhodně, může být výsledná tridiagonální matice špatně podmíněná. Pro praktické účely je výhodnější použít větu 138 a její důsledek.

Věta 138 *Uvažujme tridiagonální matici (396) s kladnými prvky na hlavní diagonále. Pak jsou-li matice*

$$\begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix}, \qquad 1 \le i < n-1, \tag{399}$$

pozitivně semidefinitní je matice T pozitivně definitní.

Důkaz Pro libovolný vektor $v \in R^n$ platí

$$\begin{aligned} v^T T v &= \sum_{i=1}^{n} \alpha_i v_i^2 + 2\sum_{i=1}^{n-1} \beta_i v_i v_{i+1} \\ &= \frac{1}{2}\alpha_1 v_1^2 + \frac{1}{2}\sum_{i=1}^{n-1} \left(\alpha_i v_i^2 + \alpha_{i+1} v_{i+1}^2 + 4\beta_i v_i v_{i+1}\right) + \frac{1}{2}\alpha_n v_n^2 \\ &= \frac{1}{2}\alpha_1 v_1^2 + \frac{1}{2}\sum_{i=1}^{n-1} [v_i, v_{i+1}] \begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} + \frac{1}{2}\alpha_n v_n^2 \end{aligned} \tag{400}$$

Jelikož matice vystupující v této rovnosti jsou podle předpokladu pozitivně semidefinitní, platí $v^T T v \ge 0$. 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 $1 \le j \le i$, 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 $v^T T v \ge 0$ a že z $v^T T v = 0$ plyne $v = 0$, takže matice T je pozitivně definitní. □

Důsledek 21 *Nechť tridiagonální matice T obsahuje hlavní diagonálu a poloviny vedlejších diagonál pozitivně definitní matice B (takže $\alpha_i = b_{i,i}$, $1 \le i \le n-1$ a $\beta_i = b_{i,i+1}/2$, $1 \le i \le n-1$). 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

$$\begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix} = \begin{bmatrix} b_{i,i} & b_{i,i+1} \\ b_{i,i+1} & b_{i+1,i+1} \end{bmatrix}, \qquad 1 \le i \le n-1.$$

Tyto matice jsou pozitivně definitní, neboť matice B je pozitivně definitní. □

Důsledek 21 můžeme použít tak, že zmenšíme prvky vedlejších diagonál matice B na polovinu. Pak je výsledná tridiagonální matice pozitivně definitní. Větu 138 můžeme použít tak, že počítáme determinanty $\alpha_i\alpha_{i+1} - 4\beta_i^2$, $1 \le i \le n-1$, a pokud pro nějaký index platí $\alpha_i\alpha_{i+1} - 4\beta_i^2 < 0$, zmenšíme mimodiagonální prvek β_i tak, aby platilo $\beta_i^2 = \alpha_i\alpha_{i+1}/4$. Pak $\alpha_i\alpha_{i+1} - 4\beta_i^2 = 0$, $1 \le i < n-1$, takže odpovídající matice jsou pozitivně semidefinitní a matice T je podle věty 138 pozitivně definitní.

Poznámka 246 Věta 138 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 138 neplatí, ale vlastní čísla zůstanou kladná. Tato výhoda se projeví, počítáme-li prvky matice T nepřesně pomocí numerického derivování (věta 143).

Poznámka 247 Podíváme-li se na poslední výraz ve vzorci (400), 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 $2 \le i < n-2$, 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 bychom potřebovali (věta 143).

Větu 138 a její důsledek lze zobecnit tak, že platí i pro obecnou pásovou matici. Ukážeme, jak to vypadá v případě pentadiagonální matice

$$P = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 & \dots & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \dots & 0 & 0 & 0 \\ \gamma_1 & \beta_2 & \alpha_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{n-2} & \beta_{n-2} & \gamma_{n-2} \\ 0 & 0 & 0 & \dots & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & 0 & \dots & \gamma_{n-2} & \beta_{n-1} & \alpha_n \end{bmatrix}. \tag{401}$$

Věta 139 *Uvažujme pentadiagonální matici (401) s kladnými prvky na hlavní diagonále. Pak, jsou-li matice*

$$\begin{bmatrix} \alpha_i & (3/2)\beta_i & 3\gamma_i \\ (3/2)\beta_i & \alpha_{i+1} & (3/2)\beta_{i+1} \\ 3\gamma_i & (3/2)\beta_{i+1} & \alpha_{i+2} \end{bmatrix}, \qquad 1 \le i < n-2, \tag{402}$$

pozitivně semidefinitní, je matice P pozitivně definitní.

Důkaz Pro libovolný vektor $v \in R^n$ platí

$$\begin{aligned} v^T P v &= \sum_{i=1}^{n} \alpha_i v_i^2 + 2\sum_{i=1}^{n-1} \beta_i v_i v_{i+1} + 2\sum_{i=1}^{n-2} \gamma_i v_i v_{i+1} \\ &= \frac{1}{3}\alpha_1 v_1^2 + \frac{1}{3}(\alpha_1 v_1^2 + \alpha_2 v_2^2) + \beta_1 v_1 v_2 \\ &+ \frac{1}{3}\sum_{i=1}^{n-2} \left(\alpha_i v_i^2 + \alpha_{i+1} v_{i+1}^2 + \alpha_{i+2} v_{i+2}^2 + 3\beta_i v_i v_{i+1} + 3\beta_{i+1} v_{i+1} v_{i+2} + 6\gamma_i v_i v_{i+2}\right) \\ &+ \frac{1}{3}(\alpha_{n-1} v_{n_1}^2 + \alpha_n v_n^2) + \beta_{n-1} v_{n-1} v_n + \frac{1}{3}\alpha_n v_n^2 \\ &= \frac{1}{3}\alpha_1 v_1^2 + \frac{1}{3}\,[v_1, v_2] \begin{bmatrix} \alpha_1 & (3/2)\beta_1 \\ (3/2)\beta_1 & \alpha_2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ &+ \frac{1}{3}\sum_{i=1}^{n-2} [v_i, v_{i+1}, v_{i+2}] \begin{bmatrix} \alpha_i & (3/2)\beta_i & 3\gamma_i \\ (3/2)\beta_i & \alpha_{i+1} & (3/2)\beta_{i+1} \\ 3\gamma_i & (3/2)\beta_{i+1} & \alpha_{i+2} \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \\ v_{i+2} \end{bmatrix} \\ &+ \frac{1}{3}\,[v_{n-1}, v_n] \begin{bmatrix} \alpha_{n-1} & (3/2)\beta_{n-1} \\ (3/2)\beta_{n-1} & \alpha_2 \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} + \frac{1}{3}\alpha_n v_n^2 \end{aligned}$$

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 138, musí být $v_1 = 0$ a $v_2 = 0$. Předpokládejme, že $v_j = 0$ pro $1 \le j \le i+1$, kde $i < n-1$. 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 22 *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. □

Poznámka 248 Nechť C je 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 $1 \le i \le k-1$), je výsledná matice pozitivně definitní. Důkaz tohoto tvrzení je podobný důkazu důsledku 22 (používá se analogie věty 139).

Důsledek 22 můžeme použít tak, že v matici B zmenšíme prvky prvních vedlejších diagonál na dvě třetiny a prvky druhých vedlejších diagonál na třetinu. Pak je výsledná tridiagonální matice pozitivně definitní. Také můžeme testovat pozitivní definitnost matic (402) a měnit jejich vedlejší diagonály pouze tehdy, nejsou-li pozitivně definitní. Větu 139 můžeme použít tak, že nejprve počítáme subdeterminanty $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2$, $1 \le i \le n-1$, a pokud je některý z nich záporný, zmenšíme mimodiagonální prvek β_i tak, aby platilo $\beta_i^2 = (4/9)\alpha_i\alpha_{i+1}$. Pak počítáme determinanty matic (402) a je-li některý z nich záporný, upravíme odpovídající prvek γ_i podle věty 140 a poznámky 249.

Věta 140 *Determinanty Δ_i matic (402) spočteme podle vzorce*

$$\Delta_i = \alpha_{i+1}\left(\alpha_i\alpha_{i+2} - 9\gamma_i^2\right) - \frac{9}{4}\left(\alpha_i\beta_{i+1}^2 + \alpha_{i+2}\beta_i^2 - 6\beta_i\beta_{i+1}\gamma_i\right). \tag{403}$$

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 $\Delta_i = 0$. Přitom

$$D_i = \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right)$$

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 (403) dostaneme snadno vyčíslením příslušného determinantu. Jelikož kvadratický člen v (403) 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 (403) se diskriminant této rovnice (vydělený číslem 36) rovná

$$\begin{aligned} D_i &= \frac{81}{16}\beta_i^2\beta_{i+1}^2 - \frac{9}{4}\alpha_i\alpha_{i+1}\beta_{i+1}^2 - \frac{9}{4}\alpha_{i+1}\alpha_{i+2}\beta_i^2 + \alpha_i\alpha_{i+1}^2\alpha_{i+2} \\ &= \frac{9}{4}\beta_{i+1}^2\left(\frac{9}{4}\beta_i^2 - \alpha_i\alpha_{i+1}\right) - \alpha_{i+1}\alpha_{i+2}\left(\frac{9}{4}\beta_i^2 - \alpha_i\alpha_{i+1}\right) \\ &= \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right). \end{aligned}$$

□

Poznámka 249 Věta 140 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

$$\gamma_i = \frac{1}{2}(\underline{\gamma}_i + \overline{\gamma}_i) = \frac{3}{4}\frac{\beta_i\beta_{i+1}}{\alpha_{i+1}}. \tag{404}$$

Další možností, jak konstruovat pásové předpodmiňovače, je předpokládat, že Hessova matice má pásovou strukturu a určovat její prvky numerickým derivováním. 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 diferencí gradientů, tedy spočítat v každém kroku Newtonovy metody k gradientů navíc. Vyšetříme opět tři speciální případy.

Poznámka 250 Předpokládejme, že Hessova matice je diagonální. Pak lze všechny její prvky aproximovat pomocí jedné diference gradientů

$$G(x)v \approx g(x+v) - g(x), \qquad v = [\delta_1, \ldots, \delta_n]^T,$$

kde $\delta_1, \ldots, \delta_n$ jsou vhodné diference. K předpodmínění pak použijeme diagonální matici $C = D$, kde $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ a $Dv = g(x+v) - g(x)$. Po dosazení dostaneme $\alpha_i\delta_i = g_i(x+v) - g_i(x)$, neboli

$$\alpha_i = \frac{g_i(x+v) - g_i(x)}{\delta_i}, \qquad 1 \le i \le n.$$

Poznámka 251 Diference lze volit dvojím způsobem. V prvním případě pokládáme $\delta_i = \delta$, $1 \le i \le n$, takže $v = \delta e$, kde e je vektor, jehož všechny prvky jsou jednotkové. Pak lze (tak jako ve větě 135) volit $\delta = \sqrt{\varepsilon_M}/\|e\| = \sqrt{\varepsilon_M\, k/n}$ ($k = 1$ je počet vedlejších diagonál zvětšený o 1). Ve druhém případě pokládáme $\delta_i = \sqrt{\varepsilon_M}\max(|x_i|, 1)$, $1 \le i \le n$. Tento způsob je méně náchylný k vlivu zaokrouhlovacích chyb. V obou případech lze psát

$$\delta_i = \varepsilon\overline{\delta}_i, \qquad 1 \le i \le n, \tag{405}$$

kde $\varepsilon = \sqrt{\varepsilon_M}$ a buď $\overline{\delta}_i = \sqrt{k/n}$ nebo $\overline{\delta}_i = \max(|x_i|, 1)$ pro $1 \le i \le n$.

Nevýhodou předpodmiňovačů založených na numerickém derivování je skutečnost, že nemusí být pozitivně definitní. Uvažujme ryze konvexní kvadratickou funkci $F : R^2 \to R$, kde

$$F(x) = \frac{1}{2}x^T \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x, \qquad g(x) = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x.$$

Pak platí

$$\frac{g(x+\delta e) - g(x)}{\delta} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

takže

$$De = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

což dává $\alpha_1 = -1$, $\alpha_2 = 4$, takže matice D není pozitivně definitní. Tuto nevýhodu můžeme odstranit tak, že pokládáme

$$\alpha_i = \frac{|g_i(x+v) - g_i(x)|}{\delta_i}, \qquad 1 \le i \le n.$$

Zdůvodnění této úpravy udává následující tvrzení.

Tvrzení 7 *Nechť $\mathcal{D}_n$ je množina všech diagonálních matic řádu n a $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ je diagonální matice taková, že*

$$\alpha_i = \sum_{j=1}^{n} |G_{ij}|, \qquad 1 \le j \le n,$$

kde G_{ij}, $1 \le j \le n$, 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 položíme-li $v = \delta e$, platí $D(\delta e) = g(x + \delta e) - g(x) \approx G(\delta e)$, takže

$$\alpha_i \approx \sum_{j=1}^{n} G_{ij} = \sum_{j=1}^{n} |G_{ij}|$$

a matice D je podle tvrzení 7 ideálním diagonálním předpodmiňovačem (v l_1 normě) pro soustavu rovnic $Gs + g = 0$. Nemá-li matice G pouze kladné prvky platí

$$|\alpha_i| \approx \left| \sum_{j=1}^{n} G_{ij} \right| \le \sum_{j=1}^{n} |G_{ij}|,$$

takže prvky upravené matice D jsou dolním odhadem prvků ideálního diagonálního předpodmiňovače.

Nyní ukážeme, jak lze numerické derivování použít ke konstrukci tridiagonálního předpodmiňovače.

Věta 141 *Předpokládejme, že Hessova matice funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ je tridiagonální matice tvaru (396). Položme $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \ldots]$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \ldots]$, kde $\delta_i = \varepsilon\overline{\delta}_i$, $1 \le i \le n$ (tak jako v (405)). Pak pro $2 \le i \le n-1$ platí*

$$\begin{aligned}
\alpha_1 &= \lim_{\varepsilon\to 0} \frac{g_1(x+v_1) - g_1(x)}{\delta_1}, & \beta_1 &= \lim_{\varepsilon\to 0} \frac{g_1(x+v_2) - g_1(x)}{\delta_2}, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 1, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 0, \\
\alpha_n &= \lim_{\varepsilon\to 0} \frac{g_n(x+v_1) - g_n(x)}{\delta_n}, & & \mathrm{mod}(n,2) = 1, \\
\alpha_n &= \lim_{\varepsilon\to 0} \frac{g_n(x+v_2) - g_n(x)}{\delta_n}, & & \mathrm{mod}(n,2) = 0.
\end{aligned}$$

Důkaz Podle věty 3 platí $g(x + v_1) - g(x) = G(x)v_1 + o(\varepsilon)$, $g(x + v_2) - g(x) = G(x)v_2 + o(\varepsilon)$, takže po dosazení $G(x) = T$, kde T je tridiagonální matice tvaru (396), a po rozepsání jednotlivých složek dostaneme

$$\begin{aligned}
\frac{g_1(x+v_1) - g_1(x)}{\delta_1} &= \alpha_1 + o(1), & \frac{g_1(x+v_2) - g_1(x)}{\delta_2} &= \beta_1 + o(1), \\
\frac{g_i(x+v_1) - g_i(x)}{\delta_i} &= \alpha_i + o(1), & \frac{g_i(x+v_2) - g_i(x)}{\delta_{i+1}} &= \beta_{i+1} + \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}} + o(1), \quad \mathrm{mod}(i,2) = 1, \\
\frac{g_i(x+v_2) - g_i(x)}{\delta_i} &= \alpha_i + o(1), & \frac{g_i(x+v_1) - g_i(x)}{\delta_{i+1}} &= \beta_{i+1} + \beta_{i-1}\frac{\delta_{i-1}}{\delta_{i+1}} + o(1), \quad \mathrm{mod}(i,2) = 0, \\
\frac{g_n(x+v_1) - g_i(x)}{\delta_i} &= \alpha_n + o(1), & & \mathrm{mod}(n,2) = 1, \\
\frac{g_n(x+v_2) - g_i(x)}{\delta_i} &= \alpha_n + o(1), & & \mathrm{mod}(n,2) = 0,
\end{aligned}$$

kde $2 \le i \le n-1$. Jelikož podíly $\delta_{i-1}/\delta_{i+1} = \overline{\delta}_{i-1}/\overline{\delta}_{i+1}$ zůstávají pro $2 \le i \le n-1$ konstantní, je tvrzení věty dokázáno. □

Poznámka 252 Věta 141 udává způsob určení tridiagonálního předpodmiňovače. Zvolí se pevně číslo ε (například $\varepsilon = \sqrt{\varepsilon_M}$ jako v (405)) a prvky matice $C = T$ se vypočtou podle vzorců uvedených ve větě 141 (ve kterých je vynechán limitní přechod).

Matice $C = T$ získaná podle poznámky 252 nemusí být pozitivně definitní, i když Hessova matice je pozitivně definitní (jako příklad lze uvést ryze konvexní kvadratickou funkci tří proměnných s pozitivně definitní Hessovou maticí

$$G = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 4 & -1 \\ -2 & -1 & 8 \end{bmatrix}.$$

Způsoby korekce indefinitní tridiagonální matice jsou popsány v poznámce 155. Nyní uvedeme dvě věty, podporující volbu tridiagonálního předpodmiňovače v případech, kdy Hessova matice je pentadiagonální. Z technických důvodů budeme tridiagonální předpodmiňovač a jeho prvky označovat vlnkou.

Věta 142 *Nechť Hessova matice $G(x)$ je pentadiagonální, pozitivně definitní a diagonálně dominantní. Pak, platí-li $\delta_i = \varepsilon\overline{\delta}$, $1 \le i \le n$, a je-li číslo ε dostatečně malé, je matice $C = \tilde{T}$ získaná podle poznámky 252 pozitivně definitní a diagonálně dominantní.*

Důkaz Uvažujme pentadiagonální Hessovu matici tvaru (401), označme vlnkou prvky tridiagonální matice $\tilde{T}$ a položme pro zjednodušení zápisu $\gamma_{-1} = \gamma_0 = \beta_0 = \tilde{\beta}_0 = 0$, $\gamma_{n-1} = \gamma_n = \beta_n = \tilde{\beta}_n = 0$. Pak podle předpokladu diagonální dominance platí

$$\alpha_i > |\gamma_{i-2}| + |\beta_{i-1}| + |\beta_i| + |\gamma_i|$$

pro $1 \le i \le n$. Použijeme-li větu 141 a poznámku 252, dostaneme

$$\tilde{\alpha}_i \approx \gamma_{i-2} + \alpha_i + \gamma_i, \qquad \tilde{\beta}_{i-1} + \tilde{\beta}_i \approx \beta_{i-1} + \beta_i \tag{406}$$

pro $1 \le i \le n$. Platí tedy $\tilde{\beta}_i \approx \beta_i$ a je-li číslo ε dostatečně malé, zůstane ostrá nerovnost zachována a můžeme psát

$$\tilde{\alpha}_i \ge \tilde{\alpha}_i - |\tilde{\beta}_{i-1}| + |\tilde{\beta}_i| \approx \alpha_i + \gamma_{i-2} + \gamma_i - |\beta_{i-1}| + |\beta_i| \ge \alpha_i - |\gamma_{i-2}| - |\beta_{i-1}| - |\beta_i| - |\gamma_i| > 0$$

pro $1 \le i \le n$. □

Poznámka 253 Věta 142 vyžaduje, aby všechny diference byly stejné, což je splněno například tehdy, když $\delta_i = \sqrt{2\varepsilon_M/n}$, $1 \le i \le n$. Numerické testy však ukazují, že volba $\delta_i = \sqrt{\varepsilon}\max(|x_i|, 1)$, $1 \le i \le n$, je výhodnější.

Věta 142 používá poměrně silné předpoklady, udává však podmínky postačující, nikoliv nutné. Ukazuje se, že pro mnohé praktické úlohy je matice $\tilde{T}$ pozitivně definitní. Uvažujme okrajovou úlohu pro obyčejnou diferenciální rovnici druhého řádu

$$y''(t) = \varphi(y(t)), \quad 0 \le t \le 1, \quad y(0) = y_0, \quad y(1) = y_1,$$

kde funkce $\varphi : R \to R$ je dvakrát spojitě diferencovatelná na R. Rozdělíme-li interval $[0, 1]$ na $n+1$ částí pomocí uzlových bodů $t_i = ih$, $0 \le i \le n+1$, 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 $1 \le 1 \le n$, dostaneme soustavu n nelineárních rovnic

$$h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1} = 0,$$

kde $x_i = y(t_i)$, $0 \le 1 \le n+1$, takže $x_0 = y_0$ a $x_{n+1} = y_1$. Řešíme-li tuto soustavu metodou nejmenších čtverců (kapitola 7), má minimalizovaná funkce tvar

$$F(x) = \frac{1}{2}\sum_{i=1}^{n} f_i^2(x) = \frac{1}{2}\sum_{i=1}^{n} \left(h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1}\right)^2, \tag{407}$$

kde $x = [x_1, \ldots, x_n]^T$.

Věta 143 *Aplikujme diferenční verzi Newtonovy metody na součet čtverců (407), kde funkce $\varphi : R \to R$ je lineární. Pak, platí-li $\delta_i = \varepsilon\overline{\delta}$, $1 \le i \le n$, a je-li číslo ε dostatečně malé, je matice $C = \tilde{T}$ získaná podle poznámky 252 pozitivně definitní.*

Důkaz Zřejmě

$$\nabla f_i(x) = \begin{bmatrix} -1 \\ \psi(x_i) \\ -1 \end{bmatrix}, \qquad \nabla^2 f_i(x) = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \psi'(x_i) & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

kde $\psi(x_i) = 2 + h^2\varphi'(x_i)$ a $\psi'(x_i) = h^2\varphi''(x_i)$. Omezíme-li se pro jednoduchost na submatice řádu pět, můžeme psát

$$J(x) = \begin{bmatrix} \psi_1 & -1 & 0 & 0 & 0 \\ -1 & \psi_2 & -1 & 0 & 0 \\ 0 & -1 & \psi_3 & -1 & 0 \\ 0 & 0 & -1 & \psi_4 & -1 \\ 0 & 0 & 0 & -1 & \psi_5 \end{bmatrix}, \quad C(x) = \begin{bmatrix} f_1\psi_1' & 0 & 0 & 0 & 0 \\ 0 & f_2\psi_2' & 0 & 0 & 0 \\ 0 & 0 & f_3\psi_3' & 0 & 0 \\ 0 & 0 & 0 & f_4\psi_4' & 0 \\ 0 & 0 & 0 & 0 & f_5\psi_5' \end{bmatrix},$$

$$J^T(x)J(x) = \begin{bmatrix} \psi_1^2+1 & -(\psi_1+\psi_2) & 1 & 0 & 0 \\ -(\psi_1+\psi_2) & \psi_2^2+2 & -(\psi_2+\psi_3) & 1 & 0 \\ 1 & -(\psi_2+\psi_3) & \psi_3^2+2 & -(\psi_3+\psi_4) & 1 \\ 0 & 1 & -(\psi_3+\psi_4) & \psi_4^2+2 & -(\psi_4+\psi_5) \\ 0 & 0 & 1 & -(\psi_4+\psi_5) & \psi_5^2+1 \end{bmatrix},$$

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 funke φ), platí $C(x) = 0$, takže $G(x) = J^T(x)J(x)$ a můžeme psát $\alpha_1 = \psi_1^2 + 1$, $\alpha_n = \psi_n^2 + 1$ a

$$\begin{aligned} \alpha_i &= \psi_i^2 + 2, & 2 \le i \le n-1, \\ \beta_i &= -(\psi_i + \psi_{i+1}), & 1 \le i \le n-1, \\ \gamma_i &= 1, & 1 \le i \le n-2. \end{aligned}$$

Je-li $\tilde{T}$ matice získaná podle poznámky 252, platí (406), což dává $\tilde{\alpha}_1 \approx \psi_1^2 + 2$, $\tilde{\alpha}_2 \approx \psi_2^2 + 3$, $\tilde{\alpha}_i \approx \psi_i^2 + 4$, $3 \le i \le n-2$, $\tilde{\alpha}_{n-1} \approx \psi_{n-1}^2 + 3$, $\tilde{\alpha}_n \approx \psi_n^2 + 2$ a $\tilde{\beta}_i \approx -(\psi_i + \psi_{i+1})$, $1 \le i \le n-1$. Nyní využijeme toho, že součet (400) (kde používáme veličiny označené vlnkou) lze podobně jako v poznámce 247 upravit tak, že

$$\begin{aligned} 2v^T\tilde{T}v \;=\; & [v_1, v_2]\begin{bmatrix} 2\tilde{\alpha}_1 & 2\tilde{\beta}_1 \\ 2\tilde{\beta}_1 & \tilde{\alpha}_2 - 1 \end{bmatrix}\begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\ & + \; [v_2, v_3]\begin{bmatrix} \tilde{\alpha}_2 + 1 & 2\tilde{\beta}_2 \\ 2\tilde{\beta}_2 & \tilde{\alpha}_3 \end{bmatrix}\begin{bmatrix} v_2 \\ v_3 \end{bmatrix} \\ & + \; \sum_{i=3}^{n-3} [v_i, v_{i+1}]\begin{bmatrix} \tilde{\alpha}_i & 2\tilde{\beta}_i \\ 2\tilde{\beta}_i & \tilde{\alpha}_{i+1} \end{bmatrix}\begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\ & + \; [v_{n-2}, v_{n-1}]\begin{bmatrix} \tilde{\alpha}_{n-2} & 2\tilde{\beta}_{n-2} \\ 2\tilde{\beta}_{n-2} & \tilde{\alpha}_{n-1} + 1 \end{bmatrix}\begin{bmatrix} v_{n-2} \\ v_{n-1} \end{bmatrix} \end{aligned}$$

$$
\begin{aligned}
&+ \quad [v_{n-1}, v_n] \begin{bmatrix} \tilde{\alpha}_{n-1} - 1 & 2\tilde{\beta}_{n-1} \\ 2\tilde{\beta}_{n-1} & 2\tilde{\alpha}_n \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} \\
\approx \quad & [v_1, v_2] \begin{bmatrix} 2(\psi_1^2 + 2) & -2(\psi_1 + \psi_2) \\ -2(\psi_1 + \psi_2) & \psi_2^2 + 2 \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \\
&+ \quad \sum_{i=2}^{n-2} [v_i, v_{i+1}] \begin{bmatrix} \psi_i^2 + 4 & -2(\psi_i + \psi_{i+1}) \\ -2(\psi_i + \psi_{i+1}) & \psi_{i+1}^2 + 4 \end{bmatrix} \begin{bmatrix} v_i \\ v_{i+1} \end{bmatrix} \\
&+ \quad [v_{n-1}, v_n] \begin{bmatrix} \psi_{n-1}^2 + 2 & -2(\psi_{n-1} + \psi_n) \\ -2(\psi_{n-1} + \psi_n) & 2(\psi_n^2 + 2) \end{bmatrix} \begin{bmatrix} v_{n-1} \\ v_n \end{bmatrix} \triangleq 2v^T T v. \qquad (408)
\end{aligned}
$$

Protože

$$
\begin{aligned}
2(\psi_i^2 + 2)(\psi_{i+1}^2 + 2) - 4(\psi_i + \psi_{i+1})^2 &= 2\psi_i^2\psi_{i+1}^2 + 8 - 8\psi_i\psi_{i+1} \\
&= 2(\psi_i^2\psi_{i+1}^2 - 2)^2 \geq 0, \quad i \in \{1, n-1\}, \\
(\psi_i^2 + 4)(\psi_{i+1}^2 + 4) - 4(\psi_i + \psi_{i+1})^2 &= \psi_i^2\psi_{i+1}^2 + 16 - 8\psi_i\psi_{i+1} \\
&= (\psi_i\psi_{i+1} - 4) \geq 0, \qquad 2 \leq i \leq n-2,
\end{aligned}
$$

jsou všechny matice použité v (408) pozitivně semidefinitní. Jelikož funkce φ je lineární (má směrnici φ'), platí $\psi_i = 2 + h^2\varphi'$, $1 \leq i \leq n$. Pokud $(2 + h^2\varphi')^2 \neq 2$, jsou první a poslední matice v (408) pozitivně definitní. V opačném případě jsou všechny ostatní matice pozitivně definitní. Podle poznámky 247 je tedy matice T pozitivně definitní. Protože podle poznámky 246 neovlivní malá změna prvků matice T její pozitivní definitnost, je pro dostatečně malé ε matice $\tilde{T} \approx T$ pozitivně definitní. □

Poznámka 254 Předpoklady věty 143 jsou velmi silné. Jsou to však podmínky postačující, nikoliv nutné. Předně je velmi nepravděpodobné, že by se současně vynulovaly determinanty všech matic použitých v (408), takže můžeme předpokládat, že matice T vystupující v (408) je pozitivně definitní. Pokud se blížíme k řešení, kde $F(x) = 0$, platí $f_i \to 0$, $1 \leq i \leq n$. Navíc matice $\mathrm{diag}(\psi_1', \ldots, \psi_n')$ je obvykle malá ve srovnání s $J(x)^T J(x)$ (pokud $n \approx 1000$, je $h^2 \approx 10^{-6}$). Protože malá změna diagonálních prvků neporuší pozitivní definitnost matice T (poznámka 246), můžeme očekávat, že v dostatečné blízkosti řešení je matice $\tilde{T}$ pozitivně definitní i když funkce funkce $\varphi : R \to R$ není lineární.

Je-li Hessova matice pentadiagonální a pozitivně definitní (tak jako v předchozích dvou větách), je výhodné použít pentadiagonální předpodmiňovač vyšetřovaný v následující větě, jejíž důkaz je velmi podobný důkazu věty 141.

Věta 144 *Předpokládejme, že Hessova matice funkce $F \in \mathcal{C}^2 : \mathcal{D} \to R$ je pentadiagonální matice tvaru (401). Položme $v_1 = [\delta_1, 0, 0, \delta_4, 0, 0, \ldots]$, $v_2 = [0, \delta_2, 0, 0, \delta_5, 0, \ldots]$, $v_3 = [0, 0, \delta_3, 0, 0, \delta_6, \ldots]$, kde $\delta_i = \varepsilon\overline{\delta}_i$, $1 \leq i \leq n$ (tak jako v (405)). Pak platí*

$$
\begin{aligned}
\alpha_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_1) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon \to 0} \frac{g_i(x + v_2) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_3) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i, 3) = 1, \\
\alpha_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_2) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon \to 0} \frac{g_i(x + v_3) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i, 3) = 2, \\
\alpha_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_3) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon \to 0} \frac{g_i(x + v_1) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon \to 0} \frac{g_i(x + v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i, 3) = 0,
\end{aligned}
$$

kde veličiny s indexy $i < 1$ považujeme za nulové a veličiny, v jejichž vzorcích vystupují indexy $i > n$, nepočítáme.

Nyní se budeme zabývat posledním typem předpodmínění zmíněným na začátku tohoto oddílu. V oddílu 6.4 jsme ukázali, že symetrický Lanczosův proces je úzce spjatý s metodou sdružených gradientů a ve větě 96 jsme uvedli příslušné převodní vztahy. Nyní ukážeme, jak lze tridiagonální matici získanou podle věty 96 použít ke konstrukci předpodmiňovače pro metodu sdružených gradientů. Z důvodů větší přehlednosti budeme prvky tridiagonální matice označovat tak jako v (396) a koeficienty metody sdružených gradientů označíme vlnkou. V této konvenci mají vzorce uvedené ve větě 96 tvar $\alpha_1 = 1/\tilde{\alpha}_1$ a

$$\beta_i^2 = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i^2}, \qquad \alpha_{i+1} = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i} + \frac{1}{\tilde{\alpha}_{i+1}} \tag{409}$$

pro $1 \leq i \leq m$, kde m je číslo takové, že $\tilde{\alpha}_i > 0$ pro $1 \leq i \leq m$.

Věta 145 *Uvažujme metodu sdružených gradientů (aplikovanou na kvadratickou funkci s Hessovou maticí G) takovou, že $\tilde{\alpha}_i > 0$ pro $1 \leq i \leq m$. Pak tridiagonální matice T_m řádu m získaná podle vzorců (409) je pozitivně definitní.*

Důkaz Označme $\gamma_i = 1/\tilde{\alpha}_i$, $1 \leq i \leq m$. Pak $\gamma_1 = \alpha_1$ a podle (409) platí

$$\beta_i^2 = \tilde{\beta}_i \gamma_i^2, \qquad \alpha_{i+1} = \tilde{\beta}_i \gamma_i + \gamma_{i+1}$$

pro $1 \leq i \leq m$, takže

$$\gamma_{i+1} = \alpha_{i+1} - \frac{\beta_i^2}{\gamma_i}$$

pro $1 \leq i \leq m$, což je právě rekurentní vztah (398). Jelikož čísla $\gamma_i = 1/\tilde{\alpha}_i$, $1 \leq i \leq m$, jsou podle předpokladu kladná, je podle věty 137 matice T_m pozitivně definitní. □

Poznámka 255 Tridiagonální matice T_m má dimenzi $m \leq n$. Abychom dostali předpodmiňovač dimenze n, položíme

$$C = (I - Q_m Q_m^T) + Q_m T_m Q_m^T \tag{410}$$

kde Q_m je matice s m ortonormálními sloupci získaná symetrickým Lanczosovým procesem (poznámka 185 a věta 96). Abychom zdůvodnili tuto volbu, ukážeme, že platí

$$C = [Q_m, Q_{n-m}] \begin{bmatrix} T_m & 0 \\ 0 & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T, \tag{411}$$

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

$$\left((I - Q_m Q_m^T) + Q_m T_m Q_m^T\right) v = Q_m T_m \tilde{v_1} + v_2 \tag{412}$$

a

$$\begin{aligned} [Q_m, Q_{n-m}] \begin{bmatrix} T_m & 0 \\ 0 & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T v &= [Q_m, Q_{n-m}] \begin{bmatrix} T_m & 0 \\ 0 & I_{n-m} \end{bmatrix} [\tilde{v_1}, \tilde{v_2}]^T \\ &= Q_m T_m \tilde{v_1} + Q_{n-m} \tilde{v_2} = Q_m T_m \tilde{v_1} + v_2 \end{aligned}$$

a jelikož vektor v lze volit libovolně, jsou obě matice stejné. Dále podle (412) 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 186), 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 146 *Nechť jsou splněny předpoklady věty 145. Pak předpodmiňovač (410) je pozitivně definitní a platí*

$$C^{-1} = (I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T. \tag{413}$$

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 145 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

$$\begin{aligned}\left[(I - Q_m Q_m^T) + Q_m T_m Q_m^T\right] &\left[(I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T\right] \\ &= I - Q_m Q_m^T + Q_m T_m Q_m^T - Q_m T_m Q_m^T + Q_m T_m^{-1} Q_m^T - Q_m T_m^{-1} Q_m^T + Q_m Q_m^T = I.\end{aligned}$$

□

Nevýhoda předpodmiňovače (410) 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 189) a matice (410) nemá požadované vlastnosti. Abychom se těmto potížím vyhnuli, museli bychom použitý předpodmiňovač zahrnout do výrazu (410) (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 (410) a ten se použije v dalších krocích metody sdružených gradientů(také se lze vrátit na začátek iteračního procesu).

Závěrem je třeba se zmínit o způsobu, který nám dovolí rozhodnout, zda máme předpodmiňovač použít nebo odvrhnout (ne vždy je nalezený předpodmiňovač vhodný k použití). To se týká hlavně pásových předpodmiňovačů určených metodou BFGS nebo numerickým derivováním. Předně 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 6.6. 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 odvrhneme. 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 256 Závěrem uvedeme několik poznámek ke konstrukci předpodmiňovačů pro diferenční verzi Newtonovy metody v případě, že neznáme Hessovu matici.

- 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 ekvivalentní předpodmíněné metodě sdružených gradientů je třeba předem upravit, jinak jsou při určování Gillova–Murrayova rozkladu většinou odvrhnuty. Velmi se osvědčily úpravy založené na větě 138, kdy se nevhodné mimodiagonální prvky zmenšují tak, aby se záporné determinanty matic (399) vynulovaly, a úpravy založené na větě 139, kdy se indefinitní matice (402) upravují tak, že se prvky β_i, β_{i+1} zmenší na dvě třetiny a prvek γ_i na třetinu. Použití věty 140 a vzorce (404) je méně výhodné. Ukazuje se, že takto získané předpodmiňovače je třeba častěji odvrhovat (například volbou $\delta = 10^{-2}$).
- Pásové předpodmiňovače určené numerickým derivováním stačí upravit nanejvýš tak, že diagonální prvky nahradíme jejich absolutními hodnotami. Úpravy založené na větě 138 a na větě 139 snižují efektivitu předpodmínění. Pro odvrhování stačí většinou volba $\delta = 10^{-12}$ (kromě diagonálních předpodmiňovačů, které jsou citlivé na odvrhování).

- Předpodmiňovače získané Lanczosovou metodou není třeba korigovat (jsou podle věty 146 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).

8.5 Numerické porovnání

Metody popsané v této kapitole byly testovány pomocí souboru 71 testovacích úloh s 1000 proměnnými. Výsledky testů jsou uvedeny ve dvou tabulkách, kde `NIT` je celkový počet iterací, `NFV` je celkový počet použitých funkčních hodnot a `NFG` je celkový počet použitých gradientů. Druhá tabulka obsahuje další hodnoty, přičemž `NCG` je celkový počet vnitřních iterací (iterací metody sdružených gradientů), `NIC` je celkový počet předpodmíněných vnějších iterací (iterací Newtonovy metody) a `NCP` je počet problémů, pro které bylo nutné zvýšit mez pro odvrhování (poznámka 256). V obou tabulkách je navíc uveden celkový čas výpočtu.

V první tabulce jsou uvedeny výsledky získané metodami `VBFGS-1` (algoritmus 11), `VBFGS-2` (algoritmus 12), `VBFGS-3` (algoritmus 13), `MBFGS-1` (algoritmus 14), `MBFGS-2` (algoritmus 15), `RBFGS` (algoritmus 17 upravený podle poznámek 238, a 238), `SVDBC-1` (algoritmus 18 s volbou (390)) a `SVDBC-2` (algoritmus 18 s volbou (387)) a (391)–(392)). Pro srovnání jsou uvedeny výsledky získané metodou sdružených gradientů (algoritmus 2). `CG-1` značí metodu sdružených gradientů se speciálním výběrem délky kroku, která dává srovnatelně přesné výsledky jako metody s proměnnou metrikou s omezenou pamětí a `CG-2` značí metodu sdružených gradientů se standardním výběrem délky kroku, která dává méně přesné výsledky.

Metoda	NIT	NFV	NFG	čas
`VBFGS-1`	121314	127189	127189	39.55
`VBFGS-2`	119665	130650	130650	37.95
`VBFGS-3`	115722	122083	122083	37.44
`MBFGS-1`	122297	128251	128251	38.90
`MBFGS-2`	119056	124680	124680	36.34
`RBFGS`	123223	142535	142535	45.34
`SVDBC-1`	134838	138172	138172	57.52
`SVDBC-2`	129313	132118	132118	52.81
`CG-1`	109166	325994	325994	75.72
`CG-2`	137846	207626	207626	48.38

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 $\overline{m} = 5$, zatímco metody redukovaných Hessiánů a posunuté metody s proměnnou metrikou vyžadují více aktualizací (obvykle $\overline{m} = 10$) a proto jsou pomalejší.

- Metodu BFGS realizovanou pomocí Strangových rekurencí (algoritmus 11) lze překonat vhodnými úpravami (algoritmus 12, algoritmus 13)).

- Metody s proměnnou metrikou s omezenou pamětí realizované pomocí maticových reprezentací jsou velmi efektivní. Rekurzivní postup založený na větě 124 se zdá být stabilnější, než použití explicitního vzorce (351).

- Metody s proměnnou metrikou s omezenou pamětí jsou účinnější než metoda sdružených gradientů (co se týče přesnosti výsledků, je s metodami proměnnou metrikou srovnatelná pouze metoda `GG-1`).

Ve druhé tabulce jsou uvedeny výsledky získané diskrétní verzí Newtonovy metody (spádových směrů) s různými předpodmiňovači. Zde `LSTN` je nepředpodmíněná Newtonova metoda, `LSTNB-1` a `LSTNB-2` jsou metody předpodmíněné třemi BFGS aktualizacemi realizovanými pomocí Strangových rekurencí (316)–(317) a pomocí věty 124, `LSTNV-1`, `LSTNV-2` a `LSTNV-3` jsou metody s pásovými předpodmiňovači (diagonálním, tridiagonálním a pentadiagonálním), určenými metodou BFGS ekvivalentní předpodmíněné metodě sdružených gradientů s korekcemi popsanými v poznámce 256, `LSTND-1`, `LSTND-2` a `LSTND-3` jsou metody s pásovými předpodmiňovači (diagonálním, tridiagonálním a pentadiagonálním), určenými numerickým derivováním s náhradou diagonálních prvků jejich absolutními hodnotami a `LSTNL` je metoda s předpodmiňovačem získaným symetrickým Lanczosovým procesem.

Metoda	NIT	NFV	NFG	NCG	NIP	NCP	čas
`LSTN`	7425	11827	372789	359505	-	-	66.08
`LSTNB-1`	7270	12521	233269	219347	7270	-	42.55
`LSTNB-2`	7327	12832	225982	211668	7327	-	37.81
`LSTNV-1`	7095	10303	274344	262855	4335	37	50.43
`LSTNV-2`	6751	9252	139989	129933	4260	37	27.47
`LSTNV-3`	6803	8857	229501	219820	4027	36	51.67
`LSTND-1`	6522	8491	347384	331709	3857	40	59.51
`LSTND-2`	7573	11245	147391	119434	4409	3	25.45
`LSTND-3`	7107	10726	125262	91665	4943	4	24.57
`LSTNL`	7398	11672	352199	339081	6808	1	55.61

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 138 a 139. Často je také třeba zvýšit mez pro odvrhování 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.

9 Metody pro rozsáhlé řídké úlohy

9.1 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

$$G(x)e_i \approx \frac{g(x+\delta e_i) - g(x)}{\delta}, \qquad 1 \leq i \leq n,$$

kde δ je malá diference ($\delta = \sqrt{\varepsilon_M}$). 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:

$$G = \begin{bmatrix} G_{11}, & G_{12}, & 0, & 0, & 0 \\ G_{21}, & G_{22}, & G_{23}, & 0, & 0 \\ 0, & G_{32}, & G_{33}, & G_{34}, & 0 \\ 0, & 0, & G_{43}, & G_{44}, & G_{45} \\ 0, & 0, & 0, & G_{54}, & G_{55} \end{bmatrix}. \tag{414}$$

Nechť

$$\begin{aligned} v_1 &= [1,0,0,1,0]^T, \\ v_2 &= [0,1,0,0,1]^T, \\ v_3 &= [0,0,1,0,0]^T. \end{aligned}$$

Pak platí

$$\begin{aligned} Gv_1 &= [G_{11}, G_{21}, G_{34}, G_{44}, G_{54}]^T, \\ Gv_2 &= [G_{12}, G_{22}, G_{32}, G_{45}, G_{55}]^T, \\ Gv_3 &= [0, G_{23}, G_{33}, G_{43}, 0]^T, \end{aligned}$$

takže všechny prvky matice G můžeme určit pomocí tří diferenčních vzorců

$$\begin{aligned} \frac{g(x+\delta v_1) - g(x)}{\delta} &\approx Gv_1, \\ \frac{g(x+\delta v_2) - g(x)}{\delta} &\approx Gv_2, \\ \frac{g(x+\delta v_3) - g(x)}{\delta} &\approx Gv_3. \end{aligned}$$

Postup, který jsme použili v tomto konkrétním případě můžeme snadno zobecnit. Rozdělme sloupce matice G do k disjunktních skupin $\mathcal{S}_i \subset \{1,\ldots,n\}$, $1 \le i \le k$, 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. Pak můžeme všechny sloupce matice G určit pomocí k diferencí

$$\frac{g(x+\delta v_i) - g(x)}{\delta} \approx Gv_i, \qquad 1 \le i \le k,$$

kde v_i, $1 \le i \le k$, jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(pomocí vektoru v_i určíme prvky submatice $G(\mathcal{S}_i)$). Takto lze postupovat pro libovolnou (i nesymetrickou) matici G. Je-li matice G symetrická, můžeme její symetrii využít k dalšímu snížení počtu potřebných diferencí. Uvažujme matici

$$G = \begin{bmatrix} G_{11}, & G_{12}, & G_{13}, & G_{14}, & G_{15} \\ G_{21}, & G_{22}, & 0, & 0, & 0 \\ G_{31}, & 0, & G_{33}, & 0, & 0 \\ G_{41}, & 0, & 0, & G_{44}, & 0 \\ G_{51}, & 0, & 0, & 0, & G_{55} \end{bmatrix}. \tag{415}$$

Použijeme-li předchozí postup, potřebujeme k určení prvků matice G pět diferencí gradientů. Položíme-li však

$$\begin{aligned} v_1 &= [1,0,0,0,0]^T, \\ v_2 &= [0,1,1,1,1]^T, \end{aligned}$$

platí

$$\begin{aligned} Gv_1 &= [G_{11}, G_{21}, G_{31}, G_{41}, G_{51}]^T, \\ Gv_2 &= [*, G_{22}, G_{33}, G_{44}, G_{55}]^T, \end{aligned}$$

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. Sloupce matice G rozdělíme opět do k disjunktních skupin $\mathcal{S}_i$, $1 \leq i \leq k$. 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 $j \in S_1 \cup \ldots \cup S_{j-1}$, a nechť $G_i(\mathcal{S}_i)$ je submatice matice G_i, která obsahuje sloupce této matice s indexy $j \in \mathcal{S}_i$, takže $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$:

Rozdělíme-li sloupce matice G do k disjunktních skupin $\mathcal{S}_i$, $i \in [1,k]$, tak aby submatice $G_i(\mathcal{S}_i)$, $i \in [1,k]$, měly v každém řádku nanejvýš jeden nenulový prvek, můžeme sloupce matice G určit pomocí k diferencí

$$\frac{g(x + \delta v_i) - g(x)}{\delta} \approx Gv_i, \qquad 1 \leq i \leq k,$$

kde v_i, $1 \leq i \leq k$, jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(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).

Zatím jsme se nezabývali určováním skupin $\mathcal{S}_i$, $1 \leq i \leq k$. Je účelné volit tyto skupiny tak, aby jejich počet byl minimální. To je však složitý kombinatorický problém, který je ekvivalentní s problémem barvení jistého grafu. V praxi se obvykle používají jednoduché a dostatečně rychlé algoritmy, které najdou dostatečně malý (i když ne minimální) počet skupin. Při určování skupin $\mathcal{S}_i$, $1 \leq i \leq k$, se používá sekvenční postup. Sloupce submatice G_i se nejprve přerovnají podle nějakého pravidla a potom se probírají postupně podle vzrůstajících indexů. Index $j \in \{1, \ldots, n\} \backslash (S_1 \cup \ldots \cup S_{i-1})$ 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.

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.

Zatím jsme se zabývali přímými metodami pro výpočet prvku řídké Hessovy matice pomocí diferencí. Nyní obrátíme pozornost na substituční metody, které obvykle vyžadují menší počet diferencí než přímé metody. Uvažujme opět matici (414) a položme

$$
\begin{aligned}
v_1 &= [1,0,1,0,1]^T, \\
v_2 &= [0,1,0,1,0]^T.
\end{aligned}
$$

Pak platí

$$
\frac{g(x+\delta v_1)-g(x)}{\delta} \approx Gv_1 = \begin{bmatrix} G_{11} \\ G_{21}+G_{23} \\ G_{33} \\ G_{43}+G_{45} \\ G_{55} \end{bmatrix},
$$

$$
\frac{g(x+\delta v_2)-g(x)}{\delta} \approx Gv_2 = \begin{bmatrix} G_{12} \\ G_{22} \\ G_{32}+G_{34} \\ G_{44} \\ G_{54} \end{bmatrix}.
$$

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_{45}. Dosadíme-li hodnoty G_{21}, G_{45} zpět do uvedených rovnic, určíme hodnoty G_{23}, G_{43} a protože matice G je symetrická i hodnoty G_{32}, G_{34}. Potřebujeme k tomu pouze dvě diference gradientů (přímá metoda používá tři diference gradientů).

Postup, který jsme použili v tomto konkrétním případě můžeme snadno zobecnit. 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 $\mathcal{S}_i \subset \{1,\ldots,n\}$, $1 \le i \le k$, 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

$$
\frac{g(x+\delta v_i)-g(x)}{\delta} \approx Gv_i, \qquad 1 \le i \le k,
$$

kde v_i, $1 \le i \le k$ jsou vektory obsahující pouze nuly a jednotky takové, že

$$
(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i
$$

(pomocí vektoru v_i určíme prvky submatice $G_U(\mathcal{S}_i)$, ostatní prvky submatice $G(\mathcal{S}_i)$ jsou určeny symetrií matice G). Při určování prvků matice G je nutné postupovat podle vzrůstajících indexů:

Určujeme-li prvky v j-tém řádku matice G_U, je nutné od prvku označeného kroužkem odečíst prvky označené křížkem, jež se v důsledku symetrie rovnají již určeným prvkům ležícím v j-tém sloupci matice G_U.

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 5) nebo jako nepřesné metody s lokálně omezeným krokem (algoritmus 6). Metody s optimálním lokálně omezeným krokem vyžadují opakované řešení soustavy rovnic $(G+\lambda I)s + g = 0$ (pro různé hodnoty parametru $\lambda \ge 0$). 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 složité postupy, které minimalizují počet nově vzniklých nenulových prvků. Tyto postupy mají kombinatorický charakter a jejich popis se vymyká rozsahu této práce (jsou jim věnovány samostatné monografie). Výpočet prvků horní trojúhelníkové matice R (numerická faktorizace) se provádí podle vzorců uvedených v oddílu 6.6.

Nepřesné metody s lokálně omezeným krokem používají metodu sdružených gradientů popsanou v oddílu 3.6 (Algoritmus 3), 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 trojúhelníková část matice B a aproximace $RR^T \approx B$ je často velmi dobrá, což dává velmi účinné předpodmínění.

Nyní ukážeme, jak lze reprezentovat řídkou Hessovu matice G. Budeme přitom pracovat s horní trojúhelníkovou maticí G_U, která vznikne z matice G vynulováním všech poddiagonálních prvků.

Definice 38 *Řídkou reprezentací Hessovy matice G nazveme trojici vektorů $num(G_U) \in R^{m_U}$, $ind(G_U) \in R^{m_U}$, $adr(G_U) \in R^{n+1}$, kde m_U je počet nenulových prvků matice G_U. Vektor $num(G_U)$ obsahuje numerické hodnoty nenulových prvků matice G_U uspořádaných po řádcích. Vektor $ind(G_U)$ obsahuje sloupcové indexy těchto nenulových prvků. Vektor $adr(G_U)$ obsahuje ukazatele umístění diagonálních prvků matice G_U ve vektorech $num(G_U)$ a $ind(G_U)$. Poslední prvek vektoru $adr(G_U)$ (s indexem $n+1$) má hodnotu $m_U + 1$.*

Pro matici (414) platí

$$\begin{aligned} num(G_U) &= [G_{11}, G_{12}, G_{22}, G_{23}, G_{33}, G_{34}, G_{44}, G_{45}, G_{55}]^T, \\ ind(G_U) &= [1, 2, 2, 3, 3, 4, 4, 5, 5]^T, \\ adr(G_U) &= [1, 3, 5, 7, 9, 10]^T. \end{aligned}$$

Pro matici (415) platí

$$\begin{aligned} num(G_U) &= [G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{22}, G_{33}, G_{44}, G_{55}]^T, \\ ind(G_U) &= [1, 2, 3, 4, 5, 2, 3, 4, 5]^T, \\ adr(G_U) &= [1, 6, 7, 8, 9, 10]^T. \end{aligned}$$

9.2 Metody s proměnnou metrikou pro řídké úlohy

Metody s proměnnou metrikou pro řídké úlohy používají aktualizace, které zachovávají strukturu řídké Hessovy matice. Toto zachovávání 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 aktualizace splňovaly kvazinewtonovskou podmínku, neporušovaly symetrii a zachovávaly strukturu řídké Hessovy matice. Označme

$$\begin{aligned} \mathcal{V}_Q &= \{B \in R^{n \times n} : Bd = y\}, \\ \mathcal{V}_S &= \{B \in R^{n \times n} : B^T = B\}, \\ \mathcal{V}_G &= \{B \in R^{n \times n} : G_{ij} = 0 \Rightarrow B_{ij} = 0\} \end{aligned}$$

(předpokládáme, že $G_{ii} \neq 0 \ \forall 1 \leq i \leq n$). 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 $R^{n\times n}$, 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 $\mathcal{V}_Q \cap \mathcal{V}_S$, $\mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{V}_S \cap \mathcal{V}_G$ a $\mathcal{V}_Q \cap \mathcal{V}_S \cap \mathcal{V}_G$. 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 147 *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 Budeme postupovat podobně jako v důkazu věty 55. Vztah pro $\mathcal{P}_Q B$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned}
L &= \frac{1}{2}\|\tilde{B} - B\|_F^2 + u^T(y - \tilde{B}d) \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\tilde{B}_{ij} - B_{ij})^2 + \sum_{i=1}^{n} u_i \left(y_i - \sum_{j=1}^{n} \tilde{B}_{ij} d_j \right).
\end{aligned}$$

Derivováním Lagrangeovy funkce podle prvků matice $\tilde{B}$ dostaneme

$$\frac{\partial L}{\partial \tilde{B}_{ij}} = (\tilde{B}_{ij} - B_{ij}) - u_i d_j, \qquad 1 \leq i \leq n, \quad 1 \leq j \leq n.$$

Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $\tilde{B} = B + ud^T$, což po dosazení do kvazinewtonovské podmínky $\tilde{B}d = y$ dává $ud^T d = y - Bd$, neboli

$$u = \frac{y - Bd}{d^T d}.$$

Dosadíme-li tento výraz do vzorce $\tilde{B} = B + ud^T$, dostaneme vztah pro $\mathcal{P}_Q B$. Vztah pro $\mathcal{P}_S B$ odvodíme minimalizací funkce

$$\frac{1}{2}\left(\|\tilde{B} - B\|_F^2 + \|\tilde{B} - B^T\|_F^2\right) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left((\tilde{B}_{ij} - B_{ij})^2 + (\tilde{B}_{ij} - B_{ji})^2\right).$$

Derivujeme-li tuto funkci podle prvků matice $\tilde{B}$, a položíme-li derivace rovny nule, dostaneme

$$(\tilde{B}_{ij} - B_{ij}) + (\tilde{B}_{ij} - B_{ji}) = 0, \qquad 1 \leq i \leq n, \quad 1 \leq j \leq n,$$

což dává $\tilde{B} = (B + B^T)/2$. Vztah pro $\mathcal{P}_G B$ odvodíme minimalizací funkce

$$\frac{1}{2}\|\tilde{B} - B\|_F^2 = \frac{1}{2}\sum_{G_{ij}\neq 0}(\tilde{B}_{ij} - B_{ij})^2 + \frac{1}{2}\sum_{G_{ij}=0} B_{ij}^2,$$

neboť podle předpokladu platí $\tilde{B}_{ij} = 0$ pokud $G_{ij} = 0$. Derivujeme-li tuto funkci podle prvků matice $\tilde{B}$ a položíme-li derivace rovny nule dostaneme

$$\begin{aligned} \tilde{B}_{ij} - B_{ij} &= 0, \quad G_{ij} \neq 0, \\ \tilde{B}_{ij} &= 0, \quad G_{ij} = 0, \end{aligned}$$

kde $1 \leq i \leq n$, $1 \leq j \leq n$, což jsme měli dokázat. □

Podmínku $\tilde{B} \in \mathcal{V}_G$ lze snadno splnit tím, že položíme $\tilde{B}_{ij} = 0$, pokud $G_{ij} = 0$. Abychom nemuseli tuto podmínku přidávat k Lagrangeově funkci, budeme používat řídkou kvazinewtonovskou podmínku. Nechť $d^i \in R^n$, $1 \leq i \leq n$, jsou vektory takové, že

$$\begin{aligned} d_j^i &= d_j, \quad G_{ij} \neq 0, \\ d_j^i &= 0, \quad G_{ij} = 0, \end{aligned}$$

kde $1 \leq j \leq n$. Pak řídkou kvazinewtonovskou podmínku lze zapsat ve tvaru

$$\sum_{j=1}^{n}(\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij})d_j^i = v_i, \qquad 1 \leq i \leq n,$$

kde $v = y - (\mathcal{P}_G B)d$.

Věta 148 *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*

$$\begin{aligned} \mathcal{P}_{QS}B &= B + \frac{(y - Bd)d^T + d(y - Bd)^T}{d^T d} - \frac{(y - Bd)^T d}{d^T d}\,\frac{dd^T}{d^T d}, \\ \mathcal{P}_{QG}B &= \mathcal{P}_G(B + ud^T), \\ \mathcal{P}_{SG}B &= \mathcal{P}_S\mathcal{P}_G B = \mathcal{P}_G\mathcal{P}_S B, \end{aligned}$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = v$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n}\|d^i\|^2 e_i e_i^T.$$

Důkaz Vztah pro $\mathcal{P}_{QS}B$ plyne bezprostředně z věty 55. Stačí dosadit $W = I$ (metoda PSB). Zřejmě platí $\mathcal{P}_{QG}B = \mathcal{P}_{QG}\mathcal{P}_G B$. Vztah pro $\mathcal{P}_{QG}\mathcal{P}_G B$ odvodíme pomocí Lagrangeovy funkce

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij})^2 + \sum_{i=1}^{n} u_i\left(v_i - \sum_{j=1}^{n}(\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij})d_j^i\right),$$

obsahující řídkou kvazinewtonovskou podmínku. Derivováním Lagrangeovy funkce podle prvků matice $\tilde{B}$ dostaneme

$$\frac{\partial L}{\partial \tilde{B}_{ij}} = (\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij}) - u_i d_j^i \qquad 1 \leq i \leq n, \quad 1 \leq j \leq n.$$

Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij} = u_i d^i_j$, $1 \le i \le n$, $1 \le j \le n$, což jsme měli dokázat. Dosadíme-li toto vyjádření do řídké kvazinewtonovské podmínky, dostaneme

$$v_i = \sum_{j=1}^{n} u_i d^i_j d^i_j = u_i \|d^i\|^2, \qquad 1 \le i \le n,$$

neboli $Qu = v$, kde Q je diagonální matice vystupující v tvrzení věty (pozitivní semidefinitnost je zřejmá). Vztahy pro $\mathcal{P}_{SG}B$ plynou bezprostředně z identit $\mathcal{P}_{SG}B = \mathcal{P}_{SG}(\mathcal{P}_S B)$ a $\mathcal{P}_{SG}B = \mathcal{P}_{SG}(\mathcal{P}_G B)$. □

Věta 149 *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*

$$\mathcal{P}_{QSG}B = \mathcal{P}_G(B + ud^T + du^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = v$ 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 Zřejmě platí $\mathcal{P}_{QSG}B = \mathcal{P}_{QSG}\mathcal{P}_G B$. Jelikož matice $\tilde{B} - \mathcal{P}_G B$ je symetrická, můžeme položit $\tilde{B} - \mathcal{P}_G B = X + X^T$, kde X je zatím neznámá čtvercová matice. Použijeme Lagrangeovu funkci

$$L = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij} + X_{ji})^2 + \sum_{i=1}^{n} u_i \left(v_i - \sum_{j=1}^{n}(X_{ij} + X_{ji})d^i_j \right),$$

obsahující řídkou kvazinewtonovskou podmínku. Derivováním Lagrangeovy funkce podle prvků matice X dostaneme

$$\frac{\partial L}{\partial X_{ij}} = (X_{ij} + X_{ji}) - u_i d^i_j - u_j d^j_i, \qquad 1 \le i \le n, \quad 1 \le j \le n.$$

Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $\tilde{B}_{ij} - (\mathcal{P}_G B)_{ij} = u_i d^i_j + u_j d^j_i$, což jsme měli dokázat. Dosadíme-li toto vyjádření do řídké kvazinewtonovské podmínky, dostaneme

$$v_i = \sum_{j=1}^{n}(u_i d^i_j + u_j d^j_i)d^i_j = \|d^i\|^2 u_i + \sum_{j=1}^{n} d^j_i d^i_j u_j,$$

neboli $Qu = v$, kde Q je symetrická matice vystupující v tvrzení věty. Matice Q má zřejmě stejnou strukturu jako matice G. Nechť $z \in R^n$ je libovolný vektor. Pak platí

$$z^T Q z = \sum_{i=1}^{n}\sum_{j=1}^{n} d^j_i d^i_j z_i z_j + \sum_{i=1}^{n} \|d^i\|^2 z_i^2 = \sum_{G_{ij}\neq 0} d_i d_j z_i z_j + \sum_{G_{ij}\neq 0} d_j^2 z_i^2 = \frac{1}{2}\sum_{G_{ij}\neq 0}(d_i z_j + d_j z_i)^2 \ge 0 \quad (416)$$

(matice G_{ij} je symetrická), takže matice Q je pozitivně semidefinitní. Zbývá dokázat, že rovnice $Qu = v$ má řešení. Předpokládejme nejprve, že $\|d^i\| \neq 0$ $\forall 1 \le i \le n$. Ukážeme, že v tomto případě je matice Q pozitivně definitní. Kdyby matice Q nebyla pozitivně definitní, existoval by vektor $z \neq 0$ takový, že $z^T Q z = 0$. Pak by podle vyjádření (416) musel existovat index $1 \le i \le n$ takový, že $z_i \neq 0$ a

$$d_i z_j + d_j z_i = 0, \qquad G_{ij} \neq 0.$$

Jelikož předpokládáme, že $G_{ii} \neq 0$ muselo by nutně platit $d_i z_i = 0$, neboli $d_i = 0$, což po dosazení do poslední rovnosti dává $d_j z_i = 0$ $\forall G_{ij} \neq 0$, neboli $d_j = 0$ $\forall G_{ij} \neq 0$. To je ale ve sporu s předpokladem, že

$$\|d^i\|^2 = \sum_{j=1}^{n}(d^i_j)^2 = \sum_{G_{ij}\neq 0} d_j^2 \neq 0.$$

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

$$v_i = y_i - \sum_{G_{ij} \neq 0} B_{ij} d_j = \sum_{G_{ij} \neq 0} (\tilde{G}_{ij} - B_{ij}) d_j^i = 0$$

(matice $\tilde{G}$ je definovaná vztahem (184)). 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. □

Metoda s proměnnou metrikou, která používá aktualizaci

$$B^+ = \mathcal{P}_{QSG} B \tag{417}$$

se nazývá Tointovou metodou. Její realizace je poměrně pracná, neboť je třeba řešit dodatečnou soustavu rovnic $Qu = v$. V případě, že matice B je hustá, je tato metoda ekvivalentní metodě PSB, která není příliš efektivní. Proto byly navrženy další aktualizace, které však v jistém smyslu narušují splnění kvazinewtonovské podmínky. V této práci se budeme zabývat Marwilovou metodou s aktualizací

$$B^+ = \mathcal{P}_S \mathcal{P}_{QG} B, \tag{418}$$

Powellovou metodou s aktualizací

$$B^+ = \mathcal{P}_G \mathcal{P}_{QS} B, \tag{419}$$

a Steihaugovou metodou s aktualizací

$$B^+ = \mathcal{P}_{SG} \mathcal{P}_Q B. \tag{420}$$

Lemma 47 *Nechť B^+ je matice určená pomocí některé z aktualizací (417)–(420). Pak platí*

$$B^+ \in \mathcal{V}_S \cap \mathcal{V}_G,$$

kde $\tilde{G}$ je matice definovaná vztahem (184).

Důkaz Pro aktualizaci (417) a (420) je toto tvrzení zřejmé. V případě aktualizace (418) tvrzení plyne z toho, že projekce $\mathcal{P}_S$, určená symetrií matice, neovlivní symetrickou řídkou strukturu. V případě aktualizace (419) tvrzení plyne z toho, že projekce $\mathcal{P}_G$, určená symetrickou řídkou strukturou neovlivní symetrii matice. □

Ve vzorcích (417)–(420) 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 48 *Nechť $B^+ = \mathcal{P}_B \mathcal{P}_A B$, kde $\mathcal{P}_A, \mathcal{P}_B$ jsou operátory ortogonální projekce do $\mathcal{V}_A$, $\mathcal{V}_B$, kde $\mathcal{V}_A \subset R^{n\times n}$, $\mathcal{V}_B \subset R^{n\times n}$ 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

$$\|\mathcal{P}_B \mathcal{P}_A B - \tilde{G}\|_F^2 = \|\mathcal{P}_A B - \tilde{G}\|_F^2 - \|\mathcal{P}_A B - \mathcal{P}_B \mathcal{P}_A B\|_F^2 \leq \|\mathcal{P}_A B - \tilde{G}\|_F^2.$$

Jelikož $\mathcal{P}_A B \in \mathcal{V}_A \subset \mathcal{V}_Q$, můžeme psát $\mathcal{P}_A B d = y$, takže platí

$$\|y - Bd\| = \|(\mathcal{P}_A B - B)d\| \leq \|\mathcal{P}_A B - B\| \|d\| \leq \|\mathcal{P}_A B - B\|_F \|d\|.$$

Jelikož $\tilde{G} \in \mathcal{V}_A$ a $\mathcal{P}_A$ je operátor ortogonální projekce, můžeme psát

$$\|\mathcal{P}_A B - \tilde{G}\|_F^2 = \|B - \tilde{G}\|_F^2 - \|B - \mathcal{P}_A B\|_F^2.$$

Spojením všech dokázaných nerovností dostaneme tvrzení lemmatu. □

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 150 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s lokálně omezeným krokem (T1)–(T3) (definice 25), pro kterou platí (246). 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 (417)–(420) a množiny N_1, N_2, N_3 jsou definovány v poznámce 166). Pak jestliže funkce $F : R^n \to R$ splňuje podmínky (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 $\|B_i\| \le C_i$, $i \in N$, kde C_i, $i \in N$, jsou čísla splňující rekurentní nerovnosti

$$C_{i+1} \le C_i + \overline{C}\|d_i\| \le C_i + \overline{C}\|s_i\|. \tag{421}$$

Nechť $i \in N_2$ a nechť $\tilde{G}_i$ je matice definovaná vztahem (184). Pak platí

$$\begin{aligned} \|\tilde{G}_i - G_i\|_F &= \left\| \int_0^1 (G(x_i + \lambda d_i) - G(x_i))d\lambda \right\|_F \le \sqrt{n} \int_0^1 \|G(x_i + \lambda d_i) - G(x_i)\|d\lambda \\ &\le \overline{L}\sqrt{n}\|d_i\| \int_0^1 \lambda d\lambda = \frac{1}{2}\overline{L}\sqrt{n}\|d_i\| \end{aligned} \tag{422}$$

(používáme předpoklad (F6) a skutečnost, že Frobeniova norma není větší než $\sqrt{n}$ násobek spektrální normy). Podobným způsobem dostaneme

$$\|\tilde{G}_i - G_{i+1}\|_F \le \frac{1}{2}\overline{L}\sqrt{n}\|d_i\|. \tag{423}$$

Použijeme-li nerovnost $\|B_{i+1} - \tilde{G}_i\|_F \le \|B_i - \tilde{G}_i\|_F$, která plyne z lemmatu 48, můžeme podle (422) a (423) psát

$$\begin{aligned} \|B_{i+1} - G_{i+1}\|_F &\le \|B_{i+1} - \tilde{G}_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \le \|B_i - \tilde{G}_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \\ &\le \|B_i - G_i\|_F + \|\tilde{G}_i - G_i\|_F + \|\tilde{G}_i - G_{i+1}\|_F \\ &\le \|B_i - G_i\|_F + \overline{L}\sqrt{n}\|d_i\|. \end{aligned}$$

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

$$\|B_{i+1} - G_{i+1}\|_F \le \|B_1 - G_1\|_F + \overline{L}\sqrt{n}\sum_{j=1}^{i}\|d_j\|.$$

neboli

$$\|B_{i+1}\| \le \|B_{i+1}\|_F \le 2\overline{G}\sqrt{n} + \|B_1\|\sqrt{n} + \overline{L}\sqrt{n}\sum_{j=1}^{i}\|d_j\|.$$

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 (421) (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 168 dostaneme

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty,$$

takže můžeme použít větu 75. □

Věta 151 *Nechť jsou splněny předpoklady věty 150 a nechť $x_i \to x^*$ a $\|\omega_i(s_i)\| \to 0$. Jestliže funkce $F : R^n \to R$ splňuje podmínky (F4), (F5) a (F6), pak $x_i \to x^*$ Q-superlineárně.*

Důkaz Nechť $i \in N_2$. Použijeme-li lemma 48 a nerovnosti (422), (423), 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 77 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 77, dostaneme

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{\|y_i - B_i d_i\|^2}{\|d_i\|^2} &\leq \sum_{i=1}^{\infty} \left(\|B_i - G_i\|_F^2 - \|B_{i+1} - G_{i+1}\|_F^2\right) + \overline{M}\sum_{i=1}^{\infty}\|d_i\| \\
&\leq \|B_1 - G_1\|_F^2 + \overline{M}\sum_{i=1}^{\infty}\|d_i\| \leq \|B_1 - G_1\|_F^2 + \overline{M}\sum_{i=1}^{\infty}\|s_i\| < \infty. \qquad (424)
\end{aligned}$$

Dále podle (F5) a (422) 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 77 a $\|y_i - B_i d_i\|/\|d_i\| \to 0$ podle (424). Jelikož $\|\omega_i(s_i)\| \to 0$ jsou splněny předpoklady věty 79 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 3). 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 39). 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.

Metody popsané v této kapitole byly testovány pomocí souboru 71 řídkých úloh s 1000 proměnnými. Jsou to tytéž úlohy, které byly použity v kaptole 8. Nyní se však využívá struktura řídkosti řídkých Hessových matic. Výsledky testů jsou uvedeny v tabulce, kde `NIT` je celkový počet iterací, `NFV` je celkový počet použitých funkčních hodnot a `NFG` je celkový počet použitých gradientů. V tabulce jsou uvedeny výsledky získané metodami `TRNMS-5`–`TRNMS-7` (diferenční verze Newtonovy metody pro řídké úlohy realizovaná jako metoda s lokálně omezeným krokem pomocí algoritmu 5–algoritmu 7) a `TRVMS-6` (metoda s proměnnou metrikou pro řídké úlohy s Marwilovou projekcí realizovaná jako metoda s lokálně omezeným krokem pomocí algoritmu 6). Pro srovnání jsou uvedeny výsledky získané metodami `LSTND-2`, `MBFGS-2`, `CG-1` testovanými v oddílu 8.5.

Metoda	NIT	NFV	NFG	NCG	čas
`TRNMS-5`	6375	6609	32193	–	11.78
`TRNMS-6`	8014	8760	45063	–	13.19
`TRNMS-7`	8983	9582	51044	58200	14.47
`TRVMS-6`	58772	69021	58842	–	73.84
`LSTND-2`	7573	11245	147391	89144	25.45
`MBFGS-2`	119056	124680	124680	–	36.34
`CG-1`	109166	325994	325994	–	75.72

10 Metody pro rozsáhlé separovatelné úlohy

10.1 Diferenční verze Newtonovy metody pro separovatelné úlohy

Rozsáhlé úlohy jsou často formulovány tak, že platí

$$F(x) = \sum_{k=1}^{m} f_k(x), \tag{425}$$

kde $m = O(n)$ a kde každá z funkcí $f_k : R^n \to R$, $1 \le k \le m$, závisí na $n_k = O(1)$ proměnných. Pak výpočet hodnoty a gradientu funkce $F(x)$ spotřebuje $O(n)$ operací a Hessova matice této funkce obsahuje $O(n)$ nenulových prvků. Gradient a Hessovu matici funkce $F : R^n \to R$ můžeme vyjádřit ve tvaru

$$g(x) = \sum_{k=1}^{m} g_k(x),$$

$$G(x) = \sum_{k=1}^{m} G_k(x),$$

kde gradienty $g_k(x)$ a Hessovy matice $G_k(x)$ funkcí $f_k : R^n \to R$, $1 \le k \le m$, obsahují $O(1)$ nenulových prvků, takže je lze uchovávat v úsporném tvaru. 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 $e = [1, \ldots, 1]^T \in R^m$ je vektor, který obsahuje samé jednotky. Jacobiova matice $J(x)$ je řídká (její k-tý řádek $g_k^T(x)$ obsahuje $n_k = O(1)$ nenulových prvků, $1 \le k \le m$). Hessova matice $G(x)$ má stejnou strukturu jako matice $J^T(x)J(x)$. Struktura řídké úlohy je tedy plně určena strukturou Jacobiovy matice.

Definice 39 *Řídkou reprezentací Jacobiovy matice J nazveme trojici vektorů num $(J) \in R^{\hat{n}}$, ind $(J) \in R^{\hat{n}}$, ord $(J) \in R^{m+1}$, kde*

$$\hat{n} = \sum_{k=1}^{m} n_k$$

je počet nenulových prvků matice J. Vektor num (J) obsahuje numerické hodnoty nenulových prvků matice J uspořádaných po řádcích. Vektor ind (J) obsahuje indexy těchto nenulových prvků. Vektor ord (J) obsahuje ukazatele umístění prvních nenulových prvků v řádcích matice J (ukazatele umístění ve vektorech num (J) a ind (J)), takže

$$adr(J)_k = 1 + \sum_{i=1}^{k-1} n_k, \qquad 1 \le k \le m+1.$$

V dalším výkladu budeme používat redukované gradienty $\hat{g}_k(x) \in R^{n_k}$, které obsahují pouze nenulové prvky gradientů $g_k(x) \in R^n$, $1 \le k \le m$, a redukované Hessovy matice $\hat{G}_k(x) \in R^{n_k \times n_k}$, které obsahují pouze nenulové prvky Hessových matic $G_k(x) \in R^{n \times n}$, $1 \le k \le m$.

Definice 40 *Nechť N_k, $1 \le k \le m$, jsou množiny indexů proměnných vystupujících ve funkcích $f_k(x)$, $1 \le k \le m$, a nechť $Z_k \in R^{n \times n_k}$ jsou matice, jejichž sloupce tvoří ortonormální báze v podprostorech určených proměnnými s indexy z N_k (jsou to sloupce jednotkové matice s indexy z N_k). Pak vektory $\hat{g}_k(x) = Z_k^T g_k(x)$, $1 \le k \le m$, nazveme redukovanými gradienty a matice $\hat{G}_k(x) = Z_k^T G_k(x) Z_k$, $1 \le k \le m$, nazveme redukovanými Hessovými maticemi funkcí $f_k(x)$, $1 \le k \le m$.*

Zřejmě platí

$$num(J) = [\hat{g}_1^T, \ldots, \hat{g}_m^T]^T.$$

Diferenční verze Newtonovy metody pro separovatelné úlohy jsou založeny na numerickém výpočtu prvků redukovaných Hessových matic. Používají se přitom diferenční vzorce

$$\hat{G}_k(x)\hat{e}_j \approx \frac{\hat{g}_k(x + \delta Z_k \hat{e}_j) - \hat{g}_k(x)}{\delta},$$

kde $\hat{e}_j$, $1 \le j \le n_k$, jsou sloupce jednotkové matice řádu n_k. K určení prvků redukovaných Hessových matic je tedy zapotřebí

$$\sum_{k=1}^{m} n_k^2 = mO(1) = O(n)$$

operací.

Poznámka 257 Redukované gradienty $\hat{g}_k(x)$ a redukované Hessovy matice $\hat{G}_k(x)$, jednoznačně určují gradient g a řídkou Hessovu matici G. Platí

$$g(x) = \sum_{k=1}^{m} Z_k \hat{g}_k(x), \quad G(x) = \sum_{k=1}^{m} Z_k \hat{G}_k(x) Z_k^T.$$

Známe-li řídkou reprezentaci Jacobiovy matice (definice 39) a numerické hodnoty redukovaných Hessových matic, můžeme snadno určit řídkou reprezentaci Hessovy matice (definice 38). Redukované Hessovy matice je možné zpracovávat sekvenčně (není třeba je ukládat současně v paměti počítače).

Diferenční verze Newtonovy metody pro separovatelné úlohy se liší od diferenčních verzí Newtonovy metody pro řídké úlohy pouze způsobem získání řídké Hessovy matice $G(x)$. 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. 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. Oba dva typy metod vyžadují přibližně stejný počet iterací.

10.2 Metody s proměnnou metrikou pro separovatelné úlohy

Metody s proměnnou metrikou pro separovatelné úlohy používají místo redukovaných Hessových matic $\hat{G}_k(x)$, $1 \leq k \leq m$, jejich aproximace $\hat{B}_k$, $1 \leq k \leq m$, které se aktualizují pomocí metod s proměnnou metrikou.

$$\hat{B}_k^+ = \frac{1}{\hat{\gamma}_k} \left(\hat{B}_k + \frac{\hat{\gamma}_k}{\hat{\rho}_k} \frac{1}{\hat{b}_k} \hat{y}_k \hat{y}_k^T - \frac{1}{\hat{c}_k} \hat{B}_k \hat{d}_k \left(\hat{B}_k \hat{d}_k \right)^T + \frac{\hat{\beta}_k}{\hat{c}_k} \left(\frac{\hat{c}_k}{\hat{b}_k} \hat{y}_k - \hat{B}_k \hat{d}_k \right) \left(\frac{\hat{c}_k}{\hat{b}_k} \hat{y}_k - \hat{B}_k \hat{d}_k \right)^T \right),$$

kde $\hat{y}_k = \hat{g}_k^+ - \hat{g}_k$ a $\hat{d}_k = Z_k^T d$ 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 $\hat{\gamma}_k$, $\hat{\rho}_k$, $\hat{\beta}_k$ jsou volné parametry.

Uvedeme nejprve několik poznámek k metodám s proměnnou metrikou pro separovatelné úlohy:

- Metody s proměnnou metrikou pro separovatelné úlohy jsou účinnější než metody s proměnnou metrikou pro řídké úlohy, jak je zřejmé z numerického porovnání uvedeného v závěru tohoto oddílu.
- Vzhledem k tomu, že redukované matice $\hat{B}_k$ se aktualizují pomocí vektorů $\hat{y}_k$, $\hat{d}_k$, je účelné aby platilo $\hat{B}_k \to \hat{G}_k$, takže se obyčejně pokládá $\hat{\gamma}_k = 1$, $\hat{\rho}_k = 1$, $1 \leq k \leq m$ (jiné volby těchto volných parametrů obyčejně zhoršují rychlost konvergence).
- Dá se dokázat, že metody s proměnnou metrikou pro separovatelné úlohy jsou Q-superlineárně konvergentní. Kupodivu obtížnější je dokázat globální konvergenci těchto metod, což se zatím bez zavedení dodatečných předpokladů nepodařilo. Souvisí to se skutečností, že není obecně zaručena platnost nerovností $\hat{y}_k^T \hat{d}_k > 0$, $1 \leq k \leq m$, takže některé z matic $\hat{B}_k^+$, $1 \leq k \leq m$, nemusí být pozitivně definitní.

Popíšeme nyní efektivní realizaci metod s proměnnou metrikou pro separovatelné úlohy. Tato realizace je metodou spádových směrů (definice 15) a aktualizace se provádí podle vzorců

$$\begin{aligned} \hat{B}_k^+ &= \hat{B}_k + \frac{1}{\hat{b}_k} \hat{y}_k \hat{y}_k^T - \frac{1}{\hat{c}_k} \hat{B}_k \hat{d}_k \left(\hat{B}_k \hat{d}_k \right)^T, \quad \hat{y}_k^T \hat{d}_k > 0, \\ \hat{B}_k^+ &= \hat{B}_k \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad , \quad \hat{y}_k^T \hat{d}_k \leq 0, \end{aligned}$$

kde $1 \leq k \leq m$ (metoda BFGS). Tyto vzorce zaručují pozitivní definitnost matic $\hat{B}_k^+$, $1 \leq k \leq m$. Je-li matice $\hat{B}_k$ pozitivně definitní a platí-li $\hat{y}_k^T \hat{d}_k \leq 0$, je matice $\hat{B}_k^+ = \hat{B}_k$ pozitivně definitní. Je-li matice $\hat{B}_k$ pozitivně definitní a platí-li $\hat{y}_k^T \hat{d}_k > 0$, je matice $\hat{B}_k^+$ pozitivně definitní podle věty 45.

Známe-li aproximace $\hat{B}_k$, $1 \leq k \leq m$ redukovaných Hessových matic $\hat{G}_k$, $1 \leq k \leq m$, můžeme podle poznámky 257 zkonstruovat řídkou aproximaci Hessovy matice G. Metody s proměnnou metrikou však mají jednu nevýhodu, která spočívá v tom, že je třeba ukládat současně všechny matice $\hat{B}_k$, $1 \leq k \leq m$. To vyžaduje rezervaci dalších

$$\hat{m} = \sum_{k=1}^{n} \frac{1}{2} \hat{n}_k (\hat{n_k} + 1)$$

míst v paměti počítače (číslo $\hat{m}$ je obvykle značně větší než počet nenulových prvků řídké Hessovy matice G).

Metody popsané v této kapitole byly testovány pomocí souboru 71 rozložitelných úloh s 1000 proměnnými. Jsou to tytéž úlohy, které byly použity v kaptole 8. Nyní se však využívá toho, že testovací úlohy lze vyjádřit ve tvaru (425). Výsledky testů jsou uvedeny v tabulce, kde `NIT` je celkový počet iterací, `NFV` je celkový počet použitých funkčních hodnot a `NFG` je celkový počet použitých gradientů. V tabulce jsou uvedeny výsledky získané metodami `TRNMS-5`–`TRNMS-7` (diferenční verze Newtonovy metody pro řídké úlohy realizovaná jako metoda s lokálně omezeným krokem pomocí algoritmu 5–algoritmu 7), `TRNMP-5`–`TRNMP-7` (diferenční verze Newtonovy metody pro rozložitelné úlohy realizovaná jako metoda s lokálně omezeným krokem pomocí algoritmu 5–algoritmu 7) a `LSVMP-1` nebo `LSVMP-1` (metoda s proměnnou metrikou pro rozložitelné úlohy realizovaná jako metoda s spádových směrů pomocí Gillova-Murrayova rozkladu nebo algoritmu 3). Pro srovnání jsou uvedeny výsledky získané metodami `LSTND-2`, `MBFGS-2`, `CG-1` testovanými v oddílu 8.5.

Metoda	NIT	NFV	NFG	NCG	čas
`TRNMS-5`	6138	6403	31994	–	19.25
`TRNMS-6`	7861	8575	46422	–	23.25
`TRNMS-7`	8510	9105	50985	53574	22.81
`TRNMP-5`	6303	6577	21485	–	44.08
`TRNMP-6`	7834	8509	28848	–	49.53
`TRNMP-7`	8777	9409	33599	54152	52.86
`LSVMP-1`	13136	20685	20685	–	24.62
`LSVMP-3`	12782	17206	17206	239518	34.27
`LSTND-2`	7388	10919	142098	114858	47.13
`MBFGS-2`	112027	117651	117651	–	83.28
`CG-1`	102434	202213	112069	–	103.66

10.3 Modifikace Gaussovy–Newtonovy metody pro řídký součet čtverců

Předpokládejme, že účelová funkce $F(x)$ má tvar

$$F(x) = \frac{1}{2} \sum_{k=1}^{m} f_k^2(x),$$

kde $m = O(n)$ a kde každá z funkcí $f_k : R^n \to R$, $1 \leq k \leq m$, závisí 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.

Gaussovu-Newtonovu (GN) metodu můžeme realizovat buď pomocí řídké reprezentace Hessovy matice (řešením normální soustavy rovnic (NE)) nebo pomocí řídké reprezentace Jacobiovy matice (řešením přeurčené soustavy rovnic (OE)). První způsob je založen na použití matice $B = J^T J$, která má stejnou strukturu jako matice G a která se snadno sestrojuje. Známe-li matici B, můžeme GN metodu realizovat buď jako metodu s optimálním lokálně omezeným krokem nebo jako nepřesnou metodu s lokálně omezeným krokem (tak jako diferenční verzi Newtonovy metody pro řídké úlohy).

Protože GN metoda může selhávat v případě úloh s velkými rezidui, je výhodné kombinovat tuto metodu s jinými metodami (oddíl 7.2). V praxi se používají tři hybridní metody pro řídký součet čtverců.

1) Kombinace GN metody s Marwilovou metodou. V prvním iteračním kroku pokládáme $B = J^T J$. Po skončení každého iteračního kroku pokládáme

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= \mathcal{P}_S \mathcal{P}_{QG} B \quad , \quad F - F_+ \leq \underline{\vartheta} F \end{aligned}$$

(viz (418) v oddílu 9.2), kde $J_+ = J(x_+)$. Globální konvergence této metody plyne z věty 105 a věty 150. Superlineární konvergence této metody plyne z věty 105 a věty 151.

2) Kombinace GN metody s diferenční verzí Newtonovy metody. V prvním iteračním kroku pokládáme $B = J^T J$. Po skončení každého iteračního kroku pokládáme

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ G_k^+ \quad , \quad F - F_+ \leq \underline{\vartheta} F, \end{aligned}$$

kde $J_+ = J(x_+)$ a $f_k^+ = f_k(x_+)$, $G_k^+ = G_k(x_+)$, $1 \leq k \leq m$ ($G_k(x_+)$ je diferenční aproximace Hessovy matice funkce $f_k(x_+)$). Globální a superlineární konvergence této metody plyne z věty 82 a věty 105.

3) Kombinace GN metody s metodou hodnosti 1. V prvním iteračním kroku pokládáme $B = J^T J$ a $B_k = I_k$, $1 \leq k \leq m$ (B_k je aproximace Hessovy matice G_k a I_k se od jednotkové matice liší pouze tím, že $(I_k)_{ii} = 0$, pokud $(G_k)_{ii} = 0$). Po skončení každého iteračního kroku pokládáme

$$\begin{aligned} B_k^+ &= B_k + \frac{w_k w_k^T}{d_k^T w_k} \quad , \quad |d_k^T w_k| > \underline{\delta}, \\ B_k^+ &= B_k \quad , \quad |d_k^T w_k| \leq \underline{\delta} \end{aligned}$$

pro $1 \leq k \leq m$, a

$$\begin{aligned} B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F, \\ B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ B_k^+ \quad , \quad F - F_+ \leq \underline{\vartheta} F. \end{aligned}$$

Přitom $w_k = y_k - B_k d_k$ a $y_k = g_k(x_+) - g_k(x)$, $d_k = x_+ - x$, $1 \leq k \leq m$. Ačkoliv pro tuto metodu nejsou dokázány konvergenční věty, jsou její numerické vlastnosti velmi dobré. Jedinou nevýhodou této metody (podobně jako metod s proměnnou metrikou pro separovatelné úlohy) je nutnost ukládat současně všechny matice B_k, $1 \leq k \leq m$ (ve skutečnosti se pracuje se redukovanými maticemi $\hat{B}_k$, $1 \leq k \leq m$).

Následující tabulka ukazuje srovnání jednotlivých hybridních metod používajících řídkou reprezentaci Hessovy matice s ostatními metodami pro řídké a separovatelné úlohy při minimalizaci 22 testovacích

funkcí se 1000 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG jakož i celkový počet selhání a celkový čas výpočtu

Metoda	NIT	NFV	NFG	selhání	čas
TRGN-5	8687	9074	8746	1	25.44
TRGN-6	9780	10200	9838	–	13.75
TRGNS-5	6358	6730	6418	–	19.43
TRGNS-6	7226	7456	7286	–	9.45
TRGNP-5	7089	7750	7149	–	21.34
TRGNP-6	8972	9893	9032	–	16.28
TRGNN-5	5515	5819	7281	–	17.39
TRGNN-6	7403	7785	8234	–	11.27
TRMNS-5	7929	53435	53267	1	63.72
TRMNS-6	11329	99518	98491	2	66.76
LSVMP-1	11217	30393	30393	1	19.27

Selhání znamená, že nestačilo 1000 iterací nebo 2000 vyčíslení součtu čtverců pro vyřešení úlohy. Z této tabulky je patrné, že pro řídké nejmenší čtverce jsou modifikace GN metody mnohem efektivnější než obecné metody pro řídké nebo separovatelné úlohy.

10.4 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. Pracujeme-li s maticí J, jsou možnosti použití informací druhého řádu značně omezené a zde se jimi zabývat nebudeme. Zaměříme se pouze na úpravy metody sdružených gradientů pro řešení normální soustavy rovnic $J^T J s + J^T f = 0$.

Nejjednodušší úpravou metody CG pro řešení normální soustavy rovnic je metoda CGNE.

Definice 41 *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

$$v_i = J p_i \quad \alpha_i = \|g_i\|^2 / \|v_i\|^2,$$

$$s_{i+1} = s_i + \alpha_i p_i, \quad u_{i+1} = u_i + \alpha_i v_i,$$

$$g_{i+1} = J^T u_{i+1}, \quad \beta_i = \|g_{i+1}\|^2 / \|g_i\|^2,$$

$$p_{i+1} = -g_{i+1} + \beta_i p_i$$

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 3.6. 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 konvenčních důvodů budeme v následující definici používat koeficienty α_i, β_i, $1 \le i \le n$, které nemají nic společného se stejně označenými koeficienty vystupujícími v definici 41.

Definice 42 *Nechť $J \in R^{m\times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$\beta_1 u_1 = f, \qquad \alpha_1 q_1 = J^T u_1$$

a

$$\beta_{i+1} u_{i+1} = J q_i - \alpha_i u_i,$$
$$\alpha_{i+1} q_{i+1} = J^T u_{i+1} - \beta_{i+1} q_i$$

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 bidiagonalizačním Lanczosovým procesem (BL) určeným maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$.

Poznámka 258 Nechť $\alpha_i \neq 0$, $\beta_i \neq 0$, $1 \le i \le k$ pro nějaký index $1 \le k \le n$. Pak podle definice 42 platí $f = U_{k+1}(\beta_1 e_1)$ a

$$JQ_k = U_{k+1} B_k, \tag{426}$$

$$J^T U_{k+1} = Q_k B_k^T + \alpha_{k+1} q_{k+1} e_{k+1}^T, \tag{427}$$

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

$$B_k = \begin{bmatrix} \alpha_1, & 0, & \ldots, & 0 \\ \beta_2, & \alpha_2, & \ldots, & 0 \\ --, & --, & --, & -- \\ 0, & 0, & \ldots, & \alpha_k \\ 0, & 0, & \ldots, & \beta_{k+1} \end{bmatrix}$$

(matice $B_k \in R^{(k+1)\times k}$ je bidiagonální).

Věta 152 *Uvažujme bidiagonalizační Lanczosův proces určený maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$. Nechť $\alpha_i \neq 0$, $\beta_i \neq 0$, $1 \le i \le k$, pro nějaký index $1 \le k \le n$. Pak vektory q_i, $1 \le i \le k$, 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, $1 \le i \le k$, 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 $1 \le k < n$ a že $\alpha_{k+1} \neq 0$, $\beta_{k+1} \neq 0$. Použijeme-li vztahy (426)–(427), dostaneme

$$J^T J Q_k = J^T U_{k+1} B_k = Q_k B_k^T B_k + \alpha_{k+1} q_{k+1} e_{k+1}^T B_k = Q_k T_k + \alpha_{k+1}\beta_{k+1} q_{k+1} e_k^T,$$

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 (426)–(427), kde $B = J^T J$, $T_k = B_k^T B_k$ a $\gamma_i = \alpha_i^2 + \beta_{i+1}^2$, $\delta_i = \alpha_i\beta_i$, $1 \le i \le k$, a můžeme použít větu 95, podle které tvoří vektory q_i, $1 \le i \le k+1$ 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 (426), dostaneme

$$U_{k+1}^T J Q_k = U_{k+1}^T U_{k+1} B_k$$

a podle (427) platí

$$U_{k+1}^T J Q_k = B_k Q_k^T Q_k + \alpha_{k+1} e_{k+1} q_{k+1}^T Q_k = B_k,$$

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 259 Z důkazu věty 152 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 260 Bidiagonalizační Lanczosův proces můžeme použít k řešení soustavy rovnic $J^T J s + J^T g = 0$. 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, jestliže $s = Q_i z$, 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 $J^T J s + J^T f = 0$ (plyne to z poznámky 187 a poznámky 259).

Poznámka 261 Vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 260 jsou shodné s 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 96 a poznámky 259).

Výhodou bidiagonalizačního Lanczosova procesu je skutečnost, že vektory s_{i+1}, $1 \leq i \leq k$, mohou být určeny pomocí stabilních operací (Givensových elementárních rotací popsaných v oddílu 7.3). To tvoří základ metody LSQR. Princip metody LSQR spočívá v tom, že se rekurentně určují rozklady

$$P_i B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \quad P_i(\beta_1 e_1) = \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix},$$

kde

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \ldots, & 0 \\ 0, & \rho_2, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \rho_i \end{bmatrix}, \quad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \ldots \\ \eta_i \end{bmatrix}.$$

Přitom $P_i \in R^{i \times i}$, $1 \leq i \leq k$, jsou ortogonální matice (součiny Givensových elementárních rotací) a $R_i \in R^{i \times i}$, $1 \leq i \leq k$, jsou horní bidiagonální matice. Použité Givensovy elementární rotace mají následující vlastnosti

Lemma 49 *Nechť*

$$\overline{P}_i = \frac{1}{\sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2}} \begin{bmatrix} \overline{\rho}_i & \beta_{i+1} \\ -\beta_{i+1} & \overline{\rho}_i \end{bmatrix}.$$

Pak matice $\overline{P}_i$ je ortogonální a platí

$$\overline{P}_i \begin{bmatrix} \overline{\rho}_i \\ \beta_{i+1} \end{bmatrix} = \begin{bmatrix} \sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2} \\ 0 \end{bmatrix},$$

$$\overline{P}_i \begin{bmatrix} \overline{\eta}_i \\ 0 \end{bmatrix} = \frac{1}{\sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2}} \begin{bmatrix} \overline{\rho}_i \overline{\eta}_i \\ -\beta_{i+1} \overline{\eta}_i \end{bmatrix} \triangleq \begin{bmatrix} \eta_i \\ \overline{\eta}_{i+1} \end{bmatrix},$$

$$\overline{P}_i \begin{bmatrix} 0 \\ \alpha_{i+1} \end{bmatrix} = \frac{1}{\sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2}} \begin{bmatrix} \beta_{i+1} \alpha_{i+1} \\ \overline{\rho}_i \alpha_{i+1} \end{bmatrix} \triangleq \begin{bmatrix} \sigma_{i+1} \\ \overline{\rho}_{i+1} \end{bmatrix}.$$

Důkaz Zřejmě

$$\overline{P}_i^T \overline{P}_i = \frac{1}{\overline{\rho}_i^2 + \beta_{i+1}^2} \begin{bmatrix} \overline{\rho}_i & -\beta_{i+1} \\ \beta_{i+1} & \overline{\rho}_i \end{bmatrix} \begin{bmatrix} \overline{\rho}_i & \beta_{i+1} \\ -\beta_{i+1} & \overline{\rho}_i \end{bmatrix} = I.$$

Zbylé vztahy dostaneme snadno dosazením a roznásobením. □

Ortogonální matice P_i, $1 \leq i \leq k$, budeme hledat ve tvaru $P_1 = \overline{P}_1$ a

$$P_i = \begin{bmatrix} I & 0 \\ 0 & \overline{P}_i \end{bmatrix} \begin{bmatrix} P_{i-1} & 0 \\ 0 & 1 \end{bmatrix},$$

pro $1 < i \leq k$, kde I je jednotková matice řádu $i-2$. Pak

$$P_i B_i = \begin{bmatrix} I & 0 \\ 0 & \overline{P}_i \end{bmatrix} \begin{bmatrix} \rho_1 & \sigma_2 & \dots & 0 & 0 \\ 0 & \rho_2 & \dots & 0 & 0 \\ -- & -- & -- & -- & -- \\ 0 & 0 & \dots & \rho_{i-1} & \sigma_i \\ 0 & 0 & \dots & 0 & \overline{\rho}_i \\ 0 & 0 & \dots & 0 & \beta_{i+1} \end{bmatrix} = \begin{bmatrix} \rho_1 & \sigma_2 & \dots & 0 & 0 \\ 0 & \rho_2 & \dots & 0 & 0 \\ -- & -- & -- & -- & -- \\ 0 & 0 & \dots & \rho_{i-1} & \sigma_i \\ 0 & 0 & \dots & 0 & \rho_i \\ 0 & 0 & \dots & 0 & 0 \end{bmatrix}$$

pro $1 < i \leq k$. Z těchto vztahů a z lemmatu 49 dostaneme rekurentní vztahy

$$\overline{\rho}_1 = \alpha_1, \quad \overline{\eta}_1 = \beta_1$$

a

$$\begin{aligned} \rho_i = \sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2} \quad &, \quad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \quad \tau_i = \frac{\beta_{i+1}}{\rho_i}, \\ \overline{\rho}_{i+1} = \lambda_i \alpha_{i+1} \quad &, \quad \sigma_{i+1} = \tau_i \alpha_{i+1}, \\ \eta_i = \lambda_i \overline{\eta}_i \quad &, \quad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i \end{aligned}$$

pro $1 \leq i \leq k$. Nyní odvodíme rekurentní vztahy pro vektory s_{i+1}, $1 \leq i \leq k$. Jelikož

$$P_i(B_i z + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

a $P_i^T P_i = I$, můžeme položit $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i} \|R_i z + h_i\|.$$

Jelikož matice $R_i \in R^{i \times i}$ je regulární, musí platit $R_i z_i + h_i = 0$. Vzhledem k jednoduché struktuře matic R_i, $1 \leq i \leq k$, můžeme vektory z_i, $1 \leq i \leq k$, a tudíž i vektory s_{i+1}, $1 \leq i \leq k$, určovat rekurentně.

Lemma 50 *Vektory $s_{i+1} = Q_i z_i$, $1 \leq i \leq k$, kde $R_i z_i + h_i = 0$, lze určit pomocí rekurentních vztahů*

$$s_1 = 0, \quad p_1 = q_1$$

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 $1 \leq i \leq k$.

Důkaz Platí $R_1 = [\rho_1]$, $R_1^{-1} = [1/\rho_1]$ a

$$R_i = \begin{bmatrix} R_{i-1} & \sigma_i e_{i-1} \\ 0 & \rho_i \end{bmatrix}, \quad R_i^{-1} = \begin{bmatrix} R_{i-1}^{-1} & -(\sigma_i/\rho_i)R_{i-1}^{-1}e_{i-1} \\ 0 & 1/\rho_i \end{bmatrix}$$

pro $1 < i \leq k$, kde e_{i-1} je poslední sloupec jednotkové matice řádu $i-1$ (vztah pro R_i^{-1} můžeme ověřit dosazením do rovnosti $R_i R_i^{-1} = I$). Označíme-li w_i poslední sloupec matice R_i^{-1} (takže $w_i = R_i^{-1}e_i$) a položíme-li $z_i = -R_i^{-1}h_i$, dostaneme z předchozích rovností rekurentní vztahy $w_1 = [1/\rho_1]$, $z_1 = -[\eta_1/\rho_1]$ a

$$w_i = \begin{bmatrix} -(\sigma_i/\rho_i)w_{i-1} \\ 1/\rho_i \end{bmatrix}, \quad z_i = \begin{bmatrix} z_{i-1} + (\sigma_i/\rho_i)\eta_i w_{i-1} \\ -\eta_i/\rho_i \end{bmatrix}.$$

pro $1 < i \leq k$, takže

$$\begin{aligned} p_i &\stackrel{\Delta}{=} \rho_i Q_i w_i = q_i - \frac{\sigma_i}{\rho_{i-1}}\rho_{i-1}Q_{i-1}w_{i-1} = q_i - \frac{\sigma_i}{\rho_{i-1}}p_{i-1}, \\ s_{i+1} &= Q_i z_i = Q_{i-1}z_{i-1} + \frac{\eta_i}{\rho_i}\sigma_i Q_{i-1}w_{i-1} - \frac{\eta_i}{\rho_i}q_i \\ &= s_i - \frac{\eta_i}{\rho_i}(q_i - \frac{\sigma_i}{\rho_{i-1}}\rho_{i-1}Q_{i-1}w_{i-1}) = s_i - \frac{\eta_i}{\rho_i}p_i. \end{aligned}$$

□

Definice 43 *Nechť $J \in R^{m\times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \quad \beta_1 u_1 = f, \quad \alpha_1 q_1 = J^T u_1, \quad \overline{\eta}_1 = \beta_1, \quad \overline{\rho}_1 = \alpha_1, \quad p_1 = q_1$$

a

$$\beta_{i+1}u_{i+1} = Jq_i - \alpha_i u_i,$$

$$\alpha_{i+1}q_{i+1} = J^T u_{i+1} - \beta_{i+1}q_i,$$

$$\rho_i = \sqrt{\overline{\rho}_i^2 + \beta_{i+1}^2}, \quad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \quad \tau_i = \frac{\beta_{i+1}}{\rho_i},$$

$$\overline{\rho}_{i+1} = \lambda_i \alpha_{i+1}, \quad \sigma_{i+1} = \tau_i \alpha_{i+1},$$

$$\eta_i = \lambda_i \overline{\eta}_i, \quad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i,$$

$$s_{i+1} = s_i - \frac{\eta_i}{\rho_i}p_i,$$

$$p_{i+1} = q_{i+1} - \frac{\sigma_{i+1}}{\rho_i}p_i$$

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 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 261 generují obě metody stejné vektory s_{i+1}, $1 \leq i \leq k$, kde $k \leq n$ a $J^T J s_{k+1} + J^T f = 0$. Ukážeme ještě, jak je možné odhadovat přesnost řešení.

Věta 153 *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 (426)–(427) a poznámky 260 dostaneme

$$\begin{aligned} J^T(Js_{i+1} + f) &= J^T(JQ_iz_i + f) = J^TU_{i+1}(B_iz_i + \beta_1e_1) = \\ &= (Q_iB_i^T + \alpha_{i+1}q_{i+1}e_{i+1}^T)(B_iz_i + \beta_1e_1) = \alpha_{i+1}q_{i+1}e_{i+1}^TB_iz_i = \\ &= \alpha_{i+1}\beta_{i+1}q_{i+1}e_i^Tz_i, \end{aligned}$$

neboť $B_i^T(B_iz_i + \beta_1e_1) = 0$ podle definice vektoru z_i, $e_{i+1}^Te_1 = 0$ a $e_{i+1}^TB_i = \beta_{i+1}e_i^T$. Ale $Q_i^TQ_i = I$ a tudíž $Q_i^Ts_{i+1} = Q_i^TQ_iz_i = z_i$, takže $e_i^Tz_i = e_i^TQ_i^Ts_{i+1} = q_i^Ts_{i+1}$, což spolu s $\|q_{i+1}\| = 1$ dává

$$\|J^T(Js_{i+1} + f)\| = \alpha_{i+1}\beta_{i+1}|q_i^Ts_{i+1}|.$$

Ale

$$q_i^Ts_{i+1} = q_i^Ts_i + \frac{\eta_i}{\rho_i}q_i^Tp_i = q_i^TQ_{i-1}z_{i-1} - \frac{\eta_i}{\rho_i}q_i^Tq_i + \frac{\eta_i\sigma_i}{\rho_i\rho_{i-1}}q_i^Tp_{i-1} = -\frac{\eta_i}{\rho_i},$$

neboť $q_i^TQ_{i-1} = 0$, $q_i^Tq_i = 1$ a vektor p_{i-1} je lineární kombinací sloupců matice Q_{i-1}, tudíž $q_i^Tp_{i-1} = 0$. Jestliže $\alpha_{i+1} = 0$, $\beta_{i+1} = 0$, platí $\|J^T(Js_{i+1} + f)\| = 0$ (poznámka 260). □

Větu 153 můžeme využít k zastavení iteračního procesu (není třeba počítat reziduum $\|J^T(Js_{i+1}+f)\|$). Následující tabulka ukazuje srovnání nepřesné GN metody s lokálně omezeným krokem realizované pomocí řídké reprezentace Hessovy matice a pomocí metody CG se dvěma nepřesnými GN metodami s lokálně omezeným krokem realizovanými pomocí řídké reprezentace Jacobiho matice a pomocí metod CGNE nebo LSQR. Je opět použito 22 testovacích funkcí s 1000 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG jakož i celkový počet selhání a celkový čas výpočtu

Metoda	NIT	NFV	NFG	NCG	čas
`TRGNA-3`	13932	14464	13977	1232187	119.24
`TRGNA-4`	7470	8008	7523	1226757	156.78
`TRGNR-3`	9931	10488	9985	1224959	115.61
`TRGNR-4`	7089	7629	7143	1225386	156.01
`TRGNV-3`	13640	14163	13692	1120004	109.41
`TRGN-7`	7255	7782	7308	1180653	125.95
`TRGNN-7`	6774	7258	7313	1087690	120.30

Z této tabulky je patrné, že pokud nejsou řádky Jacobiovy matice příliš zaplněny, je výhodnější pracovat s řídkou reprezentací Hessovy matice (GN+CG), která pracuje s méně zaplněnou maticí B. V opačném případě se rozhodujeme podle složitosti optimalizačního kriteria. Metoda CGNE používá jednodušší maticové operace a metoda LSQR potřebuje méně iterací a méně vyčíslení optimalizačního kriteria.

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 5). 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 $\mathcal{D}_F(\overline{F}) \subset \mathcal{D} \subset \mathcal{D}_F$. 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 $F(x) = (1/2)\|f(x)\|^2$ (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 154 *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

$$g(x^*) = J^T(x^*)f(x^*).$$

Jelikož matice $J(x^*)$ je regulární, můžeme psát

$$f(x^*) = (J^T(x^*))^{-1}g(x^*),$$

takže $f(x^*) = 0$ právě tehdy, jestliže $g(x^*) = 0$, 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 44 *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, $1 \leq j \leq i$, 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 45 *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) .

Při vyšetřování konvergence metod pro řešení soustav nelineárních rovnic budeme používat předpoklady (J1)–(J6) uvedené v oddílu 7. Na rozdíl od optimalizačních metod popsaných v kapitolách 2 a 5, kde se v důkazech globální konvergence nepoužívá podmínka (F5), budeme nyní (v případě Newtonovy metody) potřebovat nějakou analogii podmínky (J5). Podmínka (J5) je velmi silná, jak ukazuje tato věta.

Věta 155 *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 $f(x^*) = 0$. Platí-li (J5), pak podle (291) pro $x \in \mathcal{D}$, $x \neq x^*$, dostaneme $\|f(x)\| = \|f(x) - f(x^*)\| \geq \underline{J}\|x - x^*\| > 0$. □

Poznámka 262 Protože je obtížné zajisitit platnost podmínek (J4)–(J5) na příliš velké otevřené množině $\mathcal{D}$, budeme v důkazech globální konvergence Newtonovy metody popužívat slabší podmínky

$$\|J(x_i)s\| \leq \overline{J}\|s\| \qquad \forall i \in N \qquad \forall s \in R^n, \tag{J4a}$$

$$\|J(x_i)s\| \geq \underline{J}\|s\| \qquad \forall i \in N \qquad \forall s \in R^n. \tag{J5a}$$

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 podmínky

$$\|A_i - J_i\| \leq \overline{\vartheta}, \tag{A3a}$$

$$\|A_i s\| \leq \overline{A}s \qquad \forall s \in R^n, \tag{A4a}$$

$$\|A_i s\| \geq \underline{A}s \qquad \forall s \in R^n. \tag{A5a}$$

Podmínka (A4a) je ekvivalentní podmínce $\|A_i\| \leq \overline{A}$. Podobně podmínka (A5a) je ekvivalentní podmínce $\|A_i^{-1}\|^{-1} \geq \underline{A}$.

Poznámka 263 Poznamenejme, že z (J4a) a (A3a) plyne

$$\|A_i s\| \leq \|J_i s\| + \|(A_i - J_i)s\| \leq (\overline{J} + \overline{\vartheta})\|s\|,$$

takže platí (A4a) s $\overline{A} = \overline{J} + \overline{\vartheta}$. Pokud $\overline{\vartheta} < \underline{J}$, pak z (J5a) a (A3a) plyne

$$\|A_i s\| \geq \|J_i s\| - \|(A_i - J_i)s\| \geq (\underline{J} - \overline{\vartheta})\|s\|,$$

takže platí (A5a) s $\underline{A} = \underline{J} - \overline{\vartheta}$. Platí-li (J4a), (J5a) a (A3a), budeme předpokládat, že $\overline{\vartheta} < \underline{J}$ a že čísla $\overline{A}$, $\underline{A}$ použitá v (A4a), (A5a) 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 platilo (A5a), ale v důkazech globální konvergence metod pro řešení soustav nelineárních rovnic budeme potřebovat silnější podmínku.

$$\overline{\vartheta} < \frac{1-\overline{\omega}}{2}\underline{J}, \tag{428}$$

kde číslo $\overline{\omega}$ udává přesnost výpočtu směrového vektoru (definice 46 a definice 47).

Podmínky (A3a)–(A5a) mohou být nahraženy slabšími podmínkami

$$\|(A_i - J_i)^T f_i\| \leq \overline{\vartheta}\|f_i\|, \tag{A3b}$$

$$\|A_i s_i\| \leq \overline{A} s_i, \tag{A4b}$$

$$\|A_i s_i\| \geq \underline{A} s_i, \tag{A5b}$$

kde A_i je regulární matice, f_i je hodnota zobrazení f v bodě x_i a s_i je použitý směrový vektor. Poznamenejme, že z (J4a), (J5a) a (A3b) neplyne žádná z podmínek (A4b), (A5b) (tyto podmínky musí být splněny nezávisle na (A3b)).

Poznámka 264 Je zřemé, že z (A3a)–(A5a) plyne (A3b)–(A5b). Navíc podmínka (428) je ekvivalentní nerovnosti

$$\overline{\vartheta} < \frac{1-\overline{\omega}}{1+\overline{\omega}}\underline{A} \tag{429}$$

(pokud $\underline{A} = \underline{J} - \overline{\vartheta}$). Význam podmínek (A3b) s (429) a (A4b)–(A5b) spočívá v tom, že některé metody splňují podmínku (A3b) s (429) automaticky (sdružená kvazinewtonovská metoda uvedená v poznámce 283 splňuje podmínku (A3b) s $\overline{\vartheta} = 0$) a podmínky (A4b)–(A5b) lze snadno zajistit algoritmicky. Navíc podmínky (A3b) s (429) 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 46, implikuje nerovnost

$$h_i^T s_i = f_i^T A_i s_i = f_i^T (A_i s_i + f_i) - f_i^T f_i \leq \overline{\omega}\|f_i\|^2 - \|f_i\|^2 = -(1-\overline{\omega})\|f_i\|^2 < 0. \tag{430}$$

Definice 46 *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

$$F_{i+1} - F_i \leq -2\underline{\rho}(1-\overline{\omega})\alpha_i F_i, \tag{$\overline{\text{S2b}}$}$$

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 $0 < \underline{\rho} < 1$.

V dalším výkladu budeme předpokládat, že parametr $\underline{\rho}$ v ($\overline{\text{S2}}$) splňuje nerovnost $0 < \underline{\rho} < 1 - \lambda$, kde

$$\lambda = \frac{\overline{\vartheta}}{\underline{A}} \frac{1+\overline{\omega}}{1-\overline{\omega}} \tag{431}$$

(takže podle (429) platí $0 \leq \lambda < 1$).

Lemma 51 *(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 podmínky (A3b) s (429) a (A5b). Pak lze v každém iteračním kroku metody spádových směrů (definice 46) nalézt směrový vektor $s_i \in R^n$ vyhovující podmínce ($\overline{S1}$) a délku kroku $\alpha_i > 0$ vyhovující libovolné z podmínek ($\overline{S2}$) (s $0 < \underline{\rho} < 1 - \lambda$). Pokud $\overline{\vartheta} = 0$, platí $\lambda = 0$ a nepotřebujeme, aby byla splněna podmínka (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

$$(1-\overline{\omega})\|f_i\| \leq \|A_i s_i\| \leq (1+\overline{\omega})\|f_i\|, \tag{432}$$

$$(1-\overline{\omega})\|f_i\|^2 \leq -h_i^T s_i \leq (1+\overline{\omega})\|f_i\|^2, \tag{433}$$

$$|h_i^T s_i - g_i^T s_i| \leq \overline{\vartheta}\|f_i\|\|s_i\|. \tag{434}$$

Nerovnost (432) spolu s podmínkou (A5b) dává

$$\|s_i\| \leq \frac{1+\overline{\omega}}{\underline{A}}\|f_i\|, \tag{435}$$

Použijeme-li vztahy (430), (431) a nerovnosti (434), (435), dostaneme

$$\begin{aligned} -g_i^T s_i &\geq -h_i^T s_i - \overline{\vartheta}\|f_i\|\|s_i\| \geq -h_i^T s_i - \overline{\vartheta}\frac{1+\overline{\omega}}{\underline{A}}\|f_i\|^2 \\ &\geq -h_i^T s_i + \frac{\overline{\vartheta}}{\underline{A}}\frac{1+\overline{\omega}}{1-\overline{\omega}} h_i^T s_i = -(1-\lambda)h_i^T s_i \geq (1-\lambda)(1-\overline{\omega})\|f_i\|^2 > 0, \end{aligned} \tag{436}$$

takže podle lemmatu 3 existuje pro libovolné číslo $0 < \varepsilon_1 < 1$ délka kroku $\alpha_i > 0$ určená Armijovým výběrem (poznámka 20) 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 3 jsou splněny, neboť podle věty 102 z (J1), (J4), (J6) plyne (F3)). Položme $\underline{\rho} = \varepsilon_1(1-\lambda)$, takže $0 < \underline{\rho} < 1 - \lambda \leq 1$. Pak

$$F_{i+1} - F_i \leq \underline{\rho}\alpha_i h_i^T s_i \leq -2\underline{\rho}(1-\overline{\omega})\alpha_i F_i,$$

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

$$\|f_{i+1}\| - \|f_i\| \leq \frac{F_{i+1} - F_i}{\|f_i\|} \leq -2\underline{\rho}(1-\overline{\omega})\alpha_i \frac{F_i}{\|f_i\|} = -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|.$$

□

Poznámka 265 Poznamenejme, že kromě konzistence podmínek ($\overline{\text{S2a}}$)–($\overline{\text{S2c}}$) jsme dokázali implikace ($\overline{\text{S2a}}$) $\Rightarrow$ ($\overline{\text{S2b}}$) $\Rightarrow$ ($\overline{\text{S2c}}$) (pro stejnou hodnotu parametru $\underline{\rho}$). Ukážeme, že platí také opačné implikace ($\overline{\text{S2c}}$) $\Rightarrow$ ($\overline{\text{S2b}}$) $\Rightarrow$ ($\overline{\text{S2a}}$), 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 (433), 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 52 *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům (J1), (J4), (J6). Nechť matice A_i, $i \in N$, jsou regulární a splňují podmínky (A3b) s (429) a (A5b). Pak existuje konstanta $0 < \underline{\alpha} \leq 1$ taková, že délky kroku určené metodou spádových směrů (definice 46) splňují podmínku $0 < \underline{\alpha} \leq \alpha_i \leq 1$, $i \in N$.*

Důkaz Podle poznámky 265 se můžeme omezit na metody spádových směrů používající podmínku ($\overline{\text{S2a}}$). Při výběru délky kroku podle ($\overline{\text{S2a}}$) 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 $F(x_i + \alpha_i^{k-1}s_i) - F(x_i) \geq \underline{\rho}\alpha_i^{k-1}h_i^T s_i$. Pokud $\alpha_i < 1$, můžeme psát

$$F(x_i + \frac{\alpha_i}{\beta}s_i) - F(x_i) \geq \underline{\rho}\frac{\alpha_i}{\beta}h_i^T s_i.$$

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 (293) 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

$$\underline{\rho}h_i^T s_i \leq g_i^T s_i + \frac{\alpha_i}{\beta}(\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2$$

a použijeme-li (435) a (436), můžeme psát

$$\frac{\alpha_i}{\beta}\left(\overline{J}^2 + \overline{G}\,\overline{f}\right)\frac{(1+\overline{\omega})^2}{\underline{A}^2}\|f_i\|^2 \geq \underline{\rho}h_i^T s_i - g_i^T s_i \geq (\underline{\rho}-(1-\lambda))h_i^T s_i \geq (1-\overline{\omega})(1-\underline{\rho}-\lambda)\|f_i\|^2$$

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

$$\underline{\alpha} = \min\left(1, \frac{\underline{\beta}(1-\overline{\omega})(1-\underline{\rho}-\lambda)\underline{A}^2}{(\overline{J}^2+\overline{G}\,\overline{f})(1+\overline{\omega})^2}\right), \tag{437}$$

platí $0 < \underline{\alpha} \leq \alpha_i \leq 1 \ \forall i \in N$. □

Poznámka 266 Podmínku (A5b) potřebujeme pouze k tomu, abychom mohli použít nerovnost (435). Proto můžeme podmínku (A5b) 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 53 *Uvažujme základní metodu pro řešení soustav nelineárních rovnic (definice 44) takovou, že $\|s_i\| \leq \overline{c}\|f_i\|$, $i \in N$, a délky kroku splňují některou z podmínek ($\overline{S2}$) 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 265). Podle ($\overline{\text{S2c}}$) platí

$$\|f_{i+1}\| \leq (1-\underline{\rho}(1-\overline{\omega})\alpha_i)\|f_i\| \leq (1-\underline{\rho}(1-\overline{\omega})\underline{\alpha})\|f_i\| \triangleq q\|f_i\|,$$

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

$$\sum_{i=1}^{\infty}\|f_i\| \leq \frac{1}{1-q}\|f_1\| < \infty,$$

což implikuje $\|f_i\| \to 0$. Použijeme-li nerovnosti $\|s_i\| \leq \overline{c}\|f_i\|$, $i \in N$, můžeme psát

$$\sum_{i=1}^{\infty}\|s_i\| \leq \overline{c}\sum_{i=1}^{\infty}\|f_i\| < \infty,$$

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 156 *(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 podmínky (A3b) s (429) a (A5b). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů (definice 46). Pak $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 52 a lemmatu 53. Předpoklady lemmatu 53 jsou splněny, neboť podle (435) pro $i \in N$ platí $\|s_i\| \leq \overline{c}\|f_i\|$, kde $\overline{c} = (1+\overline{\omega})/\underline{A}$. □

Poznámka 267 Z odhadu $F_{i+1} \leq qF_i$, $i \in N$, kde $0 < q < 1$, plyne, že $x_i \to x^\star$ R-lineárně.

Nyní se budeme zabývat superlineární konvergencí metod spádových směrů. Budeme přitom používat podmínku ($\overline{\text{S2c}}$) s $0 < \underline{\rho} < 1$.

Věta 157 *(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*

$$\lim_{i\to\infty}\frac{\|A_i s_i + f_i\|}{\|f_i\|} = 0 \tag{438}$$

a

$$\lim_{i\to\infty}\frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0. \tag{439}$$

Pak existuje index $k \in N$ takový, že $\alpha_i = 1$, pokud $i \geq k$ 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 182 a 185. To znamená, že v částech (a)–(b) budeme místo (438) 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 (438), 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 (439)) 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 (439)) 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

$$\frac{\|f(x_i + s_i)\|}{\|f_i\|} \le \|\omega_i\| + \|\vartheta_i\|(1+\overline{\omega})/\underline{J} + o(\|f_i\|)/\|f_i\|, \tag{440}$$

takže $\limsup_{i\to\infty}(\|f(x_i + s_i)\| - \|f_i\|)/\|f_i\| \le -(1-\overline{\omega})$ (podle ($\overline{\text{S1}}$) a (439)), 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 (438)–(439). Pomocí věty 5 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 (438)–(439) a (440) 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 268 Věta 157 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 265. Například z ($\overline{\text{S2b}}$) plyne $2\underline{\rho}\alpha_i \leq 1 - F_{i+1}/F_i < 1$ (předpokládáme, že $F_i > 0\ \forall i \in N$), takže $\alpha_i = 1$ lze volit pouze tehdy, když $2\underline{\rho} < 1$.

Poznámka 269 Položíme-li $A_i = J(x_i)$, $i \in N$, dostaneme Newtonovu metodu. V tomto případě podmínky (J4a)–(J5a) implikují (A4a)–(A5a) a (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 (platí-li (J1), (J4), (J5a) a (J6)).

Následující lemma ukazuje, jak lze vlastnosti libovolné metody spádových směrů odvodit z vlastností Newtonovy metody.

Lemma 54 *Nechť matice $J(x_i)$, $i \in N$, splňují podmínku (J5a) a matice A_i, $i \in N$, splňují podmínku (A3a) s (428). Nechť s_i je směrový vektor vyhovující podmínce ($\overline{\text{S1}}$). Pak platí*

$$\|J_i s_i + f_i\| \leq \tilde{\omega}\|f_i\|,$$

kde $\tilde{\omega} = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$. Jinými slovy, platí-li ($\overline{\text{S1}}$) a (A3a) s (428), 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 (432) 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

$$\|J_i s_i + f_i\| \leq \|A_i s_i + f_i\| + \|(J_i - A_i)s_i\| \leq \overline{\omega}\|f_i\| + \overline{\vartheta}\|s_i\| \leq \frac{\underline{J}\overline{\omega} + \overline{\vartheta}}{\underline{J} - \overline{\vartheta}}\|f_i\| \triangleq \tilde{\omega}\|f_i\|.$$

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 lemmatech 51 a 54 vyžadují splnění podmínky (A3a) (s vhodnou hodnotou $\overline{\vartheta} > 0$, která může vycházet velmi malá). Tato podmínka má velký teoretický význam, ale v praxi ji 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 v podstatě o to aby byla splněna některá z podmínek ($\overline{\text{S2a}}$)–($\overline{\text{S2c}}$)). V případě, že neplatí ($\overline{\text{S2}}$) pro α_i větší než zadaná dolní mez, provede se restart, což znamená, že se spočte matice J a použije se krok Newtonovy metody. Tyto úvahy jsou shrnuty ve formě algoritmu.

Algoritmus 19 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 kvazinewtonovské aktualizace), položíme $i := i+1$, $l := l + 1$ a přejdeme na krok 2.

Věta 158 *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 19, 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 (437), ve kterém $\underline{A} = \underline{J}$. Pak nutně libovolná vybraná podmínka z ($\overline{\text{S2}}$) je splněna pro $k \leq \overline{k}$, takže algoritmus 19 nemůže skončit v kroku 5c. Může skončit v kroku 2, pokud $f(x_i) = 0$. V opačném případě podle lemmatu 52 a lemmatu 53 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 47 *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

$$\|s_i\| \leq \Delta_i, \tag{$\overline{\text{T1a}}$}$$

$$\|s_i\| < \Delta_i \Rightarrow \|A_i s_i + f_i\| \leq \overline{\omega}_i \|f_i\|, \tag{$\overline{\text{T1b}}$}$$

$$-L_i(s_i) \geq \underline{\sigma}\|A_i s_i\|, \tag{$\overline{\text{T1c}}$}$$

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

$$\rho_i(s_i) \leq 0 \Rightarrow \alpha_i = 0, \tag{$\overline{\text{T2a}}$}$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1. \tag{$\overline{\text{T2b}}$}$$

(3) Meze $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se určují tak, že

$$\rho_i(s_i) < \overline{\rho} \Rightarrow \underline{\beta}\|s_i\| \le \Delta_{i+1} \le \overline{\beta}\|s_i\|, \qquad (\overline{\text{T3a}})$$

$$\rho_i(s_i) \ge \overline{\rho} \Rightarrow \Delta_i \le \Delta_{i+1} \le \overline{\Delta}, \qquad (\overline{\text{T3b}})$$

kde $0 < \underline{\beta} < \overline{\beta} < 1$ *a* $0 < \overline{\rho} < 1/2$.

Poznámka 270 Při vyšetřování metod s lokálně omezeným krokem budeme používat označení

$$N_1 = \{i \in N : \|s_i\| < \Delta_i\},$$

$$N_2 = \{i \in N : \rho_i(s_i) > 0\},$$

$$N_3 = \{i \in N : \rho_i(s_i) \ge \overline{\rho}\}.$$

Jelikož $\overline{\rho} > 0$, platí $N_3 \subset N_2$.

Lemma 55 *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 podmínky (A3b)–(A5b) s (429). Nechť* $x_i \in R^n$, $i \in N$, *je posloupnost generovaná metodou s lokálně omezeným krokem* $(\overline{T1})$–$(\overline{T3})$ *(s* $0 < 2\overline{\rho} < 1 - \lambda$*). Pak existuje konstanta* $\underline{c} > 0$ *taková, že*

$$\|s_i\| \ge \underline{c}\|f_i\| \quad \forall i \in N.$$

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 $i \notin N_3$. Z $(\overline{\text{T1c}})$ plyne, že $L_i(s_i) \le 0$, takže

$$\begin{aligned} L_i(s_i)\|f_i\| &= \left(\|A_i s_i + f_i\| - \|f_i\|\right)\|f_i\| \ge \left(\|A_i s_i + f_i\|^2 - \|f_i\|^2\right) \\ &= 2\left(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i\right) \triangleq 2Q_i(s_i). \end{aligned} \qquad (441)$$

Jestliže $\|f(x_i + s_i)\| \le \|f(x_i)\|$, pak nerovnost $\rho_i(s_i) < \overline{\rho}$ spolu s (441) dává

$$\begin{aligned} F(x_i + s_i) - F(x_i) &= \frac{1}{2}\left(\|f(x_i + s_i)\|^2 - \|f(x_i)\|^2\right) \\ &\ge \left(\|f(x_i + s_i)\| - \|f(x_i)\|\right)\|f(x_i)\| \\ &\ge \overline{\rho} L_i(s_i)\|f_i\| \ge 2\overline{\rho} Q_i(s_i). \end{aligned}$$

Jestliže $\|f(x_i + s_i)\| \ge \|f(x_i)\|$, platí tato nerovnost triviálně. Můžeme tedy psát

$$F(x_i + s_i) - F(x_i) \ge 2\overline{\rho} Q_i(s_i).$$

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 (436), 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 $h_i^T s_i = f_i^T A_i s_i \leq Q_i(s_i)$ a podle (293) platí

$$\|g(x_i + d_i) - g(x_i)\| \leq (\overline{J}^2 + \overline{G}\,\overline{f})\|d_i\| \leq (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$2\overline{\rho}Q_i(s_i) \leq (1-\lambda)Q_i(s_i) + (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2,$$

neboli

$$-(1-\lambda-2\overline{\rho})Q_i(s_i) \leq (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2.$$

Podmínky ($\overline{\text{T1c}}$) a (A5a) spolu s nerovností (441) dávají

$$-Q_i(s_i) \geq -\frac{1}{2}L_i(s_i)\|f_i\| \geq \frac{\underline{\sigma}}{2}\|A_i s_i\|\|f_i\| \geq \frac{\underline{\sigma}}{2}\underline{A}\|s_i\|\|f_i\|.$$

Dosadíme-li tento vztah do předchozí nerovnosti, dostaneme

$$\frac{\underline{\sigma}\,\underline{A}}{2}(1-\lambda-2\overline{\rho})\|s_i\|\|f_i\| \leq -(1-\lambda-2\overline{\rho})Q_i(s_i) \leq (\overline{J}^2 + \overline{G}\,\overline{f})\|s_i\|^2,$$

neboli

$$\|s_i\| \geq \frac{\underline{\sigma}\,\underline{A}(1-\lambda-2\overline{\rho})}{2(\overline{J}^2 + \overline{G}\,\overline{f})}\,\|f_i\|,$$

(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 $i \notin N_1$, $i \in N_3$ a $i \neq 1$. Nechť $k < i$ je maximální index, pro který současně neplatí $k \notin N_1$, $k \in N_3$ 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

$$\underline{c} = \underline{\beta}\min\left(\frac{1-\overline{\omega}}{\overline{A}}, \frac{\underline{\sigma}\,\underline{A}(1-\lambda-2\overline{\rho})}{2(\overline{J}^2 + \overline{G}\,\overline{f})}, \frac{\|s_1\|}{\|f_1\|}\right).$$

□

Věta 159 *(globální konvergence). Nechť jsou splněny předpoklady lemmatu 55. Pak $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz (a) Nejprve ukážeme, že $\underline{f}_i \to 0$. 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 55 platí

$$\|s_i\| \geq \underline{c}\,\underline{\varepsilon}, \quad \forall i \in N.$$

Předpokládejme nejprve, že množina N_3 je nekonečná. Protože $N_3 \subset N_2$, můžeme psát

$$\begin{aligned}\|f_i\| - \|f_{i+1}\| &= \|f(x_i)\| - \|f(x_i + s_i)\| \geq -\overline{\rho}L_i(s_i) \\ &\geq \overline{\rho}\,\underline{\sigma}\|A_i s_i\| \geq \overline{\rho}\,\underline{\sigma}\,\underline{A}\,\underline{c}\,\underline{\varepsilon}, \quad \forall i \in N_3.\end{aligned}$$

Odtud plyne

$$\begin{aligned}\|f_1\| &\geq \lim_{i\to\infty}(\|f_1\|-\|f_{i+1}\|) = \sum_{i=1}^{\infty}(\|f_i\|-\|f_{i+1}\|)\\ &\geq \sum_{i\in N_3}(\|f_i\|-\|f_{i+1}\|) \geq \sum_{i\in N_3}\overline{\rho}\,\underline{\sigma}\,\underline{A}\,\underline{c}\,\underline{\varepsilon} = \infty,\end{aligned}$$

což dává spor. Předpokládejme nyní, že množina N_3 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 $\|s_i\| \geq \underline{c\varepsilon}$ $\forall i \in N$.

(b) Použitím ($\overline{\text{T1c}}$) dostaneme $L_i(s_i) = \|A_i s_i + f_i\| - \|f_i\| \leq 0$, takže

$$\|f_i\| \geq \|A_i s_i + f_i\| \geq \|A_i s_i\| - \|f_i\|.$$

Tato nerovnost implikuje $\|A_i s_i\| \leq 2\|f_i\|$, takže

$$\underline{A}\|s_i\| \leq \|A_i s_i\| \leq 2\|f_i\|. \tag{442}$$

Nyní ukážeme, že $\sum_{i=1}^{\infty}\|s_i\| < \infty$. Je-li množina N_3 konečná, existuje index $l \notin N_3$ takový, že $i \notin N_3$ $\forall i \geq l$. 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_3 nekonečná, můžeme tak jako v (a) psát

$$\begin{aligned}\|f_1\| &\geq \sum_{i=1}^{\infty}(\|f_i\|-\|f_{i+1}\|) \geq \sum_{i\in N_3}(\|f_i\|-\|f_{i+1}\|)\\ &\geq \overline{\rho}\,\underline{\sigma}\sum_{i\in N_3}\|A_i s_i\| \geq \overline{\rho}\,\underline{\sigma}\,\underline{A}\sum_{i\in N_3}\|s_i\|.\end{aligned}$$

Označme $N_3 = \{l_1, l_2, l_3, \ldots\}$. Použijeme-li (442) a lemma 55, 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 $\|s_{l_j+k}\| \leq \overline{\beta}\|s_{l_j+k-1}\|$ $\forall 2 \leq k \leq l_{j+1} - l_j - 1$. Platí tedy

$$\begin{aligned}\sum_{i=1}^{\infty}\|s_i\| &= \sum_{i=1}^{l_1-1}\|s_i\| + \sum_{j=1}^{\infty}\left[\|s_{l_j}\| + \sum_{k=1}^{l_{j+1}-l_j-1}\|s_{l_j+k}\|\right]\\ &\leq (l_1-1)\overline{\Delta} + \sum_{j=1}^{\infty}\|s_{l_j}\|\left[1 + \frac{2}{\underline{c}\,\underline{A}}\sum_{k=1}^{l_{j+1}-l_j-1}\overline{\beta}^{k-1}\right]\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{c}\,\underline{A}}\frac{1}{1-\overline{\beta}}\right]\sum_{i\in N_3}\|s_i\|\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{c}\,\underline{A}}\frac{1}{1-\overline{\beta}}\right]\frac{\|f_1\|}{\overline{\rho}\,\underline{\sigma}\,\underline{A}} < \infty.\end{aligned}$$

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 160 *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)–(T3) taková, že $x_i \to x^*$, kde $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Nechť*

$$\lim_{i\to\infty} \overline{\omega}_i = 0 \tag{443}$$

a

$$\lim_{i\to\infty} \frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0. \tag{444}$$

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

$$-L_i(s_i) \geq \underline{\sigma J}\|s_i\|$$

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 $\|\vartheta_i\| \to 0$, $J_i \to J(x^*)$ 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

$$-L_i(s_i) \geq \underline{\sigma}\|A_i s_i\| \geq \underline{\sigma J}\|s_i\|.$$

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 $i \in N_3$, 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

$$\begin{aligned} \rho_i(s_i) &= \frac{\|f(x_i)\| - \|f(x_i + s_i)\|}{-L_i(s_i)} \geq \frac{-L_i(s_i) - \|\vartheta_i\|\|s_i\| + o(\|s_i\|)}{-L_i(s_i)} \geq \\ &\geq 1 - \frac{\|\vartheta_i\|\|s_i\| + o(\|s_i\|)}{\underline{\sigma J}\|s_i\|} \to 1, \end{aligned}$$

neboť $\|\vartheta_i\| \to 0$. Jelikož $\overline{\rho} < 1$, existuje index $k_2 \geq k_1$ takový, že $\rho_i(s_i) \geq \overline{\rho}$, 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 $i \in N_3$ $\forall i \geq k_2$. 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 (443), (444) a ($\overline{\text{T1b}}$) platí $\|\omega_i\| \to 0$ a $\|\vartheta_i\| \to 0$, takže stejným způsobem jako v důkazu věty 157 (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 $i \in N_3 \subset N_2$. 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

$$\begin{aligned} \|s_{i+1}\| &\leq \frac{1}{\underline{J}} \|f_{i+1}\| \leq \frac{1}{\underline{J}} \left(\|f_{i+1} - f_i - A_i s_i\| + \|A_i s_i + f_i\| \right) \leq \\ &\leq \frac{\overline{J}}{\underline{J}} \left(\|\lambda_i\| + \|\omega_i\| \right) \|s_i\| < \left(\frac{1}{2} + \frac{1}{2} \right) \|s_i\| = \|s_i\|. \end{aligned}$$

Jelikož $i \in N_3$ 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

$$\frac{\|f_{i+1}\|}{\|f_i\|} \leq \frac{\|f_{i+1} - f_i - A_i s_i\| + \|A_i s_i + g_i\|}{\|f_i\|} \leq \|\lambda_i\| + \|\omega_i\|,$$

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 (J4a)–(J5a) plyne platnost podmínek (A4a)–(A5a).

Věta 161 *Nechť jsou splněny podmínky (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 156 a věty 159. Superlineární konvergence plyne bezprostředně z věty 157 a věty 160, neboť $\vartheta_i = 0$ $\forall i \in N$. □

Poznámka 271 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 56 *Nechť je splněn předpoklad (J6) a nechť*

$$Ae_j = \frac{f(x+\delta e_j) - f(x)}{\delta} \tag{445}$$

pro $1 \le j \le n$, *kde* e_j, $1 \le j \le n$, *jsou sloupce jednotkové matice řádu* n. *Pak platí*

$$\|A - J(x)\| \le \frac{1}{2}\overline{G}\sqrt{n}\delta.$$

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 162 *Nechť jsou splněny předpoklady (J5a) a (J6). Je-li matice* A *určena podle vzorce (445), 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 56 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 lemmatu 54 □

Poznámka 272 Věta 162 ukazuje, že lze zvolit diferenci $\delta > 0$ tak, aby matice určená podle vztahu (445) 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 $\underline{J}$ v (J5a) a čím větší je číslo $\overline{G}$ v (J6).

11.5 Kvazinewtonovské metody

Definice 48 *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 44) je kvazinewtonovskou metodou, jestliže*

$$A_i s_i + f_i = 0, \tag{446}$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i + u_i v_i^T, \tag{447}$$

kde $u_i \in R^n$, $v_i \in R^n$, a vyhovující podmínce

$$A_{i+1} d_i = y_i, \tag{448}$$

kde $y_i = f_{i+1} - f_i$, $d_i = x_{i+1} - x_i$.

Poznámka 273 V tomto oddílu se budeme zabývat pouze přesnými kvazinewtonovskými metodami (podmínka (446)), takže $(A_i s_i + f_i)/\|f_i\| = 0 \ \forall i \in N$. Neplatí však $(A_i - J_i)s_i/\|s_i\| = 0 \ \forall i \in N$ (matice A_i se mohou od matic J_i dosti lišit).

Věta 163 *Nechť $A_+ = A + uv^T$ a $Ad \neq y$. Pak $A_+ d = y$ právě tehdy, jestliže $v^T d \neq 0$ a $u = (y - Ad)/v^T d$, takže*

$$A_+ = A + \frac{(y - Ad)v^T}{v^T d}. \tag{449}$$

Jestliže $Ad = y$ stačí položit $u = v = 0$, takže $A_+ = A$.

Důkaz Z podmínky $A_+ d = y$ dostaneme $A_+ d = Ad + uv^T d = y$. Jestliže $Ad = y$, stačí položit $u = v = 0$, takže $A_+ = A$. Jestliže $Ad \neq y$, musí platit $v^T d \neq 0$ a $u = (y - Ad)/v^T d$. □

Poznámka 274 Položíme-li $v = d$ dostaneme Broydenovu dobrou metodu

$$A_+ = A + \frac{(y - Ad)d^T}{d^T d}. \tag{450}$$

Položíme-li $v = A^T y$, dostaneme Broydenovu špatnou metodu

$$A_+ = A + \frac{(y - Ad)y^T A}{y^T Ad}. \tag{451}$$

Nechť

$$e_k^T d = \max_{1 \leq i \leq n} e_i^T d.$$

Položíme-li $v = e_k$, dostaneme přímou metodu aktualizace sloupců

$$A_+ = A + \frac{(y - Ad)e_k^T}{e_k^T d}, \tag{452}$$

která aktualizuje vždy pouze jeden sloupec matice A.

Věta 164 *Nechť A je regulární matice a nechť platí (449). Pak matice A_+ je regulární právě tehdy, jestliže $v^T A^{-1} y \neq 0$.*

Důkaz Nechť $A_+ = A + uv^T$. Pak podle Shermanova-Morrisonova vzorce (poznámka 84) platí

$$A_+^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}, \tag{453}$$

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, jestliže $v^TA^{-1}y \neq 0$. □

Poznámka 275 Věta 164 opodstatňuje použití Broydenovy špatné metody. Jestliže $y \neq 0$ a matice A je regulární, pak volba $v = A^Ty$ dává $v^TA^{-1}y = y^TAA^{-1}y = y^Ty = \|y\|^2 \neq 0$.

Věta 165 *(Aktualizace matice $S = A^{-1}$). Nechť jsou splněny předpoklady věty 164. Nechť $S = A^{-1}$ a $S_+ = A_+^{-1}$, kde A_+ je matice určená podle aktualizace (449) s $v^TA^{-1}y \neq 0$. Pak platí*

$$S_+ = S + \frac{(d - Sy)v^TS}{v^TSy}, \tag{454}$$

neboli

$$S_+ = S + \frac{(d - Sy)z^T}{z^Ty}, \tag{455}$$

kde $z = S^Tv$.

Důkaz Podle (453) platí

$$S_+ = S - \frac{Suv^TS}{\delta} = S + \frac{(d - Sy)v^TS}{\delta v^Td},$$

kde δ je zatím neznámé číslo. Z rovnice $S_+y = d$ však plyne

$$S_+y = Sy + \frac{v^TSy}{\delta v^Td}(d - Sy) = d.$$

takže nutně $\delta = v^TSy/v^Td$. □

Poznámka 276 Položíme-li $v = d$, dostaneme Broydenovu dobrou metodu

$$S_+ = S + \frac{(d - Sy)d^TS}{d^TSy}. \tag{456}$$

Položíme-li $v = (S^{-1})^Ty$, neboli $z = y$, dostaneme Broydenovu špatnou metodu

$$S_+ = S + \frac{(d - Sy)y^T}{y^Ty}. \tag{457}$$

Nechť

$$e_k^Ty = \max_{1\leq i\leq n} e_i^Ty.$$

Položíme-li $S^Tv = e_k$, neboli $z = e_k$, dostaneme inverzní metodu aktualizace sloupců

$$S_+ = S + \frac{(d - Sy)e_k}{e_k^Ty}. \tag{458}$$

Poznámka 277 (Dualita). Vztah (454) dostaneme ze vztahu (449) 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 278 Prakticky použitelná je pouze dobrá Broydenova metoda a přímá metoda aktualizace sloupců. Metody k nim duální (špatná Broydenova metoda a inverzní metoda aktualizace sloupců) jsou méně efektivní.

Poznámka 279 Kvazinewtonovské metody lze také odvozovat pomocí minimalizačních principů. Podle věty 147 pro dobrou Broydenovu metodu platí

$$\|A_+ - A\|_F = \min_{\tilde{A}d=y} \|\tilde{A} - A\|_F.$$

Podobně pro špatnou Broydenovu metodu platí

$$\|S_+ - S\|_F = \min_{\tilde{S}y=d} \|\tilde{S} - S\|_F.$$

Kvazinewtonovské metody splňují kvazinewtonovskou podmínku podobně jako metody s proměnnou metrikou (stačí porovnat (448) a (104)). 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

$$J^*(x - x^*) = 0 \tag{459}$$

s regulární maticí J^* také po konečném počtu kroků. Při důkazu tohoto tvrzení budeme používat vyjádření

$$x_{i+1} = x_i - S_i f_i \tag{460}$$

a

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} \tag{461}$$

$\forall i \in N$, kde S_i jsou regulární matice, $f_i \neq 0$ a $z_i^T y_i \neq 0$ $\forall i \in N$ (zde $z_i = S_i^T v_i$).

Lemma 57 *Uvažujme iterační proces (460), (461) aplikovaný na soustavu lineárních rovnic (459) 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 (459) a (460) plyne

$$y_i = f_{i+1} - f_i = J^* d_i = -J^* S_i f_i, \tag{462}$$

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 (461) a (462), 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 58 *Nechť jsou splněny předpoklady lemmatu 57 a nechť $i \in N$ je index takový, že vektor f_{i+1} není násobkem vektoru f_i. Pak vektory $(J^*S_{i-2l+2})^k f_{i-2l+2}$, $0 \le k \le l$, jsou lineárně nezávislé pro každé číslo $l \in N$ takové, že $2l \le i+1$.*

Důkaz (indukcí). Předpokládejme, že vektory $(J^*S_{i-2l+2})^k f_{i-2l+2}$, $0 \le k \le l$, jsou lineárně nezávislé pro nějaké číslo $l \in N$ takové, že $2l \le i-1$. Platí to zcela jistě pro $l = 1$, neboť podle (462) 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 57 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}, $0 \le j \le k$. Jelikož $l+1$ lineárně nezávislých vektorů $(J^*S_{i-2l+2})^k f_{i-2l+2}$, $0 \le k \le l$, vyjadřujeme pomocí $l+1$ vektorů $(I - J^*S_{i-2l+1})(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, musí být tyto vektory také lineárně nezávislé. Odtud bezprostředně plyne, že i vektory $(J^*S_{i-2l+1})^k f_{i-2l+1}$, $0 \le k \le l$, jsou lineárně nezávislé.
(b) Použijeme-li (462), 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

$$(I - J^*S_{i-2l+1})y_{i-2l} = 0. \tag{463}$$

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 (463), což je spor s lineární nezávislostí těchto vektorů (viz (a)).
(c) Podle lemmatu 57 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 166 *Nechť jsou splněny předpoklady lemmatu 57. 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 (459).*

Důkaz Předpokládejme, že pro $i = 2n-1$ není vektor f_{i+1} násobkem vektoru f_i. Pak podle lemmatu 58 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 (461) a (462) 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 (459). □

Nevýhodou kvazinewtonovských metod realizovaných standardním způsobem (definice 46 a definice 47) je to, že není zaručena jejich globální konvergence (matice A_i, $i \in N$, mohou být obecně špatnými aproximacemi Jacobiových matic J_i, $i \in N$). Proto je třeba tyto metody kombinovat s diferenční verzí Newtonovy metody. Kvazinewtonovské metody spádových směrů se obvykle realizují tak, že se pokládá $A_1 = J_1$ a kdykoliv nelze splnit podmínku ($\overline{\text{S2a}}$) (nebo ($\overline{\text{S2b}}$), nebo ($\overline{\text{S2c}}$)), iterační proces se přeruší a položí se $A_{i+1} = J_{i+1}$ (algoritmus 19). Kvazinewtonovské metody s lokálně omezeným krokem se obvykle realizují tak, že se pokládá $A_1 = J_1$ a v případě ($\overline{\text{T3a}}$), se položí $A_{i+1} = J_{i+1}$ zatímco v případě ($\overline{\text{T3b}}$) se matice A_{i+1} aktualizuje podle (449). Tyto úpravy mají své opodstatnění, neboť platí tvrzení, které je speciálním případem vět 182 a 183.

Tvrzení 8 *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 dobrou Broydenovou metodou (450) s jednotkovým výběrem délky kroku ($\alpha_i = 1$, $i \in N$), konverguje Q-superlineárně k bodu $x^* \in R^n$.*

Následující tabulka ukazuje srovnání několika metod pro řešení soustav nelineárních rovnic: MNDER – Newtonova metoda s Jacobiovou maticí počítanou analyticky, MNDIF – Newtonova metoda s Jacobiovou maticí počítanou numericky, QNDER – Broydenova dobrá metoda s Jacobiovou maticí počítanou analyticky, QNDIF – Broydenova dobrá metoda s Jacobiovou maticí počítanou numericky; při řešení 62 testovacích rovnic s 100 neznámými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a Jacobiových matic NFJ, jakož i počet selhání F a celkový čas výpočtu). Tyto metody byly realizovány jako metody s lokálně omezeným krokem. Newtonova metoda používá LU rozklad a Broydenova dobrá metoda používá aktualizaci QR rozkladu.

Metoda	NIT	NFV	NFJ	F	čas
NMDER	1133	1334	1333	1	1.53
NMDIF	1034	104543	0	1	8.24
QMDER	1901	2205	225	0	2.42
QNDIF	1636	19343	0	0	3.07

11.6 Nemonotonní kvazinewtonovské metody

Matice A_i, $i \in N$, generované kvazinewtonovkými metodami obecně nesplňují podmínky (A3b)–(A5b) s (429). 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

$$\sum_{i=1}^{\infty} \eta_i = \overline{\eta} < \infty. \qquad (464)$$

Lemma 59 *(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 (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 ($\overline{\text{S2e}}$) 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 49 *Nemonotonní metodou spádových směrů nazveme metodu spádových směrů (definice 46), kde podmínka (S2c) je nahražena podmínkou (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 (464).*

Věta 167 *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 49), 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 (S2c) a (464) 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 52 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 (464) 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 (S2c) však platí

$$\|f_{k_2}\| \leq \|f_{k_1}\| + \sum_{i=k_1}^{k_2} \eta_i \leq \underline{\varepsilon} + \sum_{i=k}^{\infty} \eta_i < \underline{\varepsilon} + (\overline{\varepsilon} - \underline{\varepsilon}) = \overline{\varepsilon},$$

což je ve sporu s předpokladem, že $\|f_{k_2}\| \geq \overline{\varepsilon}$. □

Poznámka 280 V předpokladech věty 167 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 (S1)).

Poznámka 281 Věta 167 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

$$M = \{l \in N :\ l = (j-1)m + 1,\ j \in N\}, \tag{465}$$

a kde matice A_i, $i \notin M$, splňují slabý princip omezeného znehodnocení

$$\|A_i - J_i\| \leq c_1\|A_{i-1} - J_{i-1}\| + c_2\|x_i - x_{i-1}\| \tag{466}$$

($c_1 > 0$ a $c_2 > 0$ jsou konstanty nezávislé na indexu $i \notin M$).

Lemma 60 *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 49), pro kterou platí (465) a (466). 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 $1 \le k \le m$. Dokážeme indukcí, že pro libovolný index $1 \le k \le m$ platí

$$\lim_{l\to\infty} \|A_{l+k-1} - J_{l+k-1}\| = 0, \qquad \lim_{l\to\infty} \|x_{l+k} - x_{l+k-1}\| = 0. \tag{467}$$

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 (435), podle které

$$\|x_{l+1} - x_l\| = \|\alpha_l s_l\| \le \|s_l\| \le \frac{1+\overline{\omega}}{\underline{J}} \|f_l\|$$

(neboť $0 < \alpha_l \le 1$), což spolu s $\|f_l\| \to 0$ (věta 167) dává $\|x_{l+1} - x_l\| \to 0$. Předpokládejme nyní, že (467) platí pro nějaký index $1 \le k < m$ (dokázali jsme to pro $k = 1$). Použijeme-li (466) dostaneme

$$\lim_{l\to\infty} \|A_{l+k} - J_{l+k}\| \le c_1 \lim_{l\to\infty} \|A_{l+k-1} - J_{l+k-1}\| + c_2 \lim_{l\to\infty} \|x_{l+k} - x_{l+k-1}\| = 0.$$

To znamená, že pro dostatečně velký index $l \in M$ splňuje matice A_{l+k} podmínku (A3a) s (428) a (A5a). Můžeme tedy použít (435), 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 (467) plyne existence čísel $n_k \in N$, $1 \le k \le m$, takových, že pro $l \ge n_k$ 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 $i \ge n$ platí $\|A_i - J_i\| < \varepsilon$ a $\|x_{i+1} - x_i\| < \varepsilon$ a jelikož číslo ε bylo zvoleno libovolně, dostaneme tvrzení lemmatu. □

Věta 168 *Nechť jsou splněny předpoklady lemmatu 60. 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 (S2d), pak $x_i \to x^*$ Q-superlineárně.*

Důkaz (a) Nechť $\underline{J} \le (1/2)\|J^{-1}(x^*)\|^{-1}$. 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 $J(x)d \ge \underline{J}\|d\|$ $\forall x \in B(x^*, \delta)$. Nechť $0 < \varepsilon < \delta$ a $P(x^*, \varepsilon, \delta) = \{x \in R^n : \varepsilon < \|x - x^*\| < \delta\}$. Pak podle věty 155 $P(x^*, \varepsilon, \delta)$ 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 $i \ge k_1$. Jelikož $x_i \to x^*$, existuje index $k_2 \ge k_1$ takový, že $\|x_i - x^*\| < \varepsilon$, pokud $i \ge k_2$ a jelikož $\|x_{i+1} - x_i\| \to 0$, existuje index $k \ge k_2$ takový, že $\|x_{i+1} - x_i\| < \delta - \varepsilon$, pokud $i \ge k$. Pak ale $\|x_{i+1} - x^*\| \le \|x_i - x^*\| + \|x_{i+1} - x_i\| < \delta$, pokud $i \ge k$, a jelikož $x_{i+1} \notin P(x^*, \varepsilon, \delta)$, musí být $\|x_{i+1} - x^*\| < \varepsilon$, pokud $i \ge k$. Postupujeme-li takto dále, dostaneme $\|x_i - x^*\| < \varepsilon$ $\forall i \ge k$ a jelikož číslo $\varepsilon > 0$ bylo vybráno libovolně, platí $x_i \to x^*$.

(b) Podle lemmatu 60 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 157, takže $x_i \to x^*$ Q-superlineárně. □

Na závěr ukážeme, že kvazinewtonovské metody většinou splňují slabý princip omezeného znehodnocení, takže pro ně platí věta 168.

Věta 169 *Nechť zobrazení $f : R^n \to R^n$ splňuje podmínku (J6) a nechť*

$$A_{i+1} = A_i + \frac{(y_i - A_i d_i)v_i^T}{v_i^T d_i}, \tag{468}$$

kde $d_i = x_{i+1} - x_i$, $y_i = f_{i+1} - f_i$ a $|v_i^T d_i| \ge \underline{\gamma}\|v_i\|\|d_i\|$. Pak

$$\|A_{i+1} - J_{i+1}\| \le c_1\|A_i - J_i\| + c_2\|d_i\|.$$

kde $c_1 = 1 + 1/\underline{\gamma}$ a $c_2 = \overline{G}(1 + 1/\underline{\gamma})$.

Důkaz Použijeme-li (468), 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 5) a (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 11.8). 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 $J_+^T v$, 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 282 Nahradíme-li v aktualizaci (450) vektor y vektorem $J_+ d$, dostaneme

$$A_+ = A - \frac{(A - J_+)dd^T}{d^T d}.$$

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 283 Položíme-li

$$A_+ = A - \frac{f_+ f_+^T (A - J_+)}{f_+^T f_+},$$

platí $f_+^T A_+ = f_+^T J_+$, takže vektor $A_+^T f_+$ 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

$$g_+^T s_+ = f_+^T J_+ s_+ = f_+^T A_+ s_+ = -f_+^T f_+ < 0,$$

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 50 *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 44) je sdruženou kvazinewtonovskou metodou, jestliže*

$$A_i s_i + f_i = 0,$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i - \frac{w_i w_i^T (A_i - J_{i+1})}{w_i^T w_i}, \tag{469}$$

kde $w_i \in R^n$ je vektor takový, že $w_i^T w_i \neq 0$.

Poznámka 284 Položíme-li

$$A_+ = A - \frac{(A - J_+) d\, f_+^T (A - J_+)}{f_+^T (A - J_+) d},$$

platí současně $A_+ d = J_+ d$ a $f_+^T A_+ = f_+^T J_+$.

Metody tohoto typu nazýváme oboustrannými kvazinewtonovskými metodami. Obecný tvar oboustranných kvazinewtonovských metod je definován takto.

Definice 51 *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic (definice 44) je oboustrannou kvazinewtonovskou metodou, jestliže*

$$A_i s_i + f_i = 0,$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i - \frac{(A_i - J_{i+1}) d_i w_i^T (A_i - J_{i+1})}{w_i^T (A_i - J_{i+1}) d_i}, \tag{470}$$

kde $w_i \in R^n$ je vektor takový, že $w_i^T (A_i - J_{i+1}) d_i \neq 0$.

Oboustranné kvazinewtonovské metody mají důležitou vlastnost lineárního ukončení (naleznou řešení soustavy lineárních rovnic po nejvýše $n + 1$ krocích).

Věta 170 *(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

$$(A_i - J) d_j = 0, \tag{471}$$

$$w_j^T (A_i - J) = 0. \tag{472}$$

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

$$f_{i+1} = J(x_i + d_i - x^*) = f_i + J d_i \neq 0,$$

musí platit

$$(A_i - J) d_i = A_i s_i + f_i - J d_i - f_i = -(f_i + J d_i) \neq 0,$$

takže vektor d_i nemůže být lineární kombinací vektorů d_j, $1 \leq j < i$, pro které platí (471).

(b) Použijeme-li (470), můžeme psát

$$A_{i+1} - J = A_i - J - \frac{(A_i - J)d_i w_i^T (A_i - J)}{w_i^T (A_i - J)d_i}. \tag{473}$$

Z (471) a (473) 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 (472) a (473) plyne, že $w_j^T (A_{i+1} - J) = 0$ pro $1 \leq j < i$. Dále platí

$$w_i^T (A_{i+1} - J)) = w_i^T (A_i - J) - w_i^T (A_i - J) = 0,$$

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 (471) 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 285 Vlastnost (471) (nebo (472)) se nazývá dědičností. Kvazinewtonovské metody vyšetřované v oddílu 11.5 ani obecné sdružené kvazinewtonovské metody tuto vlastnost nemají.

Poznámka 286 Sdružené kvazinewtonovské metody vyhovují sdruženému minimalizačnímu principu. Platí

$$\|A_+ - A\|_F = \min_{\tilde{A}^T w = J_+^T w} \|\tilde{A} - A\|_F.$$

Plyne to z věty 147 aplikované na transponovanou matici.

Poznámka 287 Nejznámější sdružené kvazinewtonovské metody dostaneme, položíme-li

$$w = (A - J_+)d \tag{474}$$

(tečná sdružená Broydenova metoda), nebo

$$w = Ad - (f_+ - f) \tag{475}$$

(sečná sdružená Broydenova metoda), nebo

$$w = f_+ \tag{476}$$

(reziduální sdružená Broydenova metoda). Poznamenejme, že tečná sdružená Broydenova metoda je oboustrannou sdruženou kvazinewtonovskou metodou tvaru (470), kde $w = (A - J_+)d$. Reziduální sdružená Broydenova metoda je uvedená v poznámce 283. Tato metoda je ekvivalentní sečné sdružené Broydenově metodě, pokud $Ad = f$ (řešíme-li přesné soustavu lineárních rovnic a používáme-li jednotkovou délku kroku). Všechny sdružené Broydenovy metody používají vektor $J_+^T w$ a tečná sdružená Broydenova metoda navíc vektor $J_+ d$.

Další užitečnou vlastností sdružených i oboustranných kvazinewtonovských metod je jejich invariantnost vzhledem k lineární transformaci proměnných.

Věta 171 *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 (469), dostaneme

$$A_+ = A - \frac{ww^T(A - J_+)}{w^T w} = \tilde{A}T^{-1} - \frac{ww^T(\tilde{A} - \tilde{J}_+)T^{-1}}{w^T w} = \tilde{A_+}T^{-1}.$$

Použijeme-li vztah $J_+ = \tilde{J}_+ T^{-1}$ a (470), dostaneme

$$A_+ = A - \frac{(A - J_+)dw^T(A - J_+)}{w^T(A - J_+)d} = \tilde{A}T^{-1} - \frac{(\tilde{A} - \tilde{J}_+)T^{-1}T\tilde{d}w^T(\tilde{A} - \tilde{J}_+)T^{-1}}{w^T(\tilde{A} - \tilde{J}_+)T^{-1}T\tilde{d}} = \tilde{A_+}T^{-1},$$

neboť $d = x_+ - x = T(\tilde{x}_+ - \tilde{x}) = T\tilde{d}$. □

Poznámka 288 Broydenova dobrá metoda uvedená v oddílu 11.5 ani metoda uvedená v poznámce 282 nejsou invariantní vzhledem k lineární transformaci proměnných.

Vynikající vlastností reziduální sdružené Broydenovy metody je snadné určení přesného gradientu funkce $F = (1/2)\|f\|^2$. Podle poznámky 283 platí $h_i = A_i^T f_i = J_i^T f_i = g_i$ $\forall i \in N$, takže je splněna podmínka (A3b) s $\overline{\vartheta} = 0$. V důsledku toho platí tvrzení lemmatu 51 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 266). 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 (S2a)–(S2c) a položíme $x_{i+1} = x_i + \alpha_i s_i$ a $A_{i+1} = A_i - f_{i+1}f_{i+1}^T(A_i - J_{i+1})/f_{i+1}^T f_{i+1}$. V opačném případě (pokud $\|s_i\| > \overline{c}\|f_i\|$), položíme $x_{i+1} = x_i$ a $A_{i+1} = J_{i+1}$. Výsledkem těchto úvah je následující věta.

Věta 172 *Nechť zobrazení $f : \mathcal{D} \to R^n$ vyhovuje předpokladům (J1), (J4), (J5a) a (J6). Pak upravená reziduální sdružená Broydenova metoda je globálně konvergentní.*

Ostatní sdružené kvazinewtonovské metody tuto výjimečnou vlastnost nemají. Proto je musíme realizovat pomocí algoritmu 19. Tento přístup má své opodstatnění, neboť platí tvrzení, které je speciálním případem vět 185 a 186.

Tvrzení 9 *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 (AB1), (AB2) (nebo (AB3), když $\overline{\omega} = 0$) s jednotkovým výběrem délky kroku ($\alpha_i = 1$, $i \in N$) konverguje Q-superlineárně k bodu $x^* \in R^n$.*

11.8 Tenzorové metody

V oddílu 7.2 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 7.2 jsme tenzor T nazaváděli, pracovali jsme pouze s maticemi B_k, $1 \le k \le m$). 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 $q(x+s) = f(x) + J(x)s + (1/2)s^T Ts = 0$. 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 $B_k^i \approx G_k(x_i)$, $1 \le k \le m$, splňují p interpolačních podmínek

$$f_k(x_i + d_j^i) = f_k(x_i) + g_k^T(x_i)d_j^i + \frac{1}{2}(d_j^i)^T B_k^i d_j^i, \qquad 1 \le j \le p,$$

kde $d_j^i = x_{i-j} - x_i$, $1 \le j \le p$. 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 173 *Matice B má minimální Frobeniovu normu na množině zadané interpolačními rovnostmi*

$$z_j = 2(f(x+d_j) - f(x) - g^T(x)d_j = d_j^T B d_j, \qquad 1 \le j \le p$$

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

$$M = \begin{bmatrix} (d_1^T d_1)^2, & \ldots, & (d_1^T d_n)^2 \\ \ldots, & \ldots, & \ldots \\ (d_n^T d_1)^2, & \ldots, & (d_n^T d_n)^2 \end{bmatrix}.$$

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 173. Pak lze předchozí soustavu rovnic zapsat ve tvaru $Mu = z$ takže $u = M^{-1}z$. □

Poznámka 289 Vzorce uvedené ve větě 173 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

$$T = \sum_{l=1}^{p} (Ue_l) \times d_l \times d_l, \tag{477}$$

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 290 Použijeme-li vzorec (477) (s indexem j místo l), můžeme soustavu kvadratických rovnic, sloužících k určení směrového vektoru s, zapsat ve tvaru

$$q(x+s) = f(x) + J(x)s + \frac{1}{2}\sum_{j=1}^{p} Ue_j (d_j^T s)^2 = 0. \tag{478}$$

Předpokládejme, že matice J je regulární a označme $\beta_j = d_j^T s$, $1 \le j \le p$. Pak platí

$$s = -J^{-1}\left(f - \frac{1}{2}\sum_{j=1}^{p} Ue_j \beta_j^2\right). \tag{479}$$

Vynásobíme-li tento vztah postupně vektory d_i, $1 \le i \le p$, dostaneme soustavu kvadratických rovnic

$$w_i^T f + \beta_i + \frac{1}{2}\sum_{j=1}^{p} w_i^T U e_j \beta_j^2 = 0, \qquad 1 \le i \le p \tag{480}$$

(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 290 předpokládáme, že matice J je regulární a soustava rovnic (480) má řešení. V tomto případě má i soustava (478) řešení, které dostaneme, dosadíme-li β_j, $1 \le j \le p$ do (479). Nemá-li soustava (478) 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 $w_i = (J^{-1})^T d_i$, $1 \le i \le p$ (takže $W = (J^{-1})^T D$ a $W^T J s = D^T s = \beta$). Pak můžeme soustavu (480) zapsat ve tvaru

$$\tilde{q}(\beta) = W^T q(x+s) = W^T f + \beta + \frac{1}{2} W^T U \beta^2 = 0, \tag{481}$$

kde $\beta = [\beta_1, \ldots, \beta_p]$ a $\beta^2 = [\beta_1^2, \ldots, \beta_p^2]$.

Věta 174 *Nechť matice J je regulární a nechť vektor $\beta \in R^p$ minimalizuje normu $\|(W^TW)^{-1/2}\tilde{q}(\beta)\|$. Pak vektor*

$$s = -J^{-1}\left(f + \frac{1}{2}U\beta^2 - W(W^TW)^{-1}\tilde{q}(\beta)\right) \tag{482}$$

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

$$V = W(W^TW)^{-1/2}, \qquad W^T Z = D^T J^{-1} Z = 0$$

(vynásobením se snadno přesvědčíme, že $V^TV = I$). Použijeme-li (481), dostaneme

$$Qq(x+s) = \begin{bmatrix} V^Tq(x+s) \\ Z^Tq(x+s) \end{bmatrix} = \begin{bmatrix} (W^TW)^{-1/2}W^Tq(x+s) \\ Z^Tq(x+s) \end{bmatrix} = \begin{bmatrix} (W^TW)^{-1/2}\tilde{q}(\beta) \\ Z^Tq(x+s) \end{bmatrix}.$$

Z tohoto vyjádření je patrné, že pokud vektor β minimalizuje normu $\|(W^TW)^{-1/2}\tilde{q}(\beta)\|$ a pokud s je vektor takový, že $D^Ts = \beta$ a $Z^Tq(x+s) = 0$, minimalizuje tento vektor normu $\|Qq(x+s)\| = \|q(x+s)\|$. Položme

$$s = J^{-1}W(W^TW)^{-1}\beta + J^{-1}Zr, \tag{483}$$

kde $r \in R^{n-p}$. Pak platí

$$D^Ts = D^TJ^{-1}W(W^TW)^{-1}\beta + D^TJ^{-1}Zr = W^TW(W^TW)^{-1}\beta + W^TZr = \beta,$$

a

$$\begin{aligned} Z^Tq(x+s) &= Z^T\left(f + Js + \frac{1}{2}U\beta^2\right) = Z^Tf + Z^TW(W^TW)^{-1}\beta + Z^TZr + \frac{1}{2}Z^TU\beta^2 \\ &= Z^Tf + r + \frac{1}{2}Z^TU\beta^2, \end{aligned}$$

takže $Z^Tq(x+s) = 0$, pokud $r = -Z^T(f + (1/2)U\beta^2)$. Z ortogonality matice Q plyne, že $VV^T + ZZ^T = I$, takže $ZZ^T = I - VV^T = I - W(W^TW)^{-1}W^T$ a tedy

$$Zr = -ZZ^T\left(f + \frac{1}{2}U\beta^2\right) = -\left(f + \frac{1}{2}U\beta^2\right) + W(W^TW)^{-1}W^T\left(f + \frac{1}{2}U\beta^2\right)$$

Dosadíme-li tento vektor do (483) a použijeme-li (481), dostaneme

$$\begin{aligned} s &= J^{-1}W(W^TW)^{-1}\beta - J^{-1}\left(f + \frac{1}{2}U\beta^2\right) + J^{-1}W(W^TW)^{-1}W^T\left(f + \frac{1}{2}U\beta^2\right) \\ &= -J^{-1}\left(f + \frac{1}{2}U\beta^2 + W(W^TW)^{-1}\tilde{q}(\beta)\right). \end{aligned}$$

□

Poznámka 291 Víme-li, že existuje vektor s takový, že $Z^Tq(x+s) = 0$, můžeme vzorec (482) odvodit jednodušším způsobem. Jelikož

$$q(x+s) = Q^TQq(x+s) = [V, Z]\begin{bmatrix} (W^TW)^{-1/2}\tilde{q}(\beta) \\ 0 \end{bmatrix} = V(W^TW)^{-1/2}\tilde{q}(\beta) = W(W^TW)^{-1}\tilde{q}(\beta),$$

platí

$$f(x) + J(x)s + \frac{1}{2}U\beta^2 = W(W^TW)^{-1}\tilde{q}(\beta),$$

odkud bezprostředně plyne (482).

Poznámka 292 Nechť $W^TW = R^TR$, kde R je horní trojúhelníková matice. Pak minimalizace normy $\|(W^TW)^{-1/2}\tilde{q}(\beta)\|$ je ekvivalentní minimalizaci funkce

$$\frac{1}{2}\|(W^TW)^{-1/2}\tilde{q}(\beta)\|^2 = \frac{1}{2}\tilde{q}^T(\beta)(W^TW)^{-1}\tilde{q}(\beta) = \frac{1}{2}\tilde{q}^T(\beta)(R^TR)^{-1}\tilde{q}(\beta),$$

což je vážený součet čtverců, jehož minimum lze nalézt metodami popsanými v kapitole 7. Zdůrazněme, že vektor proměnných β má dimenzi $p \ll n$.

Poznámka 293 Je-li matice J singulární, řešíme příslušné soustavy lineární rovnic ve smyslu nejmenších čtverců, takže místo matice J^{-1} používáme pseudoinverzi $J^\dagger$ (věta 109). Pak můžeme psát $W = (J^\dagger)^T D$ a $s = -J^\dagger \left(f + (1/2)U\beta^2 - W(W^TW)^{-1}\tilde{q}(\beta)\right)$.

Poznámka 294 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^TW 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 $U = [u_1] = u = z/(d^Td)^2$, takže

$$\tilde{q}(\beta) = w^T f + \beta + \frac{1}{2} w^T u \beta^2. \tag{484}$$

Věta 175 *Nechť $p = 1$. Pak pro směrový vektor určený tenzorovou metodou platí*

$$\begin{aligned} s &= -J^{-1}\left(f + \frac{1}{2}u\beta^2\right), & w^Tuw^Tf \leq 1, \\ s &= -J^{-1}\left(f + \frac{1}{2}u\beta^2 - \frac{w}{w^Tw}\tilde{q}(\beta)\right), & w^Tuw^Tf > 1, \end{aligned}$$

kde

$$\begin{aligned} \beta &= \frac{w^Tf}{1 + \sqrt{1 - w^Tuw^Tf}}, & w^Tuw^Tf \leq 1, \\ \beta &= -\frac{1}{w^Tu}, & w^Tuw^Tf > 1. \end{aligned}$$

Důkaz Vzorce pro směrový vektor plynou bezprostředně z (482). Podle (484) je rovnice $\tilde{q}(\beta) = 0$ kvadratickou rovnicí, která má řešení pokud její diskriminant $1 - w^Tuw^Tf$ je nezáporný. V tomto případě platí

$$\beta = \frac{-1 \pm \sqrt{1 - w^Tuw^Tf}}{w^Tu} = \frac{w^Tf}{1 \pm \sqrt{1 - w^Tuw^Tf}}.$$

(znaménko bereme tak, aby číslo β bylo v absolutní hodnotě co nejmenší). Pokud $1 - w^Tuw^Tf < 0$, rovnice $\tilde{q}(\beta) = 0$ nemá řešení a číslo β musíme určit minimalizací funkce $\tilde{q}(\beta)(w^Tw)^{-1}\tilde{q}(\beta)$, 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 $\tilde{q}'(\beta) = 1 + w^Tu\beta = 0$, což dává $\beta = -1/w^Tu$. □

Abychom dostali účinný algoritmus odolný vůči selhání, je třeba tenzorovou metodu kombinovat s Newtonovou metodou. Jedna z možností je použita v následujícím algoritmu.

Algoritmus 20 Data $0 < \underline{\rho} < 1$, $0 < \underline{\beta} \leq \overline{\beta} < 1$, $\overline{\varepsilon} > 0$, $0 < \underline{k} \leq \overline{k}$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$, vypočteme $f_1 = f(x_1)$ a položíme $i = 1$.

Krok 2 Pokud $\|f_i\| \leq \overline{\varepsilon}$, 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 $s_i^S = s_i^N - J^{-1}(f + (1/2)U\beta^2 - W(W^TW)^{-1}\tilde{q}(\beta))$, kde β je vektor, který minimalizuje normu $\|(W^TW)^{-1/2}\tilde{q}(\beta))\|$ (věta 174).

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 (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 $\underline{\beta}\alpha_i^k \leq \alpha_i^{k+1} \leq \overline{\beta}\alpha_i^k$, 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 (449), 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 213). Pak řešení soustavy rovnic $QR + f = 0$ vyžaduje $O(n^2)$ operací násobení a sčítání. Ukážeme nyní, jak lze určit ortogonální rozklad matice $\bar{A} = A + uv^T$ z ortogonálního rozkladu matice A s použitím $O(n^2)$ operací násobení a sčítání.

Věta 176 *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

$$A + uv^T = Q(R + Q^T uv^T) = Q(R + \tilde{u}v^T). \tag{485}$$

Podle poznámky 209 existuje ortogonální matice $\tilde{Q}^T = \tilde{Q}_{12}^T\tilde{Q}_{23}^T \ldots \tilde{Q}_{n-1,n}^T$ 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 210 existuje ortogonální matice $\hat{Q}^T = \hat{Q}_{n-1,n}^T \ldots \hat{Q}_{23}^T\hat{Q}_{12}^T$ taková, že matice $\bar{R} = \hat{Q}^T(\tilde{R} + \|\tilde{u}\|e_1 v^T)$ je horní trojúhelníková. Po dosazení do (485) dostaneme

$$\begin{aligned} \bar{A} &= A + uv^T = Q(R + \tilde{u}v^T) = Q\tilde{Q}(\tilde{Q}^T R + \tilde{Q}^T\tilde{u}v^T) = Q\tilde{Q}(\tilde{R} + \|\tilde{u}\|e_1 v^T) \\ &= Q\tilde{Q}\hat{Q}\hat{Q}^T(\tilde{R} + \|\tilde{u}\|e_1 v^T) = Q\tilde{Q}\hat{Q}\bar{R} = \bar{Q}\bar{R}. \end{aligned}$$

□

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 8–10.

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

$$M = \{l \in N : \ l = (j-1)m + 1, \ j \in N\}, \tag{486}$$

kde $m \in N$, vzorce (454), (455), a standardní označení $d_i = x_{i+1} - x_i$, $y_i = f_{i+1} - f_i$, $i \in N$.

Definice 52 *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*

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) v_i^T S_i}{v_i^T S_i y_i} = (I + u_i v_i^T) S_i, \qquad u_i = \frac{d_i - S_i y_i}{v_i^T S_i y_i}$$

pro $l \le i \le l + m - 2$, kde $v_i \in R^n$ je zvolený vektor.

Poznámka 295 Pokud položíme $v_i = d_i$, dostaneme Broydenovu dobrou metodu. Položíme-li $v_i = e_k$, dostaneme přímou metodu aktualizace sloupců.

Definice 53 *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*

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} = S_i + u_i z_i^T, \qquad u_i = \frac{d_i - S_i y_i}{z_i^T y_i}$$

pro $l \le i \le l + m - 2$, kde $z_i \in R^n$ je zvolený vektor.

Poznámka 296 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 177 *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 $l \le j \le i$ (platí to zcela jistě pro $j = l$). Podle indukčního předpokladu a podle definice 52 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 178 *Nechť $p_l^{i+1} = S_l f_{i+1}$ a*

$$p_{j+1}^{i+1} = p_j^{i+1} + u_j z_j^T f_{i+1}, \qquad l \le j \le i,$$

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 $p_{j+1}^{i+1} = S_{j+1} f_{i+1}$, $l \le j \le i$, 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 $l \le j \le i$ (platí to zcela jistě pro $j = l$). Podle indukčního předpokladu a podle definice 52 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 297 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. Při odvozování maticových reprezentací budeme používat označení $D_k = [d_1, \ldots, d_k]$, $Y_k = [y_1, \ldots, y_k]$, $V_k = [v_1, \ldots, v_k]$, $Z_k = [z_1, \ldots, z_k]$. Abychom zjednodušili zápis budeme v důkazech index k vynechávat a index $k+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $D = [d_1, \ldots, d_{k-1}]$, $Y = [y_1, \ldots, y_{k-1}]$, $V = [v_1, \ldots, v_{k-1}]$, $Z = [z_1, \ldots, z_{k-1}]$, takže $D_k = [D, d]$, $Y_k = [Y, y]$, $V_k = [V, v]$, $Z_k = [Z, z]$.

Nejprve odvodíme maticové reprezentace pro přímé kvazinewtonovské metody s omezenou pamětí. Pro tento účel označíme R_k horní trojúhelníkovou matici řádu k takovou, že $(R_k)_{ij} = v_i^T d_j$, $i \le j$, a $(R_k)_{ij} = 0$, $i > j$. V důkazech budeme používat označenmí

$$R_k = \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}.$$

Lemma 61 *Nechť A_1 je regulární matice a nechť platí (449) s $v_k^T d_k \neq 0$ pro libovolný index $1 \le k \le m$. Pak lze psát*

$$A_{k+1} = A_1 + (Y_k - A_1 D_k) R_k^{-1} V_k^T. \tag{487}$$

Důkaz Pro $k = 1$ je (487) ekvivalentní s (449). Dále budeme postupovat matematickou indukcí. Předpokládejme, že (487) platí pro všechny indexy menší než k. Pro index k můžeme (487) 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 (449). □

Poznámka 298 Vyjádření (487) používáme nejčastěji ve spojení s iteračním řešením soustavy rovnic $A_i s_i + f_i = 0$, $i \in N$. Pokládáme $A_i = A_l = J_l$ pro $i = l \in M$ a

$$A_i = A_l + (Y_k - A_l D_k) R_k^{-1} V_k^T \tag{488}$$

pro $l < i \leq l + m - 1$ (viz (487)), 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 $A_i p$ počítáme podle vzorce $A_i p = A_l p + U_k R_k^{-1} V_k^T p$.

Věta 179 *Nechť S_1 je regulární matice a nechť platí (454) pro libovolný index $1 \leq k \leq m$. Pak lze psát*

$$S_{k+1} = S_1 + (D_k - S_1 Y_k)(C_k - L_k + V_k^T S_1 Y_k)^{-1} V_k^T S_1, \tag{489}$$

kde L_k je dolní trojúhelníková matice taková, že $(L_k)_{ij} = 0$, $i < j$, a $(L_k)_{ij} = v_i^T d_j$, $i \geq j$, a C_k je diagonální matice řádu k taková, že $(C_k)_{ij} = v_i^T d_j$, $i = j$, a $(C_k)_{ij} = 0$, $i \neq j$.

Důkaz Přímou inverzí vztahu (487) (použitím důsledku 6, kde $H = A_1$, $U = (Y_k - A_1 D_k) R_k^{-1}$ a $V = V_k$), dostaneme

$$A_{k+1}^{-1} = A_1^{-1} - A_1^{-1}(Y_k - A_1 D_k)(R_k + V_k^T A_1^{-1}(Y_k - A_1 D_k))^{-1} V_k^T A_1^{-1}.$$

Zřejmě $R_k - V_k^T D_k = C_k - L_k$ a jelikož $A_1^{-1} = S_1$ a $A_{k+1}^{-1} = S_{k+1}$, můžeme poslední vzorec přepsat ve tvaru

$$S_{k+1} = S_1 + (D_k - S_1 Y_k)(C_k - L_k + V_k^T S_1 Y_k)^{-1} V_k^T S_1.$$

□

Nyní odvodíme maticové reprezentace pro inverzní kvazinewtonovské metody s omezenou pamětí. Tentokrát označíme R_k horní trojúhelníkovou matici řádu k takovou, že $(R_k)_{ij} = z_i^T y_j$, $i \leq j$, a $(R_k)_{ij} = 0$, $i > j$. V důkazech budeme používat označenmí

$$R_k = \begin{bmatrix} R, & Z^T y \\ 0, & z^T y \end{bmatrix}.$$

Věta 180 *Nechť S_1 je regulární matice a nechť platí (455) s $z_k^T y_k \neq 0$ pro libovolný index $1 \leq k \leq m$. Pak lze psát*

$$S_{k+1} = S_1 + (D_k - S_1 Y_k) R_k^{-1} Z_k^T. \tag{490}$$

Důkaz Tvrzení věty plyne z duality. Porovnáme-li (449) s (455), vidíme, že v (487) stačí provést záměnu $A_1 \to S_1$, $V_k \to Z_k$, $D_k \to Y_k$ a $Y_k \to D_k$. □

Poznámka 299 Vyjádření (490) používáme k určení směrového vektoru $s_i = -S_i f_i$, $i \in N$. Pokládáme $S_i = S_l = J_l^{-1}$ pro $i = l \in M$ a

$$S_i = S_l + (D_k - S_l Y_k) R_k^{-1} Z_k^T \tag{491}$$

pro $l < i \leq l + m - 1$ (viz (490), 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 300 Snadno se dokáže tvrzení, které je analogií věty 135. 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 8.4 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 181 *Předpokládejme, že Jacobiova matice zobrazení f je tridiagonální matice tvaru*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \dots & 0 & 0 \\ \gamma_2 & \alpha_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & \dots & \gamma_n & \alpha_n \end{bmatrix}. \tag{492}$$

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*

$$\begin{aligned}
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_{i-1}}, & &\mathrm{mod}(i,3) = 1, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_{i-1}}, & &\mathrm{mod}(i,3) = 2, \\
\alpha_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_i}, & \beta_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_{i+1}}, \\
\gamma_i &= \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_{i-1}}, & &\mathrm{mod}(i,3) = 0,
\end{aligned}$$

kde veličiny, v jejichž vzorcích vystupují indexy $i < 1$ nebo $i > n$, nepočítáme.

Důkaz Použitím (492) 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 141, dostaneme dokazované tvrzení. □

K předpodmínění vybrané iterační metody realizující diferenční verzi Newtonovy metody lze též použít kvazinewtonovskou metodu s omezenou pamětí popsanou v poznámce 299. 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 (486)) pokládáme $S_i = S_l = T_l^{-1}$, kde T_l je tridiagonální matice (492) spočtená podle věty 181, a pro $l \le i \le l + m - 2$ používáme vzorec (491).

12.3 Diferenční verze Newtonovy metody pro řídké úlohy

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 56).

Sloupcové metody jsou založeny na aproximaci sloupců Je_j, $1 \le j \le n$, Jacobiovy matice J pomocí diferenčních vzorců

$$J(x)e_j \approx \frac{f(x + \delta_j e_j) - f(x)}{\delta_j},$$

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 $\{1, \ldots, n\} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_l$, takového, aby počet skupin l byl minimální je složitý kombinatorický problém, jehož řešení se vymyká rozsahu tohoto textu. Efektivní algoritmy pro řešení tohoto problému, včetně zdrojových programů v jazyce Fortran jsou uvedeny v časopisu TOMS. Poznamenejme, že sloupcové metody nelze použít, má-li Jacobiova matice husté řádky.

Řádkové metody určují jednotlivé nenulové prvky Jacobiovy matice podle vzorců

$$(J(x))_{ij} \approx \frac{f_i(x + \delta_j e_j) - f_i(x)}{\delta_j}.$$

Pro každý index $1 \leq i \leq n$, se počítají jen ty diference, které odpovídají nenulovým prvkům v i-tém řádku Jacobiovy matice. 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 $(n_1 + \ldots + n_n)/n$ 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 8.4 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 148 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

$$A_+ = A + \sum_{i=1}^{n} \frac{e_i^T(y - Ad)e_i(d^i)^T}{(d^i)^T d^i}, \tag{493}$$

kde členy s $d^i = 0$ odpadnou. Metoda, která používá aktualizaci (493) se nazývá Schubertovou metodou a jelikož je zobecněním Broydenovy dobré metody, má podobné vlastnosti jako Broydenova dobrá metoda. Není zaručena globální konvergence Schubertovy metody, takže je občas třeba iterační proces přerušovat a pokládat $A_+ = J_+$. Je však možné dokázat, že Schubertova metoda konverguje lokálně Q-superlineárně.

Lemma 62 *Nechť A_+ je matice určená podle (493). 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 63 *Nechť A_+ je matice určená podle (493) a nechť platí (J6). Pak*

$$\|A_+ - J_+\|_F \le \|A - J\|_F + \overline{G}\sqrt{n}\|d\|.$$

Důkaz Podle věty o střední hodnotě (tvrzení 5) platí

$$y = f_+ - f = \tilde{J}d, \qquad \tilde{J} = \int_0^1 J(x + \lambda d)\mathrm{d}\lambda.$$

Použijeme-li (J6), můžeme psát

$$\|\tilde{J} - J\|_F \le \sqrt{n}\int_0^1 \|J(x + \lambda d) - J(x)\|\mathrm{d}\lambda \le \overline{G}\sqrt{n}\|d\|\int_0^1 \lambda d\lambda \le \frac{1}{2}\overline{G}\sqrt{n}\|d\| \tag{494}$$

a podobným způsobem dostaneme

$$\|\tilde{J} - J_+\|_F \le \frac{1}{2}\overline{G}\sqrt{n}\|d\|. \tag{495}$$

Podle lemmatu 62 platí

$$\begin{aligned} \|A_+ - J_+\|_F &\le \|A_+ - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \le \|A - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \\ &\le \|A - J\|_F + \|\tilde{J} - J\|_F + \|\tilde{J} - J_+\|_F, \end{aligned}$$

což s použitím (494)–(495) dává dokazované tvrzení. □

Věta 182 *Nechť platí (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*

$$\begin{aligned} \|A_i d_i + f_i\| &\le \overline{\omega}\|f_i\|, \\ x_{i+1} &= x_i + d_i, \\ A_{i+1} &= \mathcal{P}_{QJ} A_i \end{aligned}$$

$\forall i \in N$, kde $0 \le \overline{\omega} < 1$, konverguje posloupnost x_i, $i \in N$, k bodu $x^* \in R^n$.

Důkaz (a) Výsledky dosažené v částech (a)–(b) důkazu věty 157 můžeme přeformulovat (pomocí okolí) tak, že existují čísla $\delta > 0$, $\vartheta > 0$ taková, že pokud $\|x - x^*\| \le \delta$, $\|(A - J(x))d\| \le \vartheta\|d\|$ a $\|Ad + f(x)\| \le \overline{\omega}\|f(x)\|$, kde $0 \le \overline{\omega} < 1$, platí

$$\frac{1 - \overline{\omega}}{\overline{J}}\|f(x)\| \le \|d\| \le \frac{1 + \overline{\omega}}{\underline{J}}\|f(x)\|, \tag{496}$$

kde $0 < \underline{J} < \|J^{-1}(x^*)\|^{-1} \le \|J(x^*)\| < \overline{J}$, a

$$\|f(x + d)\| \le r\|f(x)\|, \tag{497}$$

kde $\overline{\omega} < r < 1$. Zdůrazněme, že číslo $0 \le \overline{\omega} < 1$ 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

$$\left(1 + \frac{\overline{J}}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\right)\overline{\delta} \le \delta$$

a

$$\overline{\vartheta}\sqrt{n} + \overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\overline{\delta} \le \vartheta.$$

Nechť $\|x_1 - x^*\| \le \overline{\delta}$ a $\|A_1 - J_1\| \le \overline{\vartheta}$. Dokážeme indukcí, že pro libovolný index $i \in N$ platí $\|x_i - x^*\| \le \delta$ a $\|A_i - J_i\| \le \vartheta$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme platnost tohoto tvrzení pro $1 \le i \le k$. Pak podle (496), (497) a (J6) platí

$$\sum_{i=1}^{k}\|d_i\| \le \frac{1 + \overline{\omega}}{\underline{J}}\sum_{i=1}^{k}\|f_i\| \le \frac{1 + \overline{\omega}}{\underline{J}}\|f_1\|\sum_{i=1}^{k} r^{i-1} \le \frac{1}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\|f_1\| \le \frac{\overline{J}}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\|x_1 - x^*\|, \tag{498}$$

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 63, 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 183 *Nechť jsou splněny předpoklady věty 182 a nechť navíc $\|A_i d_i + f_i\|/\|f_i\| \to 0$. Pak $x_i \to x^*$ Q-superlineárně.*

Důkaz (a) Podle lemmatu 62 platí

$$\begin{aligned}\frac{\|y-Ad\|^2}{\|d\|^2} &\leq \|A-\tilde{J}\|_F^2 - \|A_+-\tilde{J}\|_F^2 \\ &= \left(\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F\right)\left(\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F\right) \\ &\leq \overline{M}\left(\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F\right).\end{aligned}$$

Existence konstanty $\overline{M}$ plyne z toho, že

$$\begin{aligned}\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F &\leq \|A-J\|_F + \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\| \\ &\leq 2\|A-J\|_F + 2\overline{G}\sqrt{n}\|d\| \\ &\leq 2\|A-J\|_F + 2\overline{G}\sqrt{n}\left(\|x^+-x^*\| + \|x-x^*\|\right),\end{aligned}$$

takže podle části (b) důkazu věty 182 platí

$$\|A-\tilde{J}\|_F + \|A_+-\tilde{J}\|_F \leq 2\sqrt{n}\vartheta + 4\overline{G}\sqrt{n}\delta \triangleq \overline{M}.$$

Dále lze psát

$$\|A_+-J_+\|_F \leq \|A_+-\tilde{J}\|_F + \|J_+-\tilde{J}\|_F,$$

takže podle lemmatu 63 platí

$$\begin{aligned}\|A-\tilde{J}\|_F - \|A_+-\tilde{J}\|_F &\leq \|A-J\|_F + \|J-\tilde{J}\|_F - \|A_+-J_+\|_F + \|J_+-\tilde{J}\|_F \leq \\ &\leq \|A-J\|_F - \|A_+-J_+\|_F + \overline{G}\sqrt{n}\|d\|,\end{aligned}$$

což podle (498) dává

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\|y_i-A_id_i\|^2}{\|d_i\|^2} &\leq \overline{M}\|A_1-J_1\|_F + \overline{M}\,\overline{G}\sqrt{n}\sum_{i=1}^{\infty}\|d_i\| \\ &\leq \overline{M}\|A_1-J_1|_F + \overline{M}\,\overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1-x^*\| < \infty,\end{aligned}$$

neboli

$$\lim_{i\to\infty}\frac{\|y_i-A_id_i\|}{\|d_i\|} = 0 \tag{499}$$

(b) Použijeme-li (499) a (494), dostaneme

$$\frac{\|(A_i-J_i)d_i\|}{\|d_i\|} \leq \frac{\|(A_i-\tilde{J}_i)d_i\|}{\|d_i\|} + \|\tilde{J}_i-J_i\| \leq \frac{\|y_i-A_id_i\|}{\|d_i\|} + \frac{1}{2}\overline{G}\sqrt{n}\|d_i\|,$$

což spolu s (499) a $\|d_i\| \to 0$ (podle části (c) důkazu věty 182) dává $\|(A_i-J_i)d_i\|/\|d_i\| \to 0$. Jelikož předpokládáme, že také $\|A_is_i+f_i\|/\|f_i\| \to 0$, lze použít větu 157, 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

$$\begin{aligned} \mathcal{V}_A &= \{A \in R^{n\times n} : A^T w = z\}, \\ \mathcal{V}_J &= \{A \in R^{n\times n} : J_{ij} = 0 \Rightarrow A_{ij} = 0\}, \end{aligned}$$

kde $z = J_+^T w$. Tak jako v oddílu 8.4 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 184 *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*

$$\mathcal{P}_{AJ} A = \mathcal{P}_J(A + wu^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = z - (\mathcal{P}_J A)^T w$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \|w^i\|^2 e_i e_i^T.$$

kde w^i, $1 \le i \le n$, jsou vektory takové, že $w^i_j = w_j$, $J_{ji} \neq 0$ a $w^i_j = 0$, $J_{ji} = 0$. 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 $A_+ = \mathcal{P}_J(A + wu^T)$ zapsat ve tvaru

$$A_+ = A - \sum_{i=1}^{n} \frac{w^T(A - J_+)e_i w^i e_i^T}{(w^i)^T w^i}, \tag{500}$$

kde členy s $w^i = 0$ odpadnou.

Důkaz Položíme-li ve větě 148 $B = A^T$, $G = J^T$, $d = w$, $y = z$, dostaneme

$$A_+^T = \mathcal{P}_{QG} B = \mathcal{P}_G(B + ud^T) = \mathcal{P}_J(A^T + uw^T),$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = z - (\mathcal{P}_J A)^T w = (J_+ - (\mathcal{P}_J A))^T w$. Má-i matice A stejné rozložení nenulových prvků jako matice J, můžeme tento vzorec zapsat ve tvaru

$$A_+^T = A^T + \sum_{i=1}^{n} \frac{e_i^T(J_+^T - A^T)w e_i (w^i)^T}{(w^i)^T w^i},$$

což po úpravě dává (500). □

Lemma 64 *Nechť A_+ je matice určená podle (500). Pak platí*

$$\|A_+ - J_+\|_F^2 \le \|A - J_+\|_F^2 - \frac{\|(A - J_+)^T w\|^2}{\|w\|^2}.$$

Důkaz Jelikož $J_+ \in \mathcal{V}_Q \cap \mathcal{V}_J$, $\mathcal{P}_{QJ}$ je operátor ortogonální projekce do $\mathcal{V}_Q \cap \mathcal{V}_J$ a $A_+ = \mathcal{P}_{QJ}A$, můžeme použít Pythagorovu větu

$$\|A_+ - J_+\|_F^2 = \|A - J_+\|_F^2 - \|A - A_+\|_F^2.$$

Jelikož $\mathcal{V}_Q \cap \mathcal{V}_J \subset \mathcal{V}_Q$, platí $A_+^T w = J_+^T w$, takže

$$\|(A - J_+)^T w\| = \|(A - A_+)w\| \leq \|A - A_+\|\|w\| \leq \|A - A_+\|_F\|w\|,$$

což po dosazení do předchozí rovnosti dává dokazované tvrzení. □

Lemma 65 *Nechť A_+ je matice určená podle (500) a nechť platí (J6). Pak*

$$\|A_+ - J_+\|_F \leq \|A - J\|_F + \overline{G}\sqrt{n}\|d\|.$$

Důkaz Použijeme-li (J6), dostaneme

$$\|J_+ - J\|_F \leq \sqrt{n}\|J_+ - J\| \leq \overline{G}\sqrt{n}\|d\|. \tag{501}$$

Podle lemmatu 64 pak platí

$$\begin{aligned} \|A_+ - J_+\|_F &\leq \|A - J_+\|_F \leq \|A - J\|_F + \|J_+ - J\|_F \\ &\leq \|A - J\|_F + \overline{G}\sqrt{n}\|d\|. \end{aligned}$$

□

Věta 185 *Nechť platí (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 Důkaz tohoto tvrzení je prakticky stejný jako důkaz věty 182 (místo matic $\tilde{J}_i$ $i \in N$, používáme matice J_{i+1}, $i \in N$, a místo lemmatu 63 používáme lemma 65). □

Věta 186 *Nechť jsou splněny předpoklady věty 185 a nechť navíc $\|A_i d_i + f_i\|/\|f_i\| \to 0$. Pak pokud vektory w_i, $i \in N$, vybíráme podle (AB1), (AB2) (nebo (AB3), když $\overline{\omega} = 0$), konverguje $x_i \to x^*$ Q-superlineárně.*

Důkaz (a) Podle lemmatu 64 platí

$$\begin{aligned} \frac{\|(A - J_+)^T w\|^2}{\|w\|^2} &\leq \|A - J_+\|_F^2 - \|A_+ - J_+\|_F^2 \\ &= (\|A - J_+\|_F - \|A_+ - J_+\|_F)(\|A - J_+\|_F + \|A_+ - J_+\|_F) \\ &\leq \overline{M}(\|A - J_+\|_F - \|A_+ - J_+\|_F). \end{aligned}$$

Existence konstanty $\overline{M}$ 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 (498) dává

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{\|(A_i - J_{i+1})^T w_i\|^2}{\|w_i\|^2} &\leq \overline{M}\|A_1 - J_1\|_F + \overline{M}\,\overline{G}\sqrt{n}\sum_{i=1}^{\infty}\|d_i\| \\ &\leq \overline{M}\|A_1 - J_1|_F + \overline{M}\,\overline{G}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1 - x^*\| < \infty,\end{aligned}$$

neboli

$$\lim_{i\to\infty} \frac{\|(A_i - J_{i+1})^T w_i\|}{\|w_i\|} = 0. \tag{502}$$

Dále platí

$$\frac{\|(A_i - J_i)^T w_i\|}{\|w_i\|} \leq \frac{\|(A_i - J_{i+1})^T w_i\|}{\|w_i\|} + \|J_{i+1} - J_i\| \leq \frac{\|(A_i - J_{i+1})^T w_i\|}{\|w_i\|} + \overline{G}\sqrt{n}\|d_i\|,$$

což spolu s (502) a $\|d_i\| \to 0$ dává

$$\lim_{i\to\infty} \frac{\|(A_i - J_i)^T w_i\|}{\|w_i\|} = 0. \tag{503}$$

(b) Ukážeme, že pro vektory w_i, $i \in N$, zvolené podle (AB1) nebo (AB2) platí

$$\lim_{i\to\infty} \frac{\|w_i - (A_i - J_i)d_i\|}{\|d_i\|} = 0. \tag{504}$$

Použijeme-li (AB1), dostaneme

$$\begin{aligned}\|w_i - (A_i - J_i)d_i\| &= \|(A_i - J_{i+1})d_i - (A_i - J_i)d_i\| = \|(J_i - J_{i+1})d_i\| \\ &\leq \|J_i - J_{i+1}\|\|d_i\| \leq \overline{G}\sqrt{n}\|d_i\|^2,\end{aligned}$$

což spolu s $\|d_i\| \to 0$ dává (504). Použijeme-li (AB2), dostaneme

$$\begin{aligned}\|w_i - (A_i - J_i)d_i\| &= \|(A_i d_i - (f_{i+1} - f_i) - (A_i - J_i)d_i\| = \|J_i d_i - (f_{i+1} - f_i)\| \\ &= \|f_i + J_i d_i - (f_i + J_i d_i + o(\|d_i\|))\| = o(\|d_i\|)\end{aligned}$$

(používáme dva členy Taylorova rozvoje), což spolu s $\|d_i\| \to 0$ dává (504).

(c) Ukážeme, že z (503) a (504) plyne

$$\lim_{i\to\infty} \frac{\|(A_i - J_i)d_i\|}{\|d_i\|} = 0. \tag{505}$$

Platí

$$\|w_i\|^2 = w_i^T(w_i - (A_i - J_i)d_i + (A_i - J_i)d_i) \le \|w_i\|\|w_i - (A_i - J_i)d_i\| + \|(A_i - J_i)^T w_i\|\|d_i\|,$$

takže

$$\frac{\|w_i\|}{\|d_i\|} \le \frac{\|w_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|(A_i - J_i)^T w_i\|}{\|w_i\|}$$

a

$$\frac{\|(A_i - J_i)d_i\|}{\|d_i\|} \le \frac{\|w_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|w_i\|}{\|d_i\|} \le 2\frac{\|w_i - (A_i - J_i)d_i\|}{\|d_i\|} + \frac{\|(A_i - J_i)^T w_i\|}{\|w_i\|},$$

což spolu s (503) a (504) dává (505). Z (505) a $\|A_i d_i + f_i\|/\|f_i\| \to 0$ plyne superlineární konvergence (věta 157). □

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 8.3) 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

$$U_+ d = L^{-1}Py.$$

Jelikož musí být zároveň zachována struktura horní trojúhelníkové matice, můžeme použít postup popsaný v oddílu 12.4. Výsledkem je aktualizace

$$U_+ = U + \sum_{i=1}^{n} \frac{e_i(L^{-1}Py - Ud)e_i d^i}{(d^i)^T d^i}, \tag{506}$$

kde d^i, $1 \le i \le n$, jsou vektory takové, že $d^i_j = d_j$, $U_{ij} \ne 0$ a $d^i_j = 0$, $U_{ij} = 0$ (členy s $d^i = 0$ odpadnou). Metoda, která používá aktualizaci (506) 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 (506). Po m aktualizacích (506) 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

$$D_+ LUd = Py.$$

Zapíšeme-li tuto podmínku ve tvaru

$$\sum_{i=1}^{n} D_+ e_i e_i^T LUd = Py$$

a přihlédneme-li k tomu, že matice D_+ je diagonální, můžeme psát

$$e_i^T D_+ e_i e_i^T LUd = e_i^T Py,$$

$1 \leq i \leq n$, neboli

$$e_i^T D_+ e_i = \frac{e_i^T Py}{e_i^T LUd}. \tag{507}$$

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 $\{1, \ldots, n\} = \mathcal{S}_1 \cup \ldots \cup \mathcal{S}_k$ 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 (493).

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 54 *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 187 *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 $\tilde{p}_i^T q_i \neq 0$ $\forall 1 \le i \le n$. 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 $\tilde{p}_i^T q_i \neq 0$ $\forall 1 \le i \le n$. Dokážeme indukcí, že platí

$$\tilde{p}_j^T f_i = p_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1, \tag{508}$$

$$\tilde{f}_j^T f_i = f_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1, \tag{509}$$

$$\tilde{p}_j^T q_i = p_j^T \tilde{q}_i = 0 \quad \forall 1 \le j < i \le n. \tag{510}$$

Z (509) 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 $\lambda_1 f_1 + \ldots + \lambda_n f_n = 0$, 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 (510) plyne, že vektory p_i, $1 \le i \le n$ (a také $\tilde{p}_i$, $1 \le i \le n$), jsou lineárně nezávislé. Jelikož $f_{n+1} = As_{n+1} + f$ (plyne to z rekurentních vztahů metody BCG), vektory $\tilde{f}_i$, $1 \le i \le n$, jsou lineárně nezávislé a

$$\tilde{f}_j^T f_{n+1} = 0 \quad \forall 1 \le j \le n,$$

musí platit $f_{n+1} = As_{n+1} + f = 0$. Pro $i = 1$ (508)–(510) platí, neboť není co dokazovat.

(a) Nechť $i \leq n$. Podle indukčních předpokladů (508) a (510) 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 (508) a (510) 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 301 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 66 *Nechť jsou splněny předpoklady věty 187. 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*

$$\mathcal{K}_i = \text{span}\{f,\ Af, \ldots, A^{i-1}f\}$$

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

$$\tilde{\mathcal{K}}_i = \text{span}\{\tilde{f},\ (A^T)\tilde{f}, \ldots, (A^T)^{i-1}\tilde{f}\}.$$

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 187), 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 187), tvoří tam bázi. □

Poznámka 302 Nechť jsou splněny předpoklady věty 187. 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

$$\begin{aligned} \varphi_{i+1} &= \varphi_i - \alpha_i A\psi_i, \\ \psi_{i+1} &= \varphi_{i+1} + \beta_i \psi_i \end{aligned}$$

$1 \leq i \leq n$. Plyne to bezprostředně z rekurentních vztahů metody BCG. Koeficienty α_i a β_i, $1 \leq i \leq n$, lze vyjádřit pomocí polynomů φ_i a ψ_i, $1 \leq i \leq n$, 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, $1 \leq i \leq n$ lze použít také k určení polynomů $\varphi_i^2(A)$ a $\psi_i^2(A)$, $1 \leq i \leq n$, můžeme definovat nový iterační proces $\bar{s}_i \in R^n$, $1 \leq i \leq n+1$ tak, aby platilo $\bar{f}_i = A\bar{s}_i + f = \varphi_i^2(A)f$, $1 \leq i \leq n+1$.

Lemma 67 *Nechť maticové polynomy φ_i a ψ_i splňují rekurentní vztahy*

$$\varphi_1 = I, \qquad \psi_1 = I$$

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 55 *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

$$\begin{gathered}v_i = Ap_i, \quad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i,\\ q_i = u_i - \alpha_i v_i,\\ \overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i),\\ \overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i,\\ u_{i+1} = \overline{f}_{i+1} + \beta_i q_i,\\ p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)\end{gathered}$$

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 303 Jsou-li splněny předpoklady věty 187 platí

$$\|\overline{f}_i\| = \|\varphi_i^2(A)f\| \leq \|\varphi_i(A)\|\|\varphi_i(A)f\| = \|\varphi_i(A)\|\|f_i\|,$$

$1 \leq i \leq n+1$, takže metoda CGS najde řešení soustavy rovnic $As+f=0$ po nejvýše n krocích ($\|f_{n+1}\| = 0$ podle věty 187).

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 68 *Nechť $\overline{f}_i$, $i \in N$, je posloupnost reziduí určená metodou CGS. Nechť $f_1 = \overline{f}_1$ a*

$$\begin{aligned} \lambda_i &= -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2}, \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}), \end{aligned}$$

$1 \leq i \leq n$. Pak platí

$$\lambda_i = \arg\min_{\lambda \in R} \|\overline{f}_{i+1} + \lambda(f_i - \overline{f}_{i+1})\|,$$

$1 \leq i \leq n$, takže $\|f_{i+1}\| \leq \|f_i\|$ (normy reziduí monotonně klesají) a $\|f_{i+1}\| \leq \|\overline{f}_{i+1}\|$ (řešení je nalezeno po nejvýše n krocích).

Důkaz Zřejmě pro $1 \leq i \leq n$ 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 68) spolu s odpovídajícími rekurentními vztahy pro s_i jsou základem jednoduše zhlazené metody CGS.

Definice 56 *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

$$\begin{aligned} v_i &= Ap_i, \quad \alpha_i = \tilde{f}^T f_i / \tilde{f}^T v_i, \\ q_i &= u_i - \alpha_i v_i, \\ \overline{s}_{i+1} &= \overline{s}_i - \alpha_i(u_i + q_i), \\ \overline{f}_{i+1} &= \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T f_{i+1} / \tilde{f}^T \overline{f}_i, \\ u_{i+1} &= \overline{f}_{i+1} + \beta_i q_i, \\ p_{i+1} &= u_{i+1} + \beta_i(q_i + \beta_i p_i), \\ \lambda_i &= -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2}, \\ s_{i+1} &= \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}), \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) \end{aligned}$$

pro $1 \leq i \leq n$, 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 57 *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 304 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 68). 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 ortogonálními vektory q_i, $1 \leq i \leq n$, tak, že q_j $1 \leq j \leq i$, tvoří bázi v Krylovově podprostoru $\mathcal{K}_i$. Vektor $s_{i+1} \in R^n$ se volí tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \|As + f\|. \tag{511}$$

Metoda GMRES je tedy založena na minimalizačním principu, což znamená, že normy reziduí monotonně klesají.

Ortonormální vektory q_i, $1 \leq i \leq n$ se generují pomocí Gramova-Schmidtova ortogonalizačního procesu. Klasický Gramův-Schmidtův ortogonalizační proces používá rekurentní vztahy

$$\beta_1 q_1 = f$$

a

$$\begin{gathered} q_{i+1}^1 = Aq_i, \\ \left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^1, \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j, \end{aligned}\right\} 1 \leq j \leq i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \end{gathered}$$

$1 \leq i \leq n-1$, kde koeficienty β_i, $1 \leq i \leq n$ se vybírají tak, aby vektory q_i, $1 \leq i \leq n$, měly jednotkovou normu. Stabilnější je modifikovaný Gramův-Schmidtův ortogonalizační proces

$$\beta_1 q_1 = f$$

a

$$\begin{gathered} q_{i+1}^1 = Aq_i, \\ \left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j, \end{aligned}\right\} 1 \leq j \leq i \\ \beta_{i+1} q_{i+1} = q_{i+1}^{i+1}, \end{gathered}$$

$1 \leq i \leq n-1$. Gramův-Schmidtův ortogonalizační proces generující ortonormální báze Krylovových podprostorů $\mathcal{K}_i$, $1 \leq i \leq n$, 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

$$z_i = \arg\min_{z \in R^i} \|H_i z + \beta_1 e_1\|. \tag{512}$$

Věta 188 *Nechť $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j \leq i$, $\mathcal{K}_i = \mathcal{K}_{i+1}$ a nechť platí (511). Pak $As_{i+1} + f = 0$.*

Důkaz Uvažujme Arnoldiův proces určený regulární maticí $A \in R^{n\times n}$ a vektorem f. Jestliže $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j \leq i$ a $\mathcal{K}_i = \mathcal{K}_{i+1}$, pak vektory q_i, $1 \leq j \leq i$, 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

$$\|As_{i+1} + f\| = \|\overline{H}_i z_i + \beta_1 e_1\| = 0.$$

□

Důsledek 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 ($\overline{\mathrm{M}}$), 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 **??**), proto ho nebudeme znovu odvozovat. Uvedeme pouze výsledné rekurentní vztahy metody GMRES.

Definice 58 *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

$$\begin{gathered} q_{i+1}^1 = Aq_i, \\ \overline{\alpha}_{1i} = q_1^T q_{i+1}^1, \qquad q_{i+1}^2 = q_{i+1}^1 - \overline{\alpha}_{1i} q_1, \\ \left.\begin{gathered} \alpha_{ji} = q_j^T q_{i+1}^j, \qquad q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \\ \rho_{j-1i} = \lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji}, \\ \overline{\alpha}_{ji} = -\lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji}, \end{gathered}\right\} 1 < j \leq i \end{gathered}$$

$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$
$$\rho_{ii} = \sqrt{\overline{\alpha}_{ii}^2 + \beta_{i+1}^2},$$
$$\lambda_i = \frac{\overline{\alpha}_{ii}}{\rho_{ii}}, \qquad \tau_i = \frac{\beta_{i+1}}{\rho_{ii}},$$
$$\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i,$$

$1 \leq i \leq n$, *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 $|(\overline{\eta}_{i+1}| \leq \overline{\omega}\|f\|)$ 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

$$A\mathcal{K}_i = \text{span}\{Af, \ldots, A^i f\},$$

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

$$R_i = \begin{bmatrix} \beta_1 & \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1i-1} \\ 0 & \beta_2 & \alpha_{22} & \ldots & \alpha_{2i-1} \\ 0 & 0 & \beta_3 & \ldots & \alpha_{3i-1} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & \beta_i \end{bmatrix}$$

($R_i \in R^{i \times i}$ 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 59 *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$. Pak iterační proces*

$$f_1 = f, \quad \beta_1 q_1 = Af$$

a

$$\gamma_i = q_i^T f_i,$$
$$f_{i+1} = f_i - \gamma_i q_i,$$
$$q_{i+1}^1 = Aq_i,$$
$$\left.\begin{array}{l} \alpha_{ji} = q_j^T q_{i+1}^j, \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j, \end{array}\right\} 1 \le j \le i,$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1},$$

$1 \le i \le n-1$, 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 ($\|f_{i+1}\| \le \overline{\omega}\|f\|$), 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 305 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 $C \approx A$.

12.9 Metody s lokálně omezeným krokem

Poznámka 306 Zhlazenou metodu CGS nebo metodu GMRES můžeme použít ke konstrukci nepřesných metod s lokálně omezeným krokem. V tomto případě se generuje posloupnost vektorů $s_{i+1} \in R^n$, $1 \leq i \leq n$, které aproximují řešení soustavy rovnic $As + f = 0$, a pak se pokládá $s = s_{i+1}$, pokud $\|s_{i+1}\| \leq \Delta$ a $\|f_{i+1}\| \leq \overline{\omega}\|f\|$, $0 \leq \overline{\omega} < 1$, nebo $s = s_i + \alpha_i(s_{i+1} - s_i)$ a $\|s\| = \Delta$, pokud $\|s_j\| < \Delta$, $\|f_j\| > \overline{\omega}\|f\|$, $1 \leq j \leq i$, a $\|s_{i+1}\| \geq \Delta$. Tato volba zřejmě splňuje podmínky (T1a), (T1b) metody s lokálně omezeným krokem (definice 47). Navíc je třeba zformulovat předpoklady, aby byla splněna i podmínka (T1c), neboli

$$\|f\| - \|As + f\| \geq 2\underline{\sigma}\|As\|),$$

kde $\underline{\sigma}$ je nějaká konstanta. V dalším textu budeme předpokládat, že matice A splňuje podmínku $\|I - A\| \leq \overline{\nu} < 1$, což lze docílit vhodným předpodmíněním (místo matice A se používá matice AC^{-1} taková, že $\|I - AC^{-1}\| \leq \overline{\nu} < 1$).

Lemma 69 *Nechť $\|I - A\| \leq \overline{\nu} < 1$ a nechť $s_{i+1} \in R^n$, $i = 1, \ldots, n$, jsou vektory generované metodou GMRES nebo dvojnásobně zhlazenou metodou CGS. Pak*

$$\|f\|^2 - \|r_{i+1}\|^2 \geq \underline{\eta}^2\|f\|^2,$$

kde $\underline{\eta} = (1 - \overline{\nu})/(1 + \overline{\nu})$.

Důkaz (a) Nejprve ukážeme, že

$$|f^T Af| \geq \frac{1 - \overline{\nu}}{1 + \overline{\nu}}\|f\|\|Af\| = \underline{\eta}\|f\|\|Af\|.$$

Podle předpokladu platí

$$\begin{aligned} |f^T Af| &= |f^T f - f^T(I - A)f| \geq |f^T f| - |f^T(I - A)f| \\ &\geq \|f\|^2 - \|I - A\|\|f\|^2 \geq (1 - \overline{\nu})\|f\|^2 \end{aligned}$$

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 70 *Nechť jsou splněny předpoklady lemmatu 69 a nechť $s \in \mathcal{R}^n$ je vektor určený metodou GMRES nebo dvojnásobně zhlazenou metodou CGS tak jako v poznámce 306. 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 69 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 69 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 189 *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 306. 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 55, platí $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 70 a věty 159. □

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 70 nebo v důkazu lemmatu **??**), že pokud platí $\Delta \geq \underline{\gamma}\|f\|$ nebo $|f^T Af| \geq \underline{\varepsilon}\|f\|\|Af\|$, je splněna podmínka ($\overline{\text{T1c}}$).

V následující tabulka ukazuje porovnání několika metod pro řešení rozsáhlých systémů nelineárních rovnic: DNM - diskrétní Newtonova metoda, QNS - quasinewtonovská metoda se Schubertovou řídkou aktualizací, 5-QLB - pětikroková Broydenova metoda s omezenou pamětí, 5-QLC - pětikroková metoda aktualizace sloupců s omezenou pamětí, 5-QLI - pětikroková inverzní metoda aktualizace sloupců s omezenou pamětí, DNS - metoda škálování řádků LU rozkladu, DNL - Lieova nedokonalá Newtonova metoda se Schubertovou aktualizací. Tyto metody jsou realizovány buď s iteračním (CGS - dvojnásobně zhlazená metoda CGC předpodmíněná pomocí neúplného LU rozkladu) nebo s přímým (LU - úplný LU rozklad) řešením soustav lineárních rovnic. K porovnání bylo použito 44 rozsáhlých řídkých systémů nelineárních rovnic s 5000 neznámými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i počet selhání a celkový čas výpočtu).

Metoda	NIT-NFV	selhání	čas
DNM + CGS	576-2988	-	7.94
DNM + LU	570-2942	1	8.67
QNS + CGS	876-2114	-	9.09
QNS + LU	842-2212	1	8.39
5-QLB + CGS	756-2012	1	7.19
5-QLC + CGS	770-2255	-	7.39
5-QLI + CGS	972-2328	-	6.47
5-QLI + LU	747-1667	1	4.91
DNS + LU	848-3653	1	8.17
DNL + LU	672-6028	1	16.97

13 Optimalizace dynamických systémů

Uvažujeme úlohu s účelovou funkcí

$$F(x) = \int_{t_0}^{t_1} F_A(x, y(x,t), t)\,\mathrm{d}t + F_T(x, y(x,t_1)), \tag{513}$$

kde

$$\frac{\mathrm{d}y(x,t)}{\mathrm{d}t} = f_S(x, y(x,t), t), \qquad y(x,t_0) = f_I(x)\ . \tag{514}$$

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 počátečmími podmínkami. Tuto soustavu diferenciálních rovnic nazýváme stavovým systémem.

Obvykle je výhodné počítat integrál společně s řešením soustavy diferenciálních rovnic. V tomto případě pokládáme

$$F(x) = \tilde{F}_A(x, t_1) + F_T(x, y(x, t_1)), \tag{515}$$

kde

$$\begin{aligned} \frac{\mathrm{d}y(x,t)}{\mathrm{d}t} &= f_S(x, y(x,t), t), \qquad y(x, t_0) = f_I(x), \\ \frac{\mathrm{d}\tilde{F}_A(x,t)}{\mathrm{d}t} &= F_A(x, y(x,t), t), \qquad \tilde{F}_A(x, t_0) = 0. \end{aligned} \tag{516}$$

Celkem se řeší $n_S + 1$ diferenciálních rovnic v přímém směru. Stačí spočítat hodnoty na konci intervalu. Úloha (515) + (516) se řeší pomocí gradientních optimalizačních metod (CG, VM, N) proto je třeba počítat derivace účelové funkce. Předpoklady:

(A1) Existuje spojité řešení systému (514) na intervalu $[t_0, t_1]$ kdykoliv $x \in X \subset R^n$.

(A2) Funkce F_A, F_T, f_S, f_I jsou dvakrát spojitě diferencovatelné na $X \subset R^n$.

Přitom $X \subset R^n$ je oblast obsahující všechny body $x_i \in R^n$ $i \in N$, získané během iteračního procesu.

13.1 Přímý výpočet gradientu

Označme $u(x,t) = \mathrm{d}y(x,t)/\mathrm{d}x$, takže $u : R^n \times [t_0, t_1] \to R^{n_S \times n}$. Derivováním (515) a (516) dostaneme

$$g^T(x) = \tilde{g}_A^T(x, t_1) + \frac{\partial F_T(x, y(x,t_1))}{\partial x} + \frac{\partial F_T(x, y(x,t_1))}{\partial y} u(x, t_1), \tag{517}$$

kde

$$\begin{aligned} \frac{\mathrm{d}u(x,t)}{\mathrm{d}t} &= \frac{\partial f_S(x,y,t)}{\partial x} + \frac{\partial f_S(x,y,t)}{\partial y} u(x,t), \qquad u(x,t_0) = \frac{\mathrm{d}f_I(x)}{\mathrm{d}x}, \\ \frac{\mathrm{d}\tilde{g}_A^T(x,t)}{\mathrm{d}t} &= \frac{\partial F_A(x,y,t)}{\partial x} + \frac{\partial F_A(x,y,t)}{\partial y} u(x,t), \qquad \tilde{g}_A^T(x,t_0) = 0. \end{aligned} \tag{518}$$

Přitom $g^T(x) = \mathrm{d}F(x)/\mathrm{d}x$, $\tilde{g}_A^T(x,t) = \mathrm{d}\tilde{F}_A(x,t)/\mathrm{d}x$. Celkem se řeší $(n_S + 1)(n + 1)$ diferenciálních rovnic v přímém směru.

13.2 Zpětný výpočet gradientů

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 (514), takže $f_S(x,y,t) - \mathrm{d}y(x,t)/\mathrm{d}t = 0$ pro $t \in [t_0, t_1]$. Použijeme-li (513), 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

$$\begin{aligned} F(x) &= \int_{t_0}^{t_1} \left[F_A(x,y,t) + p^T(t) f_S(x,y,t) + \frac{\mathrm{d}p^T(t)}{\mathrm{d}t} y(x,t) \right] \mathrm{d}t \\ &\quad + p^T(t_0) y(x,t_0) - p^T(t_1) y(x,t_1) + F_T(x, y(x,t_1)). \end{aligned}$$

Nyní můžeme $F(x)$ derivovat podle x, takže

$$\begin{aligned} g^T(x) &= \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial x} \right. \\ &\quad \left. + \left(\frac{\partial F_A(x,y,t)}{\partial y} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial y} + \frac{\mathrm{d}p^T(t)}{\mathrm{d}t} \right) \frac{\mathrm{d}y(x,t)}{\mathrm{d}x} \right] \mathrm{d}t \\ &\quad + p^T(t_0) \frac{\mathrm{d}f_I(x)}{\mathrm{d}x} + \frac{\partial F_T(x,y(x,t_1))}{\partial x} + \left(\frac{\partial F_T(x,y(x,t_1))}{\partial y} - p^T(t_1) \right) \frac{\mathrm{d}y(x,t_1)}{\mathrm{d}x}. \end{aligned}$$

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

$$-\frac{\mathrm{d}p^T(x,t)}{\mathrm{d}t} = \frac{\partial F_A(x,y,t)}{\partial y} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial y}, \qquad p^T(x,t_1) = \frac{\partial F_T(x,y(x,t_1))}{\partial y},$$

platí

$$g^T(x) = \int_{t_0}^{t_1} \left[\frac{\partial F_A(x,y,t)}{\partial x} + p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial x} \right] \mathrm{d}t + p^T(x,t_0) \frac{\mathrm{d}f_I(x)}{\mathrm{d}x} + \frac{\partial F_T(x,y(x,t_1))}{\partial x}.$$

Dohromady to lze zapsat takto

$$g(x) = \tilde{g}_A(x,t_0) + \left(\frac{\mathrm{d}f_I(x)}{\mathrm{d}x} \right)^T p(x,t_0) + \left(\frac{\partial F_T(x,y(x,t_1))}{\partial x} \right)^T, \tag{519}$$

kde

$$\begin{aligned} -\frac{\mathrm{d}p(x,t)}{\mathrm{d}t} &= \left(\frac{\partial F_A(x,y,t)}{\partial y} \right)^T + \left(\frac{\partial f_S(x,y,t)}{\partial y} \right)^T p(x,t), \qquad p(x,t_1) = \left(\frac{\partial F_T(x,y(x,t_1))}{\partial y} \right)^T, \\ \frac{\mathrm{d}\tilde{g}_A(x,t)}{\mathrm{d}t} &= \left(\frac{\partial F_A(x,y,t)}{\partial x} \right)^T + \left(\frac{\partial f_S(x,y,t)}{\partial x} \right)^T p(x,t), \qquad \tilde{g}_A(x,t_1) = 0. \end{aligned} \tag{520}$$

Celkem se řeší $n_S + 1$ diferenciálních rovnic v přímém směru a $n_S + n$ diferenciálních rovnic ve zpětném směru.

13.3 Přímý výpočet Hessovy matice

Označme $v(x,t) = \mathrm{d}u(x,t)/\mathrm{d}x = \mathrm{d}^2y(x,t)/\mathrm{d}x^2$, takže $v : R^n \times [t_0,t_1] \to R^{n_S \times n \times n}$. Derivováním (517) a (518) dostaneme

$$\begin{aligned} G(x) &= \tilde{G}_A(x,t_1) + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x^2} \\ &+ \left[2 \frac{\partial^2 F_T(x,y(x,t_1))}{\partial x \partial y} + \frac{\partial^2 F_T(x,y(x,t_1))}{\partial y^2} u(x,t_1) \right] u(x,t_1) \\ &+ \frac{\partial F_T(x,y(x,t_1))}{\partial y} v(x,t_1), \end{aligned}$$

kde

$$\begin{aligned}
\frac{\mathrm{d}v(x,t)}{\mathrm{d}t} &= \frac{\partial^2 f_S(x,y,t)}{\partial x^2} + \left[2\,\frac{\partial^2 f_S(x,y,t)}{\partial x \partial y} + \frac{\partial^2 f_S(x,y,t)}{\partial y^2}u(x,t)\right]u(x,t) \\
&+ \frac{\partial f_S(x,y,t)}{\partial y}v(x,t), \quad v(x,t_0) = \frac{\mathrm{d}^2 f_I(x)}{\mathrm{d}x^2}, \\
\frac{\mathrm{d}\tilde{G}_A(x,t)}{\mathrm{d}t} &= \frac{\partial^2 F_A(x,y,t)}{\partial x^2} + \left[2\,\frac{\partial^2 F_A(x,y,t)}{\partial x \partial y} + \frac{\partial^2 F_A(x,y,t)}{\partial y^2}u(x,t)\right]u(x,t) \\
&+ \frac{\partial F_A(x,y,t)}{\partial y}v(x,t), \quad \tilde{G}_A(x,t_0) = 0.
\end{aligned} \tag{521}$$

Přitom $G(x) = \mathrm{d}^2F(x)/\mathrm{d}x^2$ a $G_A(x) = \mathrm{d}^2 f_A(x,t)/\mathrm{d}x^2$. Celkem se řeší $(n_S+1)(n^2+n+1)$ diferenciálních rovnic v přímém směru.

13.4 Přímá aproximace Hessovy matice (součet čtverců)

Platí

$$F_A(y,t) = \frac{1}{2}(y(x,t) - z(t))^T W(t)(y(x,t) - z(t)),$$

$$\frac{\partial F_A(y,t)}{\partial y} = W(t)(y(x,t) - z(t)), \qquad \frac{\partial^2 F_A(y,t)}{\partial y^2} = W(t)$$

a podobně

$$F_T(y(x,t_1)) = \frac{1}{2}(y(x,t_1) - z(t_1))^T W_1(y(x,t_1) - z(t_1)),$$

$$\frac{\partial F_T(y(x,t_1))}{\partial y} = W_1(y(x,t_1) - z(t_1)), \qquad \frac{\partial^2 F_T(y(x,t_1))}{\partial y^2} = W_1.$$

Přitom $z : [t_0,t_1] \to R^{n_S}$, $W : [t_0,t_1] \to R^{n_S \times n_S}$ (SPD) (obecně $W_1 \neq W(t_1)$). Jestliže $F(x) \to 0$, pak nutně $y(x,t) \to z(t)$ takže $\partial F_A(y(x,t),t)/\partial y \to 0$ a $\partial F_T(y(x,t_1))/\partial y \to 0$. Můžeme tedy zanedbat tyto členy v (521) a (521). Dostaneme tak

$$G(x) \approx B(x) = B_A(x,t_1) + u^T(x,t_1)W_1u(x,t_1), \tag{522}$$

kde

$$\frac{\mathrm{d}B_A(x,t)}{\mathrm{d}t} = u^T(x,t)W(t)u(x,t), \qquad B_A(x,t_0) = 0. \tag{523}$$

Celkem se řeší $(n_S+1)(n+1)+n^2$ diferenciálních rovnic v přímém směru.

14 Automatické derivování

Existují dva postupy:

(1) Přímé automatické derivování – výpočet derivace zobrazení.

(2) Zpětné automatické derivování – výpočet gradientu funkce.

Ukážeme na příkladech použití obou postupů.

Příklad 1 (Přímé derivování). Máme nalézt derivaci funkce $F(x) = x_1 \sin(x_2 x_3 + x_4)$ podle proměnné x_3.

x_1	$=$	x_1		$F = x_1 \sin(x_2 x_3 + x_4)$
x_1'	$=$	0		
x_2	$=$	x_2		
x_2'	$=$	0		
x_3	$=$	x_3		
x_3'	$=$	1		
x_4	$=$	x_4		
x_4'	$=$	0		
x_5	$=$	$x_2 x_3$	$=$	$x_2 x_3$
x_5'	$=$	$x_2 x_3' + x_2' x_3$	$=$	x_2
x_6	$=$	$x_4 + x_5$	$=$	$x_4 + x_2 x_3$
x_6'	$=$	$x_4' + x_5'$	$=$	x_2
x_7	$=$	$\sin(x_6)$	$=$	$\sin(x_4 + x_2 x_3)$
x_7'	$=$	$\cos(x_6) x_6'$	$=$	$\cos(x_4 + x_2 x_3) x_2$
x_8	$=$	$x_1 x_7$	$=$	$x_1 \sin(x_4 + x_2 x_3)$
x_8'	$=$	$x_1 x_7' + x_1' x_7$	$=$	$x_1 x_2 \sin(x_4 + x_2 x_3) x_2$
F	$=$	x_8	$=$	$x_1 \sin(x_4 + x_2 x_3)$
F'	$=$	x_8'	$=$	$x_1 \cos(x_4 + x_2 x_3) x_2$

Příklad 2 (Zpětné derivování). Máme nalézt gradient funkce $F(x) = x_1 \sin(x_2 x_3 + x_4)$.

x_1	$=$	x_1		$\overline{x}_i = \partial F / \partial x_i$
x_2	$=$	x_2		
x_3	$=$	x_3		
x_4	$=$	x_4		
x_5	$=$	$x_2 x_3$	$=$	$x_2 x_3$
x_6	$=$	$x_4 + x_5$	$=$	$x_4 + x_2 x_3$
x_7	$=$	$\sin(x_6)$	$=$	$\sin(x_4 + x_2 x_3)$
x_8	$=$	$x_1 x_7$	$=$	$x_1 \sin(x_4 + x_2 x_3)$
$\overline{x}_8$	$=$	1		
$\overline{x}_7$	$=$	$\overline{x}_8 x_1$	$=$	x_1
$\overline{x}_1$	$=$	$\overline{x}_8 x_7$	$=$	$\sin(x_4 + x_2 x_3)$
$\overline{x}_6$	$=$	$\overline{x}_7 \cos x_6$	$=$	$x_1 \cos(x_4 + x_2 x_3)$
$\overline{x}_5$	$=$	$\overline{x}_6$	$=$	$x_1 \cos(x_4 + x_2 x_3)$
$\overline{x}_4$	$=$	$\overline{x}_6$	$=$	$x_1 \cos(x_4 + x_2 x_3)$
$\overline{x}_3$	$=$	$\overline{x}_5 x_2$	$=$	$x_1 \cos(x_4 + x_2 x_3) x_2$
$\overline{x}_2$	$=$	$\overline{x}_5 x_3$	$=$	$x_1 \cos(x_4 + x_2 x_3) x_3$
$\partial F / \partial x_1$	$=$	$\overline{x}_1$	$=$	$\sin(x_4 + x_2 x_3)$
$\partial F / \partial x_2$	$=$	$\overline{x}_2$	$=$	$x_1 \cos(x_4 + x_2 x_3) x_3$
$\partial F / \partial x_3$	$=$	$\overline{x}_3$	$=$	$x_1 \cos(x_4 + x_2 x_3) x_2$
$\partial F / \partial x_4$	$=$	$\overline{x}_4$	$=$	$x_1 \cos(x_4 + x_2 x_3)$

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.

Tvrzení 10 *(Hahn-Banach). Nechť funkce $F : R^n \to R$ je positivně 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 funkce taková, že $l(x) \leq F(x)\ \forall x \in X$. Pak existuje vektor $g \in R^n$ tak, že $g^T x = l(x)\ \forall x \in X$ a $g^T x \leq F(x)\ \forall x \in R^n$.*

Tvrzení 11 *(Rademacher). Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská (definice 79) v oblasti $\Omega \in R^n$. Pak F je diferencovatelná skoro všude (množina $\{x \in R_n : \nabla F(x)$ neexistuje$\}$ má Lebesgueovu míru nula).*

15.1 Konvexní množiny

Definice 60 *Řekněme, že množina $C \in R^n$ je konvexní, jestliže z $x \in C$, $y \in C$ plyne*

$$\lambda x + (1 - \lambda)y \in C, \tag{524}$$

pokud $0 \leq \lambda \leq 1$.

Poznámka 307 Vztah (524) můžeme zapsat ve tvaru

$$y + \lambda(x - y) \in C.$$

Definice 61 *Nechť $m \geq 1$, $x_i \in R^n$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \ldots + \lambda_m = 1$. 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 190 *Množina $C \subset R^n$ je konvexní právě tehdy, obsahuje-li všechny konvexní kombinace svých bodů.*

Důkaz Obsahuje-li množina C všechny konvexní kombinace svých bodů, obsahuje též konvexní kombinace tvaru (524), takže je konvexní. Opačnou implikaci dokážeme indukcí. Předpokládejme, že C obsahuje všechny konvexní kombinace svých m bodů, kde $m \geq 1$ (pro $m = 1$ je to zřejmé, neboť z $x_1 \in C$ a $\lambda_1 = 1$ plyne $\lambda_1 x_1 = x_1 \in C$). Pak pro $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m + 1$, $\lambda_1 + \ldots + \lambda_{m+1} = 1$ můžeme psát

$$\sum_{i=1}^{m+1} \lambda_i x_i = \sum_{i=1}^{m} \lambda_i x_i + \lambda_{m+1} x_{m+1} = (1 - \lambda_{m+1}) x'_{m+1} + \lambda_{m+1} x_{m+1} \in C,$$

kde

$$x'_{m+1} = \sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} x_i \triangleq \sum_{i=1}^{m} \lambda'_i x_i \in C,$$

neboť $x_i \in C$, $\lambda'_i \geq 0$, $1 \leq i \leq m$, $\lambda'_1 + \ldots + \lambda'_m = 1$. □

Poznámka 308 Podobným způsobem, jaký jsme použili v důkazu věty 190, lze ukázat, že konvexní kombinace konvexních kombinací je opět konvexní kombinací.

Poznámka 309 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 a konvexní množiny. Lineárními podprostory a afinními množinami se zde zabývat nebudeme (probírají se v kurzech lineární algebry). Připomeneme pouze, že všechny uvedené množiny jsou konvexní a afinní množina je posunutým lineárním podprostorem, to znamená, že je-li množina C afinní a $x \in C$, je množina $C - x$ lineárním podprostorem. Lze tedy definovat dimenzi afinní množiny jako dimenzi odpovídajícího lineárního podprostoru a jelikož konvexní množinu lze vnořit do afinní množiny (vynecháním omezení $\lambda_i \geq 0$, $1 \leq i \leq m$) i dimenzi konvexní množiny. Z těchto úvah plyne že konvexní množina v R^n obsahuje vnitřní body právě tehdy, má-li dimenzi n.

Věta 191 *Průnik konvexních množin je konvexní množinou.*

Důkaz Nechť $C = \cap_\alpha C_\alpha$, kde $C_\alpha \subset R^n$ jsou konvexní množiny. Nechť $x \in C$, $y \in C$. Pak platí $x \in C_\alpha$ a $y \in C_\alpha$ $\forall \alpha$ a tedy $\lambda x + (1-\lambda)y \in C_\alpha$ $\forall \alpha$ pokud $0 \leq \lambda \leq 1$. Odtud plyne, že $\lambda x + (1-\lambda)y \in C$. □

Věta 192 *Lineární kombinace konvexních množin je konvexní množinou.*

Důkaz Nechť $C = \sum_{i=1}^m \lambda_i C_i$, kde $C_i \subset R^n$ jsou konvexní množiny a $\lambda_i \in R$. Nechť $x \in C$, $y \in C$ a $0 \leq \lambda \leq 1$. Pak existují body $x_i \in C_i$, $y_i \in C_i$, $1 \leq i \leq m$, 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 $x_i \in C_i$, $y_i \in C_i$, platí $z_i = \lambda x_i + (1-\lambda) y_i \in C_i$, $1 \leq i \leq m$, takže $\lambda x + (1-\lambda)y \in C$. □

Definice 62 *Konvexním obalem množiny $C \subset R^n$ nazveme průnik*

$$\text{conv } C = \bigcap_{C \subset C_\alpha} C_\alpha$$

všech konvexních množin $C_\alpha \subset R^n$ obsahujících C.

Poznámka 310 Zřejmě platí $C \subset \text{conv } C$.

Věta 193 *Konvexní obal množiny $C \subset R^n$ je množina všech konvexních kombinací bodů z C, tedy všech bodů tvaru*

$$y = \sum_{i=1}^m \lambda_i x_i, \tag{525}$$

kde $m \geq 1$, $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \ldots + \lambda_m = 1$.

Důkaz Nechť $\tilde{C}$ je množina všech konvexních kombinací bodů z C. Jelikož $\tilde{C}$ je konvexní, platí conv $C \subset \tilde{C}$. Nechť $y \in \tilde{C}$, takže $y = \lambda_1 x_1 + \ldots + \lambda_m x_m$, kde $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \ldots + \lambda_m = 1$. Jelikož $x_i \in C_\alpha$, $1 \leq i \leq m$, pro každou konvexní množinu $C_\alpha \subset R^n$ obsahující C, platí

$$y \in \text{conv } C = \bigcap_{C \subset C_\alpha} C_\alpha,$$

což dává $\tilde{C} \subset$ conv C □

Věta 194 *(Caratheodory) Nechť $y \in$ conv C, kde $C \subset R^n$. Pak existuje nejvýše $n+1$ bodů $x_i \in C$, $1 \leq i \leq n+1$, takových, že y je jejich konvexní kombinací.*

Důkaz Dokážeme, že pokud platí (525) s $m > n+1$, 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. Nechť tedy

$$y = \sum_{i=1}^{m} \lambda_i x_i,$$

kde $m > n+1$, $x_i \in C$, $\lambda_i \geq 0$, $1 \leq i \leq m$, $\lambda_1 + \ldots + \lambda_m = 1$. 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

$$\sum_{i=1}^{m} \alpha_i \hat{x}_i = 0. \tag{526}$$

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 (525) a (526) 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 $\lambda_i' \geq 0$, $1 \leq i \leq m$, $\lambda_j' = 0$, $\lambda_1' + \ldots + \lambda_m' = 1$, takže bod y je konvexní kombinací bodů $x_1, \ldots, x_{j-1}, x_{j+1}, \ldots x_m$, kterých je $m-1$. □

Věta 195 *Je-li množina C kompaktní, je i množina* conv C *kompaktní.*

Důkaz (a) Nechť $y \in$ conv C. Pak podle věty 194 existují vektory $x_i \in C$ a čísla $\lambda_i \geq 0$, $1 \leq i \leq n+1$, $\lambda_1 + \ldots + \lambda_{n+1} = 1$ takové, že $y = \lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1}$. Jelikož množina C je omezená, existuje číslo $M > 0$ takové, že $\|x_i\| \leq M$, $1 \leq i \leq n+1$. Pak ale $\|y\| = \|\lambda_1 x_1 + \ldots + \lambda_{n+1} x_{n+1}\| \leq (\lambda_1 + \ldots + \lambda_{n+1}) M = M$, takže množina conv C je omezená

(b) Nechť $\{y_i\} \subset$ conv C je posloupnost taková, že $y_i \to y \in R^n$. Máme dokázat, že $y \in$ conv C. Jelikož $y_i \in$ conv C, existují podle věty 194 vektory $x_i^k \in C$ a čísla $\lambda_i^k \geq 0$, $1 \leq k \leq n+1$, $\lambda_i^1 + \ldots + \lambda_i^{n+1} = 1$ takové, že $y_i = \lambda_i^1 x_i^1 + \ldots + \lambda_i^{n+1} x_i^{n+1}$. Protože množina C je kompaktní a číslo $n+1$ je konečné, lze vybrat podposloupnost $\{\tilde{y}_i\} \subset \{y_i\}$ takovou, že odpovídající podposloupnosti $\{\tilde{x}_i^k\} \subset \{x_i^k\}$, $\{\tilde{\lambda}_i^k\} \subset \{\lambda_i^k\}$, $1 \leq k \leq n+1$, jsou konvergentní, čili $\tilde{x}_i^k \to \tilde{x}^k \in C$, $\tilde{\lambda}_i^k \to \tilde{\lambda}^k \geq 0$, $1 \leq k \leq n+1$, $\tilde{\lambda}^1 + \ldots + \tilde{\lambda}^{n+1} = 1$. Jelikož vybraná posloupnost má stejnou limitu jako původní konvergentní posloupnost, lze psát

$$\begin{aligned} y &= \lim_{i \to \infty} y_i = \lim_{i \to \infty} \tilde{y}_i = \lim_{i \to \infty} \left(\sum_{k=1}^{n+1} \tilde{\lambda}_i^k \tilde{x}_i^k \right) \\ &= \sum_{k=1}^{n+1} \left(\lim_{i \to \infty} \tilde{\lambda}_i^k \right) \left(\lim_{i \to \infty} \tilde{x}_i^k \right) = \sum_{k=1}^{n+1} \tilde{\lambda}^k \tilde{x}^k \in \text{conv } C. \end{aligned}$$

□

Poznámka 311 Předpoklad omezenosti je ve větě 195 podstatný. Je-li C uzavřená ale neomezená, nemusí být conv C uzavřená. Nechť $C \in R^2$ a $C = C_1 \cap C_2$, kde C_1 je úsečka spojující body $x_1 = [-1, 0]$, $x_2 = [1, 0]$ a C_2 je polopřímka $[0, t]$, $t \geq 0$. Nechť $y_i = [1/i - 1, 1]$. Protože $x_1 = [-1, 0] \in C$ a $z_i = [0, i] \in C$, platí

$$y_i = [1/i - 1, 1] = [-1, 0] + \frac{1}{i}([0, i] - [-1, 0]) = x_1 + \frac{1}{i}(z_i - x_1) \in \text{conv } C.$$

Ale $y_i \to y = [-1, 1]$ a bod y nelze vyjádřit jako lineární kombinaci bodů z C, takže conv C není uzavřená.

Definice 63 *Nechť $C \subset R^n$. Pak funkci*

$$d_C(x) = \inf_{y \in C} \|y - x\|$$

nazveme vzdáleností bodu x od množiny C (nebo vzdálenostní funkcí množiny C).

Poznámka 312 Je-li množina $C \subset R^n$ uzavřená, platí

$$d_C(x) = \min_{y \in C} \|y - x\|.$$

Existuje tedy bod $y \in C$ takový, že $d_C(x) = \|y - x\|$. V dalším výkladu se omezíme na uzavřené množiny i když většina tvrzení má obecnější charakter.

Věta 196 *Nechť množina $C \subset R^n$ je uzavřená. Pak vzdálenostní funkce d_C je lipschitzovská v R^n s koeficientem $L = 1$. Je-li C konvexní, je d_C konvexní v R^n a ke každému bodu $x \in R^n$ existuje právě jeden bod $y \in C$ takový, že*

$$\|y - x\| = d_C(x).$$

Důkaz Nechť $x_1 \in R^n$, $x_2 \in R^n$. Jelikož množina C je uzavřená, existuje podle poznámky 312 bod $y \in C$ takový, že

$$\|y - x_1\| = d_C(x_1).$$

Platí tedy

$$d_C(x_2) \leq \|y - x_2\| \leq \|y - x_1\| + \|x_1 - x_2\| = d_C(x_1) + \|x_2 - x_1\|,$$

neboli

$$d_C(x_2) - d_C(x_1) \leq \|x_2 - x_1\|.$$

Protože nezáleží na pořadí bodů x_1, x_2, platí

$$|d_C(x_2) - d_C(x_1)| \leq \|x_2 - x_1\|,$$

takže funkce d_C je lipschitzovská v R^n s koeficientem $L = 1$ Nechť C je konvexní, $x_1 \in R^n$, $x_2 \in R^n$. Podle poznámky 312 existují body $y_1 \in C$, $y_2 \in C$ tak, že

$$\begin{aligned} \|y_1 - x_1\| &= d_C(x_1), \\ \|y_2 - x_2\| &= d_C(x_2). \end{aligned}$$

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 $y \in C$, takže platí

$$\begin{aligned} d_C(\lambda_1 x_1 + \lambda_2 x_2) &\leq \|y - \lambda_1 x_1 - \lambda_2 x_2\| \leq \lambda_1 \|y_1 - x_1\| + \lambda_2 \|y_2 - x_2\| \\ &= \lambda_1 d_C(x_1) + \lambda_2 d_C(x_2) \end{aligned}$$

a d_C je konvexní v R^n Nechť C je konvexní, $x \in R^n$ a $y_1 \in C$, $y_2 \in C$ jsou dva různé body takové, že $\|y_1 - x\| = d_C(x)$, $\|y_2 - x\| = d_C(x)$. 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

$$(y_2 - x)^T(y_1 - x) < d_C^2(x). \tag{527}$$

Položme nyní $y = \frac{1}{2}(y_2 + y_1)$. Jelikož C je konvexní, platí $y \in C$. Dále podle (527) platí

$$\|y - x\|^2 = \frac{1}{4}\left(\|y_2 - x\|^2 + \|y_1 - x\|^2 + 2(y_2 - x)^T(y_1 - x)\right) < d_C^2(x),$$

což je spor, neboť $y \in C$, takže podle poznámky 312 $d_C(x) \leq \|y - x\|$. □

Definice 64 *Nechť $C \subset R^n$ je uzavřená konvexní množina, $x \in R^n$ a $y \in C$ je bod takový, že $\|y - x\| = d_C(x)$. Pak řekneme, že y je projekcí bodu x do množiny C a píšeme $y = P_C(x)$.*

Věta 197 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \notin C$. Pak bod $y = P_C(x)$ je hraničním bodem množiny C.*

Důkaz Připomeňme, že bod $y \in R^n$ je hraničním bodem uzavřené množiny $C \subset R^n$, jestliže $y \in C$ a existuje posloupnost $\{x_i\} \subset R^n \setminus C$ 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 $x_i \notin C$ takovou, že $x_i \to y$. □

Lemma 71 *. Nechť $C \subset R^n$ je uzavřená konvexní množina, $x \in R^n$ a $y = P_C(x)$. Pak platí*

$$(x - y)^T(z - y) \leq 0 \quad \forall z \in C$$

Důkaz Jelikož $y \in C$, $z \in C$ a C je konvexní, platí $y + \lambda(z - y) = \lambda z + (1 - \lambda)y \in C$ $\forall 0 \leq \lambda \leq 1$. 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 $\varphi(0) = d_C(x)$ 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 $\|y + \lambda(z - y) - x\|^2 < d_C(x)$, což není možné, neboť $y + \lambda(z - y) \in C$ $\forall 0 \leq \lambda \leq 1$. □

Věta 198 *Nechť $C \subset R^n$ je uzavřená konvexní množina. Pak*

$$\|P_C(x_2) - P_C(x_1)\| \leq \|x_2 - x_1\| \quad \forall x_1, x_2 \in R^n.$$

Důkaz Nechť $y_1 = P_C(x_1)$ a $y_2 = P_C(x_2)$. Podle lemmatu 71 platí

$$\begin{aligned}(x_1 - y_1)^T(z_1 - y_1) &\leq 0 \quad \forall z_1 \in C,\\ (x_2 - y_2)^T(z_2 - y_2) &\leq 0 \quad \forall z_2 \in C.\end{aligned}$$

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 65 *Nechť $a \in R^n$ a $\alpha \in R$. Pak množinu*

$$H(a, \alpha) = \{y \in R^n : a^T y \leq \alpha\}$$

nazveme poloprostorem určeným normálovým vektorem a a číslem α.

Poznámka 313 Hranicí poloprostoru $H(a,\alpha)$ je nadrovina

$$L(a,\alpha) = H(a,\alpha) \cap H(-a,\alpha) = \{y \in R^n : a^T y = \alpha\}.$$

Číslo α určuje vzdálenost nadroviny $L(a,\alpha)$ od počátku. Tato vzdálenost se rovná podílu $\alpha/\|a\|$. Odtud plyne, že bod $y = 0$ je hraničním bodem poloprostoru $H(a,\alpha)$ (leží v hraniční nadrovině $L(a,\alpha)$) právě tehdy, když $\alpha = 0$.

Věta 199 *Poloprostor $H(a,\alpha)$ je uzavřenou konvexní množinou.*

Důkaz (a) Nechť $\{y_i\} \subset H(a,\alpha)$ 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 $y \in H(a,\alpha)$. Poloprostor $H(a,\alpha)$ je tedy uzavřený.
(b) Nechť $y_1 \in H(a,\alpha)$, $y_2 \in H(a,\alpha)$, 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 $y \in H(a,\alpha)$. Poloprostor $H(a,\alpha)$ je tedy konvexní. □

Věta 200 *Nechť C je uzavřená konvexní množina a nechť $x \notin C$. Pak existuje poloprostor $H(a,\alpha)$ takový, že $C \subset H(a,\alpha)$ a $x \notin H(a,\alpha)$. Tento poloprostor lze volit tak, že $a = x - P_C(x)$ a $\alpha = a^T P_C(x)$ (nebo $a = (x - P_C(x))/\|x - P_C(x)\|$ a $\alpha = a^T P_C(x)$). Pak $P_C(x) \in L(a,\alpha)$, takže $C \cap L(a,\alpha) \neq \emptyset$.*

Důkaz Máme dokázat, že existuje vektor $a \in R^n$ a číslo $\alpha \in R$ tak, že

$$a^T x > \alpha \geq a^T z \quad \forall z \in C.$$

Nechť $y = P_C(x)$, takže $\|y - x\| = d_C(x)$. 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 $x \neq C$, takže $\|x - y\| \neq 0$. Nerovnost $\alpha \geq a^T z \ \forall z \in C$ dokážeme sporem. Předpokládejme, že existuje bod $z \in C$ takový, že $\alpha = a^T y < a^T z$, a označme $z(\lambda) = y + \lambda b$, kde $b = z - y$. Zřejmě $z(\lambda) \in C$, pokud $0 \leq \lambda \leq 1$ (poznámka 307). Dále platí

$$\|z(\lambda) - x\|^2 = \|\lambda b - a\|^2 = \|a\|^2 - 2\lambda a^T b + \lambda^2 \|b\|^2$$

a

$$\left.\frac{d\|z(\lambda) - x\|^2}{d\lambda}\right|_{\lambda=0} = -2a^T b = -2(a^T z - a^T y) < 0.$$

Tedy $\|z(0) - x\|^2 = \|a\|^2$ a existuje číslo $0 < \overline{\lambda} \leq 1$ takové, že $\|z(\lambda) - x\|^2 < \|a\|^2 = d_C^2(x) \ \forall 0 < \lambda < \overline{\lambda}$, což je ve sporu s definicí $d_C(x)$. Jelikož jsme číslo α zvolili tak, že $\alpha = a^T y$, platí $y \in L(a,\alpha)$. □

Důsledek 23 *Nechť C_1, C_2 jsou uzavřené konvexní množiny takové, že $C_1 \cap C_2 = \emptyset$. Pak existuje poloprostor $H(a,\alpha)$ takový, že $C_1 \subset H(a,\alpha)$ a $C_2 \cap H(a,\alpha) = \emptyset$*

Důkaz Jelikož množiny C_1, C_2 jsou uzavřené, existují body $x_1 \in C_1$, $x_2 \in C_2$ takové, že

$$d(C_1, C_2) \stackrel{\Delta}{=} \inf_{y_1 \in C_1, y_2 \in C_2} \|y_2 - y_1\| = \min_{y_1 \in C_1, y_2 \in C_2} \|y_2 - y_1\| = \|x_2 - x_1\|$$

(argumentace je stejná jako v poznámce 312, dvojice $x_1 \in C_1$, $x_2 \in C_2$ nemusí být určena jednoznačně). Protože $x_2 \notin C_1$ plyne z věty 200 (a jejího důkazu), že $C_1 \subset H(a_1,\alpha_1)$ a $x_2 \notin H(a_1,\alpha_1)$, kde $a_1 = x_2 - x_1$

a $\alpha_1 = a_1^T x_1$. Podobně $C_2 \subset H(a_2, \alpha_2)$ a $x_1 \notin H(a_2, \alpha_2)$, kde $a_2 = x_1 - x_2$ a $\alpha_2 = a_2^T x_2$. Zbývá dokázat, že $H(a_1, \alpha_1) \cap H(a_2, \alpha_2) = \emptyset$ (pak lze volit $a = a_1$, $\alpha = \alpha_1$). Nechť $y \in H(a_1, \alpha_1) \cap H(a_2, \alpha_2)$. 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 201 *Uzavřená konvexní množina $C \subset R^n$ je průnikem všech poloprostorů obsahujících C.*

Důkaz Nechť $\tilde{C}$ je průnikem všech poloprostorů obsahujících uzavřenou konvexní množinu C. Jelikož kazdý poloprostor je podle věty 199 uzavřený a konvexní, je množina $\tilde{C}$ uzavřená a konvexní a platí. $C \subset \tilde{C}$. Stačí tedy dokázat, že $\tilde{C} \subset C$. Předpokládejme naopak, že existuje bod $x \in \tilde{C}$ takový, že $x \notin C$. Pak podle věty 528 existuje poloprostor H takový, že $C \subset H$ a $x \notin H$. Jelikož $C \subset H$, platí $C \subset \tilde{C} \subset H$, což je spor, neboť $x \in \tilde{C}$ a $x \notin H$. □

Definice 66 *Konvexní množina, která je průnikem konečného počtu poloprostorů, se nazývá polyedrální množinou.*

Definice 67 *Nechť C je uzavřená konvexní množina a $H(a, \alpha)$ je poloprostor s hranicí $L(a, \alpha)$ takový, že $C \subset H(a, \alpha)$ a $C \cap L(a, \alpha) \neq \emptyset$. Pak řekneme, že $H(a, \alpha)$ je tečným poloprostorem a $L(a, \alpha)$ tečnou nadrovinou množiny C.*

Poznámka 314 Ve větě 201 se můžeme omezit na tečné poloprostory (uzavřená konvexní množina je průnikem svých tečných poloprostorů). Obsahuje-li poloprostor $H(a, \alpha)$ konvexní množinu C, přižemž $C \cap L(a, \alpha) = \emptyset$, lze volbou $\alpha' = \max_{y \in C} a^T y$ docílit toho, že $C \subset H(a, \alpha') \subset H(a, \alpha)$ a $C \cap L(a, \alpha') \neq \emptyset$.

Věta 202 *Nechť bod $y \in R^n$ je hraničním bodem uzavřené konvexní množiny C. Pak existuje tečná nadrovina $L(a, \alpha)$ taková, že $y \in L(a, \alpha)$.*

Důkaz Jelikož $y \in C$ je hraničním bodem uzavřené konvexní množiny C, existuje posloupnost $\{x_i\} \subset R^n \setminus C$ taková, že $x_i \to y$. Pro každý bod $x_i \notin C$, $i \in N$, lze podle věty 200 sestrojit poloprostor $H(a_i, \alpha_i)$ takový, že $C \subset H(a_i, \alpha_i)$ a $P_C(x_i) \in L(a_i, \alpha_i)$, přičemž $a_i = (x_i - P_C(x_i))/\|x_i - P_C(x_i)\|$ a $\alpha_i = a_i^T P_C(x_i)$. Jelikož $x_i \to y$, platí podle věty 198 $P_C(x_i) \to P_C(y)$, 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 $C \subset H(a, \alpha)$ a $y \in L(a, \alpha)$. □

Definice 68 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Není-li bod x konvexní kombinací žádných bodů z C různých od x, řekneme, že x je krajním bodem nebo vrcholem množiny C.*

Poznámka 315 Z důkazu věty 190 plyne, že se v definici krajních bodů můžeme omezit na konvexní kombinace dvou bodů z C různých od x. Dále se můžeme omezit na průměry dvou bodů z 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 $x_3 \in C$ a $x = (x_1 + x_3)/2$. Bod $x \in C$ je tedy krajním bodem konvexní množiny C, neexistují-li dva body $x_1 \in C$, $x_3 \in C$ takové, že $x = (x_1 + x_3)/2$.

Věta 203 *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 $C \subset R^n$ je kompaktní konvexní množina a $x \in C$ není jejím krajním bodem.
(a) Předpokládejme nejprve, že x je hraničním bodem množiny C. Pak podle věty 202 existuje tečná nadrovina $L(a,\alpha)$ taková, že $x \in L(a,\alpha)$. Označme $\tilde{C} = C \cap L(a,\alpha)$. Jelikož C a $L(a,\alpha)$ jsou uzavřené konvexní množiny, C je kompaktní a $L(a,\alpha)$ má dimenzi nižší než n, je i množina $\tilde{C}$ kompaktní, konvexní a má dimenzi nižší než n. Podle indukčního předpokladu je tedy bod x konvexní kombinací krajních bodů množiny $\tilde{C}$. Zbývá dokázat, že krajní body množiny $\tilde{C}$ jsou také krajní body množiny C. Předpokládejme naopak, že bod y je krajním bodem množiny $\tilde{C}$, ale není krajním bodem množiny C. Pak podle poznámky 315 existují body $y_1 \in C \setminus L(a,\alpha)$, $y_2 \in C \setminus L(a,\alpha)$, takové, že $y = (y_1 + y_2)/2$. Jelikož $y \in L(a,\alpha)$, 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 $y_2 \in C$ a $C \subset H(a,\alpha)$, takže nutně $a^T y_2 \leq \alpha$.
(b) Je-li x vnitřním bodem množiny C, která je kompaktní, lze tímto bodem vést přímku, která protne hranici množiny 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 C, je i bod x konvexní kombinací krajních bodů množiny C. □

Definice 69 *Nechť $C \subset R^n$. Pak funkci*

$$\delta_C(x) = \sup_{y \in C} y^T x$$

nazveme opěrnou funkcí množiny C.

Poznámka 316 Nechť množina $C \subset R^n$ je kompaktní. Pak platí

$$\delta_C(x) = \max_{y \in C} y^T x.$$

Existuje tedy bod $y \in C$ takový, že $\delta_C(x) = y^T x$. V dalším výkladu se omezíme na kompaktní množiny i když většina tvrzení má obecnější charakter.

Věta 204 *Nechť množina $C \subset R^n$ je kompaktní. Pak opěrná funkce δ_C je pozitivně homogenní, subaditivní a lipschitzovská v R_n.*

Důkaz Podle poznámky 316 pro $x \in R^n$ a $\lambda \geq 0$ platí

$$\delta_C(\lambda x) = \max_{y \in C} y^T(\lambda x) = \lambda \max_{y \in C} y^T x = \lambda \delta_C(x),$$

takže funkce δ_C je pozitivně homogenní. Podobně pro $x_1 \in R^n$ a $x_2 \in R^n$ platí

$$\delta_C(x_1 + x_2) = \max_{y \in C} y^T(x_1 + x_2) \leq \max_{y \in C} y^T x_1 + \max_{y \in C} y^T x_2 = \delta_C(x_1) + \delta_C(x_2),$$

takže funkce δ_C je subaditivní. Ze subaditivity plyne nerovnost

$$\delta_C(x_2) \leq \delta_C(x_1) + \delta_C(x_2 - x_1)$$

a jelikož C je kompaktní existuje konstanta L taková, že $\|y\| \leq L \ \forall y \in C$. Můžeme tedy psát

$$\delta_C(x_2) - \delta_C(x_1) \leq \max_{y \in C} y^T(x_2 - x_1) \leq L\|x_2 - x_1\|.$$

Protože nezáleží na pořadí bodů x_1, x_2, platí

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

takže funkce δ_C je lipschitzovská v R^n. □

Věta 205 *Nechť množina $C \subset R^n$ je kompaktní. Pak*

$$\delta_C(x) = \delta_{\text{conv}\, C}(x) \quad \forall x \in R^n.$$

Důkaz Protože $C \subset \text{conv}C$, platí podle poznámky 316 $\delta_C(x) \leq \delta_{\text{conv}\, C}(x)\ \forall x \in R^n$. Nechť $x \in R^n$. Podle věty 194 lze každý vektor $y \in \text{conv}\ C$ vyjádřit jako konvexní kombinaci nejvýše $n+1$ vektorů $y_i \in C$, $1 \leq i \leq n+1$. Můžeme tedy psát

$$\begin{aligned} \delta_{\text{conv}\, C}(x) &= \max_{y \in \text{conv}\, C} y^T x = \max \left\{ \sum_{i=1}^{n+1} \lambda_i y_i^T x : y_i \in C, \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\} \\ &\leq \max_{y \in C} y^T x = \delta_C(x). \end{aligned}$$

□

Věta 206 *Nechť množiny $C_1 \subset R^n$, $C_2 \subset R^n$ jsou konvexní a kompaktní. Pak $C_1 \subset C_2$ platí právě tehdy, jestliže*

$$\delta_{C_1}(x) \leq \delta_{C_2}(x) \quad \forall x \in R^n.$$

Důkaz Jestliže $C_1 \subset C_2$, pak podle poznámky 316 platí $\delta_{C_1}(x) \leq \delta_{C_2}(x)\ \forall x \in R^n$. Předpokládejme, že $\delta_{C_1}(x) \leq \delta_{C_2}(x)\ \forall x \in R^n$ a existuje bod $\overline{y} \in C_1$ takový, že $\overline{y} \notin C_2$. Pak podle věty 200 existuje vektor $a \in R^n$ a číslo $\alpha \in R$ tak, že

$$a^T \overline{y} > \alpha \geq a^T y \quad \forall y \in C_2.$$

Platí tedy

$$\delta_{C_1}(a) \geq a^T \overline{y} > \delta_{C_2}(a),$$

což je ve sporu s předpokladem. □

Důsledek 24 *Nechť množina $C \subset R^n$ je konvexní a kompaktní. Pak $y \in C$ právě tehdy, jestliže*

$$y^T x \leq \delta_C(x) \quad \forall x \in R^n.$$

Věta 207 *Nechť množiny $C_1 \subset R^n$, $C_2 \subset R^n$ jsou kompaktní. Pak*

$$\delta_{C_1 + C_2}(x) = \delta_{C_1}(x) + \delta_{C_2}(x).$$

Důkaz Platí

$$\begin{aligned} \delta_{C_1+C_2}(x) &= \max_{y \in C_1 + C_2} y^T x = \max_{\substack{y_1 \in C_1 \\ y_2 \in C_2}} (y_1 + y_2)^T x = \max_{y_1 \in C_1} y_1^T x + \max_{y_2 \in C_2} y_2^T x \\ &= \delta_{C_1}(x) + \delta_{C_2}(x). \end{aligned}$$

□

Opěrná funkce množiny $C \subset R^n$ má bezprostřední vztah k poloprostorům obsahujícím tuto množinu.

Věta 208 *Množina $C \subset R^n$ leží v poloprostoru $H(a, \alpha)$ právě tehdy jestliže $\alpha \geq \delta_C(a)$, přičemž $H(a, \alpha)$ je tečným poloprostorem množiny C právě tehdy, jestliže $\alpha = \delta_C(a)$.*

Důkaz Tvrzení plyne z definice 69 a z toho, že $C \subset H(a, \alpha)$ právě tehdy, jestliže $\delta_C(a) = \sup_{y \in C} a^T y \leq \alpha$ a $C \cap L(a, \alpha) = \emptyset$, pokud $\delta_C(a) = \sup_{y \in C} a^T y < \alpha$. □

Definice 70 *Řekneme, že množina $K \subset R^n$ je kuželem, jestliže z $x \in K$ a $\lambda \geq 0$ plyne $\lambda x \in K$.*

Věta 209 *Průnik kuželů je kuželem.*

Důkaz Nechť $K = \cap_\alpha K_\alpha$, kde $K_\alpha \subset R^n$ jsou kužely. Nechť $x \in K$ a $\lambda \geq 0$. Pak platí $x \in K_\alpha$ a tedy $\lambda x \in K_\alpha\ \forall \alpha$. Odtud plyne, že $\lambda x \in K$. □

Věta 210 *Lineární kombinace kuželů je kuželem.*

Důkaz Nechť $K = \sum_{i=1}^{m} \lambda_i K_i$, kde $K_i \subset R^n$ jsou kužely a $\lambda_i \in R$. Nechť $x \in K$ a $\lambda \geq 0$. Pak existují body $x_i \in K_i$, $1 \leq i \leq m$, takové, že

$$\lambda x = \lambda \sum_{i=1}^{m} \lambda_i x_i = \sum_{i=1}^{m} \lambda_i (\lambda x_i).$$

Jelikož $x_i \in K_i$ a $\lambda \geq 0$, platí $\lambda x_i \in K_i$, $1 \leq i \leq m$, takže $\lambda x \in K$. □

Definice 71 *Kuželovým obalem množiny $C \subset R^n$ nazveme průnik*

$$\text{cone } C = \bigcap_{C \subset K_\alpha} K_\alpha$$

všech kuželů $K_\alpha \subset R^n$ obsahujících C.

Věta 211 *Nechť $C \subset R^n$. Pak platí*

$$\text{cone } C = \bigcup_{\lambda \geq 0} \lambda C = \{x \in R^n : \ x = \lambda y, \ \ y \in C, \ \ \lambda \geq 0\}$$

Kužel cone C *je tedy množinou všech nezáporných násobků bodů z C.*

Důkaz Nechť $\tilde{K} = \bigcup_{\lambda \geq 0} \lambda C$. Jelikož $\tilde{K}$ je kužel obsahující množinu C, platí cone $C \subset \tilde{K}$. Nechť naopak $y \in \tilde{K}$, takže $y = \lambda x$, kde $x \in C$ a $\lambda \geq 0$. Nechť K_α je libovolný kužel obsahující množinu C. Jelikož $x \in C$, platí $x \in K_\alpha$ a jelikož $\lambda \geq 0$, platí $y = \lambda x \in K_\alpha$. Tudíž $y \in$ cone C. □

Věta 212 *Množina $K \subset R^n$ je konvexním kuželem právě tehdy, obsahuje-li všechny nezáporné lineární kombinace svých bodů.*

Důkaz Obsahuje-li množina K všechny nezáporné lineární kombinace svých bodů, obsahuje též konvexní kombinace tvaru (524) a nezáporné násobky svých bodů, takže je konvexním kuželem. Nechť K je konvexním kuželem a $x_i \in K$, $\lambda_i \geq 0$, $1 \leq i \leq m$. Položme $\lambda = \lambda_1 + \ldots + \lambda_m$. Jestliže $\lambda = 0$, platí $\lambda_1 x_1 + \ldots \lambda_m x_m = 0 \in K$. 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 $\lambda_1' + \ldots + \lambda_m' = 1$. Jelikož množina K je konvexní, platí $x' \in K$, takže

$$x = \sum_{i=1}^{m} \lambda_i x_i = \lambda x' \in K.$$

□

Důsledek 25 *Množina $K \subset R^n$ je konvexním kuželem právě tehdy, obsahuje-li nezáporné násobky a součty svých bodů.*

Důkaz Obsahuje-li množina K nezáporné násobky a součty svých bodů, je kuželem podle definice 70 a z $x_1 \in K$, $x_2 \in K$ a $0 \leq \lambda \leq 1$ plyne $\lambda x_1 \in K$, $(1-\lambda)x_2 \in K$, takže $\lambda x_1 + (1-\lambda)x_2 \in K$ a K je konvexní množinou. Opačná implikace plyne bezprostředně z věty 212. □

Věta 213 *Množina* cone (conv C) *je množinou všech nezáporných lineárních kombinací bodů z C.*

Důkaz Nechť $\tilde{K}$ je množina všech nezáporných lineárních kombinací bodů z C. Jelikož nezáporná lineární kombinace nezáporných lineárních kombinací je opět nezápornou lineární kombinací je podle věty 212 $\tilde{K}$ konvexním kuželem. Podle definice 71 tedy platí cone (conv C) $\subset \tilde{K}$. Nechť $x \in \tilde{K}$ a $x \notin$ cone (conv C). Pak podle věty 200 existuje poloprostor $H(a,0)$ takový, že cone (conv C) $\subset H(a,0)$ a $x \notin H(a,0)$. Jelikož $H(a,0)$ je konvexním kuželem, je podle vět 191 a 209 i $H(a,0) \cap$ cone (conv C) konvexním kuželem, což je spor s minimalitou $\tilde{K}$ (definice 71). □

Jelikož uzavřený konvexní kužel je uzavřenou konvexní množinou, můžeme studovat tečné poloprostory uzavřených konvexních kuželů.

Věta 214 *Tečný poloprostor uzavřeného konvexního kuželu je uzavřeným konvexním kuželem (takže obsahuje počátek souřadnic). Jestliže $K \subset H(a,0)$, je $H(a,0)$ tečným poloprostorem uzavřeného konvexního kuželu K.*

Důkaz Nechť $H(a,\alpha)$ je tečným poloprostorem uzavřeného konvexního kuželu K. Jelikož $L(a,\alpha) \cap K \neq \emptyset$, existuje bod $y \in K$ takový, že $a^T y = \alpha$. Protože K je kuželem, musí platit $\lambda y \in K \subset H(a,\alpha)\ \forall \lambda \geq 0$, 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 $0 \in H(a,0)$ a pokud $x \in H(a,0)$, pak také $\lambda x \in H(a,0)\ \forall \lambda \geq 0$. Konvexita $H(a,0)$ plyne z věty 199. Zbytek tvrzení plyne z poznámky 314. □

Definice 72 *Nechť $C \in R^n$. Množinu*

$$C^* = \{x \in R^n : y^T x \leq 0 \quad \forall y \in C\}$$

nazveme polárním kuželem množiny C.

Poznámka 317 Z definice 72 lze snadno usoudit, že z $C_1 \subset C_2$ plyne $C_2^* \subset C_1^*$.

Věta 215 *Nechť $C \subset R^n$. Pak množina C^* je uzavřeným konvexním kuželem.*

Důkaz (a) Nechť $\{x_i\} \subset C^*$ je posloupnost taková, že $x_i \to x$. Jelikož $y^T x_i \leq 0\ \forall i \in N\ \forall y \in C$ a neostrá nerovnost je invariantní vůči limitnímu přechodu, platí

$$y^T x = \lim_{i\to\infty} y^T x_i \leq 0 \quad \forall y \in C,$$

takže $x \in C^*$.
(b) Nechť $x_1 \in C^*$, $x_2 \in C^*$. Pak platí $y^T x_1 \leq 0$, $y^T x_2 \leq 0\ \forall y \in C$. Nechť $0 \leq \lambda \leq 1$ a $x = \lambda x_1 + (1-\lambda)x_2$. Pak

$$y^T x = y^T(\lambda x_1 + (1-\lambda)x_2) = \lambda y^T x_1 + (1-\lambda) y^T x_2 \leq 0 \quad \forall y \in C,$$

takže $x \in C^*$.
(c) Nechť $x \in C^*$ a $\lambda \geq 0$. Pak platí

$$y^T(\lambda x) = \lambda y^T x \leq 0 \quad \forall y \in C,$$

takže $\lambda x \in C^*$. □

Věta 216 *Je-li $K \subset R^n$ uzavřeným konvexním kuželem, platí $(K^*)^* = K$.*

Důkaz Nechť $y \in (K^*)^*$. Pak podle definice 72 platí $y^T x \leq 0 \quad \forall x \in K^*$, takže $y \in K$. Nechť naopak $z \notin K$. Jelikož K je uzavřeným konvexním kuželem, existuje podle věty 200 a věty 212 poloprostor $H(x,0)$ takový, že $K \subset H(x,0)$ a $z \notin H(x,0)$, neboli $x^T y \leq 0\ \forall y \in K$ (takže $x \in K^*$) a $x^T z > 0$. Protože $x \in K^*$ a $x^T z > 0$, musí platit $z \notin (K^*)^*$. □

Věta 217 *Nechť $K \subset R^n$ je uzavřený konvexní kužel. Pak*

$$K^* = \{x \in R^n : P_K(x) = 0\}.$$

Důkaz Nechť $P_K(x) = 0$. Pak podle věty 200 existuje tečný poloprostor množiny K s normálovým vektorem $a = x - P_K(x) = x$ a číslem $\alpha = (x - P_K(x))^T P_K(x) = 0$, takže $x^T y \leq 0 \ \forall y \in K$, neboli $x \in K^*$. Nechť naopak $x \in K^*$. Pak $x^T y \leq 0 \ \forall y \in K$, takže $K \subset H(x, 0)$, a jelikož $P_{H(x,0)}(x) = 0$, platí též $P_K(x) = 0$. □

Věta 218 *Nechť $K \subset R^n$ je uzavřený konvexní kužel. Pak K^* je sjednocením normálových vektorů tečných poloprostorů kuželu K, neboli*

$$K^* = \bigcup_{K \subset H(a,0)} a.$$

Důkaz Označme $\tilde{K}^*$ množinu na pravé straně dokazované rovnosti. Nechť $a \in \tilde{K}^*$. Pak $H(a, 0)$ je tečným poloprostorem kuželu K, takže $a^T y \leq 0 \ \forall y \in K$, neboli $a \in K^*$. Nechť naopak $a \in K^*$. Pak $a^T y \leq 0$ $\forall y \in K$, takže $K \subset H(a, 0)$. Jelikož $H(a, 0)$ je podle věty 212 tečným poloprostorem množiny K, platí $a \in \tilde{K}^*$. □

Definice 73 *Kužel $K \in R^n$, který je průnikem konečného počtu tečných poloprostorů, se nazývá polyedrálním kuželem*

Věta 219 *Nechť $K \in R^n$ je polyedrální kužel takový, že*

$$K = \bigcap_{i=1}^{m} H(a_i, 0)$$

Pak

$$K^* = \text{cone}\,(\text{conv}\,\{a_i : 1 \leq i \leq m\}).$$

Důkaz Označme $\tilde{K}^*$ množinu na pravé straně dokazované rovnosti. Nechť $a \in \tilde{K}^*$. Pak podle věty 213 existují čísla $\lambda_i \geq 0$, $1 \leq i \leq m$, taková, že

$$a = \sum_{i=1}^{m} \lambda_i a_i.$$

Jelikož $a_i^T y \leq 0 \ \forall y \in K$ a $\lambda_i \geq 0$, platí $a^T y \leq 0 \ \forall y \in K$, takže $a \in K^*$. Nechť naopak $a \notin \tilde{K}^*$. Pak podle věty 200 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 $x \in H(a_i, 0)$, $1 \leq i \leq m$, neboli $x \in K$, a jelikož $x^T a > 0$, musí platit $a \notin K^*$. □

Definice 74 *Nechť $C \subset R^n$ je uzavřená množina a $x \in C$. Tečným kuželem množiny C v bodě x nazveme množinu*

$$T_C(x) = \{y \in R^n : \text{ existují posloupnosti } y_i \to y, \ t_i \downarrow 0 \text{ takové, že } x + t_i y_i \in C\}$$

Věta 220 *Nechť $C \subset R^n$ je uzavřená množina a $x \in C$. Pak $T_C(x)$ je uzavřeným kuželem. Je-li C konvexní, je i $T_C(x)$ konvexní.*

Důkaz (a) Nechť $y^k \in T_C(x)$, $y^k \to y$ a $\varepsilon > 0$. Pak existuje index $\overline{k} \in N$ takový, že $\|y^k - y\| < \varepsilon/2$ $\forall k \geq \overline{k}$. Jelikož $y^k \in T_C(x)$, existují posloupnosti

$$y_i^k \to y^k, \quad t_i^k \downarrow 0$$

takové, že $x + t_i^k y_i^k \in C$. Pro každé $k \in N$ tedy existuje index $\overline{i}_k \in N$ takový, že

$$\|y_i^k - y^k\| < \varepsilon/2, \quad t_i^k < 1/k, \ \text{ a } \ x + t_i^k y_i^k \in C$$

$\forall i \geq \bar{i}_k$. Zkonstruujeme-li posloupnost indexů $\{i_k\} \subset N$ rekurentním předpisem $i_1 = \bar{i}_1$ a $i_{k+1} = \max(i_k + 1, \bar{i}_{k+1})$, platí

$$\|y_{i_k}^k - y^k\| < \varepsilon/2, \quad t_{i_k}^k < 1/k \text{ a } x + t_{i_k} y_{i_k} \in C$$

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 $x + t_{i_k}^k y_{i_k}^k \in C$, což implikuje $y \in T_C(x)$, takže množina $T_C(x)$ je uzavřená.
(b) Nechť $y \in T_C(x)$ a $\lambda \geq 0$. Podle definice 74 existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in C$. Pak ale

$$\lambda y_i \to \lambda y, \quad t_i/\lambda \downarrow 0 \text{ a } x + (t_i/\lambda)\lambda y_i = x + t_i y_i \in C,$$

takže $\lambda y \in T_C(x)$ a množina $T_C(x)$ je kuželem.
(c) Nechť $y^1 \in T_C(x)$ a $y^2 \in T_C(x)$. Podle definice 74 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 $x + t_i^1 y_i^1 \in C$, $x + t_i^2 y_i^2 \in C$. Je-li C konvexní, platí podle poznámky 307 $x + t_i y_i^1 \in C$, $x + t_i y_i^2 \in C$, kde $t_i = \min(t_i^1, t_i^2)$. 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

$$x + t_i y_i = \lambda(x + t_i y_i^1) + (1-\lambda)(x + t_i y_i^2) \in C,$$

takže $y \in T_C(x)$ a množina $T_C(x)$ je konvexní. □

Věta 221 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak*

$$T_C(x) = \overline{\bigcup_{\lambda \geq 0} \lambda(C - x)}$$

Důkaz Označme

$$K = \text{cone } (C - x) = \bigcup_{\lambda \geq 0} \lambda(C - x).$$

(a) Nechť $z \in C$, $\lambda \geq 0$ a $y = \lambda(z - x)$. Nechť $y_i = y \; \forall i \in N$ a $t_i \downarrow 0$, přičemž $t_i' = \lambda t_i \leq 1$. Pak

$$x + t_i y_i = x + t_i y = x + \lambda t_i (z - x) = x + t_i'(z - x) \in C$$

podle poznámky 307, takže $y \in T_C(x)$. Platí tedy $K \subset T_C(x)$ a jelikož $T_C(x)$ je uzavřená množina, též $\overline{K} \subset T_C(x)$.
(b) Nechť naopak $y \in T_C(x)$. Pak existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in C$. Označme $z_i = x + t_i y_i \in C$. Pak $y_i = (z_i - x)/t_i$, takže $y_i \in K$. Jelikož $y_i \to y$, platí $y \in \overline{K}$, takže $T_C(x) \subset \overline{K}$. □

Věta 222 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$ je jejím hraničním bodem. Pak $T_C(x)$ je průnikem všech tečných poloprostorů množiny $C - x$ obsahujících počátek souřadnic, neboli*

$$T_C(x) = \bigcap_{C - x \subset H(a,0)} H(a, 0).$$

Je-li množina $C \in R^n$ polyedrální, je i tečný kužel $T_C(x)$ polyedrální a existují tečné poloprostory $H(a_i, 0)$, $1 \leq i \leq m$, takové, že

$$T_C(x) = \bigcap_{i=1}^{m} H(a_i, 0).$$

Důkaz (a) Označme $\overline{K}$ průnik všech tečných poloprostorů množiny $C-x$ obsahujících počátek souřadnic. Nechť $y \in \text{cone}\,(C-x)$. Pak existuje bod $z \in C-x$ takový, že $y = \lambda z$, $\lambda \geq 0$, a pro libovolný tečný poloprostor $H(a,0)$ množiny $C-x$ platí $a^T y = \lambda a^T z \leq 0$, neboli $y \in H(a,0)$. Platí tedy cone $(C-x) \subset \overline{K}$ a jelikož $\overline{K}$ je uzavřeným kuželem, též $T_C(x) \subset \overline{K}$.
(b) Nechť $y \in \overline{K}$ a $y \notin T_C(x)$. Jelikož $T_C(x)$ je uzavřená konvexní množina, existuje podle věty 200 tečný poloprostor této množiny takový, že $T_C(x) \subset H$ a $y \notin H$. Podle věty 212 je $0 \in H$, takže H je tečným poloprostorem množiny $C-x$ obsahujícím počátek souřadnic. Platí tedy $\overline{K} \subset H$ a jelikož $y \notin H$, musí být $y \notin \overline{K}$, což je ve sporu s předpokladem, že $y \in \overline{K}$.
(c) Je-li množina $C \in R^n$ polyedrální, má tuto vlastnost i množina $C-x$. Jelikož $C-x$ je průnikem konečného počtu poloprostorů, lze i v průniku definujícím $T_C(x)$ vybrat konečný počet poloprostorů. □

Definice 75 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Normálovým kuželem množiny C v bodě x nazveme množinu*

$$N_C(x) = T_C^*(x),$$

kde $T_C^*(x)$ je polární kužel tečného kuželu $T_C(x)$.

Poznámka 318 Podle věty 215 je množina $N_C(x)$ uzavřeným konvexním kuželem.

Věta 223 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak*

$$N_C(x) = \{z \in R^n : (y-x)^T z \leq 0 \quad \forall y \in C\}.$$

Důkaz Platí

$$\begin{aligned} N_C(x) &= \{z \in R^n : (y-x)^T z \leq 0 \quad \forall (y-x) \in T_C(x)\} \\ &= \{z \in R^n : (y-x)^T z \leq 0 \quad \forall (y-x) \in \overline{\bigcup_{\lambda \geq 0} \lambda(C-x)}\} \\ &= \{z \in R^n : (y-x)^T z \leq 0 \quad \forall (y-x) \in \bigcup_{\lambda \geq 0} \lambda(C-x)\} \\ &= \{z \in R^n : (y-x)^T z \leq 0 \quad \forall y \in C\}. \end{aligned}$$

První rovnost plyne z definic 72 a 75, druhá z věty 221, třetí z invariance neostré nerovnosti vůči limitnímu přechodu a poslední z invariance neostré nerovnosti vůči násobení nezáporným číslem λ. □

Věta 224 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$ je jejím hraničním bodem. Pak $N_C(x)$ je sjednocením normálových vektorů tečných poloprostorů množiny $C-x$ obsahujících počátek souřadnic, neboli*

$$N_C(x) = \bigcup_{C-x \subset H(a,0)} a.$$

Je-li množina $C \in R^n$ polyedrální, je i normálový kužel $T_C(x)$ polyedrální a existují tečné poloprostory $H(a_i,0)$, $1 \leq i \leq m$, takové, že

$$N_C(x) = \text{cone}\,(\text{conv}\,\{a_i : 1 \leq i \leq m\}).$$

Důkaz Toto tvrzení je důsledkem věty 218, věty 219 a věty 222. □

15.2 Konvexní funkce

Definice 76 *Řekneme, že funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$, jestliže existuje číslo $\varepsilon > 0$ tak, že F je definovaná v $B(x,\varepsilon) = \{y : \|y-x\| < \varepsilon\}$ a platí*

$$F(\lambda x_1 + (1-\lambda)x_2) \leq \lambda F(x_1) + (1-\lambda)F(x_2), \tag{528}$$

pokud $x_1 \in B(x,\varepsilon)$, $x_2 \in B(x,\varepsilon)$ a $0 \leq \lambda \leq 1$. Řekneme, že funkce $F : R^n \to R$ je konvexní na konvexní množině $C \subset R^n$, platí-li (528) pokud $x_1 \in C$, $x_2 \in C$ a $0 \leq \lambda \leq 1$.

Poznámka 319 Nerovnost (528) můžeme zapsat v ekvivalentním tvaru

$$F(x_2 + \lambda(x_1 - x_2)) \leq F(x_2) + \lambda(F(x_1) - F(x_2)).$$

Poznámka 320 Indukcí snadno dokážeme, že z $x_i \in C$, $\lambda_i \geq 0$ a $\sum_{i=1}^m \lambda_i = 1$ plyne

$$F\left(\sum_{i=1}^m \lambda_i x_i\right) \leq \sum_{i=1}^m \lambda_i F(x_i),$$

pokud F je konvexní na C (princip důkazu je shodný s postupem uvedeným v důkazu věty 190).

Věta 225 *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 $B(x, \varepsilon\sqrt{n+1})$ a tudíž i v nadkrychli

$$\overline{H(x,\varepsilon)} = \{y \in R^n : x_i - \varepsilon \leq y_i \leq x_i + \varepsilon,\ 1 \leq i \leq n\} \subset B(x, \varepsilon\sqrt{n+1}).$$

Nechť $y^{(k)}$, $1 \leq k \leq 2^n$, jsou vrcholy této nadkrychle. Označme

$$M = \max_{1 \leq k \leq 2^n} F(y^{(k)}).$$

Jelikož každý bod $\overline{H(x,\varepsilon)}$ lze podle věty 203 vyjádřit jako konvexní kombinaci vrcholů $y^{(k)}$, $1 \leq k \leq 2^n$, platí to i o bodech okolí $B(x,\varepsilon) \subset \overline{H(x,\varepsilon)}$. Nechť tedy $y \in B(x,\varepsilon)$. Pak platí

$$y = \sum_{k=1}^{2^n} \lambda_k y^{(k)}, \quad \sum_{k=1}^{2^n} \lambda_k = 1,$$

kde $\lambda_k \geq 0$, $1 \leq k \leq 2^n$, takže

$$F(y) = F\left(\sum_{k=1}^{2^n} \lambda_k y^{(k)}\right) \leq \sum_{k=1}^{2^n} \lambda_k F(y^{(k)}) \leq M \sum_{k=1}^{2^n} \lambda_k = M.$$

Funkce F je tedy omezená shora na $B(x,\varepsilon)$. Zvolme nyní $y \in B(x,\varepsilon)$ a $y' = 2x - y$. Pak $\|y' - x\| = \|x - y\| < \varepsilon$ takže $y' \in B(x,\varepsilon)$. Z konvexity plyne

$$F(x) = F\left(\frac{y + y'}{2}\right) \leq \frac{1}{2}(F(y) + F(y')),$$

takže

$$F(y) \geq 2F(x) - F(y') \geq 2F(x) - M$$

a funkce F je omezená zdola na $B(x,\varepsilon)$. Položme $\delta = \varepsilon/2$ a $m = 2F(x) - M$. Nechť $z \in B(x,\delta)$, $z' \in B(x,\delta)$ a $z \neq z'$. Položme

$$z'' = z' + \delta\frac{z' - z}{\|z' - z\|} \in B(x,\varepsilon).$$

Přímým výpočtem dostaneme

$$z' = \frac{\|z' - z\|}{\delta + \|z' - z\|} z'' + \frac{\delta}{\delta + \|z' - z\|} z$$

a z konvexity plyne

$$\begin{aligned} F(z') - F(z) &\leq \frac{\|z' - z\|}{\delta + \|z' - z\|} F(z'') + \frac{\delta}{\delta + \|z' - z\|} F(z) - F(z) \\ &= \frac{\|z' - z\|}{\delta + \|z' - z\|} (F(z'') - F(z)) \leq \frac{1}{\delta} \|z' - z\| (M - m). \end{aligned}$$

Jelikož nezáleží na pořadí bodů z a z', dostaneme

$$|F(z') - F(z)| \leq \frac{M - m}{\delta} \|z' - z\|,$$

takže F je lipschitzovská s konstantou $L = (M - m)/\delta$ na $B(x, \delta)$. □

Lemma 72 *Nechť funkce $\varphi : R \to R$ je konvexní na intervalu $[a, b]$ a nechť $a \leq t_1 < t_2 < t_3 \leq b$. Pak platí*

$$\frac{\varphi(t_2) - \varphi(t_1)}{t_2 - t_1} \leq \frac{\varphi(t_3) - \varphi(t_1)}{t_3 - t_1} \leq \frac{\varphi(t_3) - \varphi(t_2)}{t_3 - t_2}.$$

Důkaz Platí

$$t_2 = t_1 + \frac{t_2 - t_1}{t_3 - t_1} (t_3 - t_1),$$

kde

$$0 \leq \frac{t_2 - t_1}{t_3 - t_1} \leq 1.$$

Z konvexity funkce φ (poznámka 319) 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

$$\varphi(t_3) \leq \varphi(t_2) + \frac{t_3 - t_2}{t_3 - t_1} (\varphi(t_3) - \varphi(t_1)).$$

□

Definice 77 *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*

$$F'(x, h) = \lim_{t \downarrow 0} \frac{F(x + th) - F(x)}{t}. \tag{529}$$

Věta 226 *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$ (takže je lipschitzovská s nějakou konstantou L v okolí tohoto bodu). Pak:*

(a) Směrová derivace $F'(x,h)$ existuje pro každé $h \in R^n$. Navíc existuje číslo $\varepsilon > 0$ takové, že

$$F'(x,h) = \inf_{0<t\|h\|<\varepsilon} \frac{F(x+th)-F(x)}{t}.$$

(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

$$\limsup_{i\to\infty} F'(x_i,h_i) \le F'(x,h),$$

kdykoliv $x_i \to x$ a $h_i \to h$.

Důkaz (a) Nechť funkce F je konvexní v $B(x,\varepsilon)$. Podle lemmatu 72 je funkce

$$\varphi(t) = \frac{F(x+th)-F(x)}{t}$$

neklesající (levá nerovnost) a zdola omezená pro $0 < t\|h\| < \varepsilon$ (spojením obou nerovností dostaneme $(F(x+th)-F(x))/t \ge (F(x)-F(x-t'h))/t'$ pro libovolné $0 < t'h < \varepsilon$, přičemž výraz na levé straně poslední nerovnosti je konečný, neboť funkce F je spojitá). Existuje tedy limita (529). Zbytek tvrzení (a) plyne z toho, že $\varphi(t)$ je neklesající pro $0 < t\|h\| < \varepsilon$.
(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 $x + t_i h$, $x + t_i h_i$ a $x_i + t_i h_i$ leží v $B(x,\varepsilon)$. Pak podle (a) platí

$$F'(x_i,h_i) \le \frac{F(x_i + t_i h_i) - F(x_i)}{t_i} = \frac{F(x + t_i h) - F(x)}{t_i} + \frac{F(x_i + t_i h_i) - F(x + t_i h)}{t_i} + \frac{F(x) - F(x_i)}{t_i}.$$

Ale

$$\frac{|F(x_i + t_i h_i) - F(x + t_i h)|}{t_i} \le \frac{L(\|x_i - x\| + t_i\|h_i - h\|)}{t_i} \le L(\sqrt{\|x_i - x\|} + \|h_i - h\|) \to 0$$

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

$$\limsup_{i\to\infty} F'(x_i,h_i) \le \limsup_{i\to\infty} \frac{F(x + t_i h) - F(x)}{t_i} = \lim_{t\downarrow 0} \frac{F(x + th) - F(x)}{t} = F'(x,h)$$

□

Poznámka 321 Z části (b) důkazu věty 226 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 322 Podle definice 77 platí $F'(x,0) = 0$, takže podle věty 226 (b) dostaneme

$$|F'(x,h)| = |F'(x,h) - F'(x,0)| \le L\|h\|.$$

Definice 78 *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak množinu*

$$\partial F(x) = \left\{ g \in R^n : F'(x,h) \ge g^T h \ \ \forall h \in R^n \right\}$$

nazveme subdiferenciálem funkce F v bodě x. Elementy $g \in \partial F(x)$ budeme nazývat subgradienty funkce F v bodě x.

Věta 227 *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$ (takže je v tomto okolí lipschitzovská s nějakou konstantou L). Pak:*

(a) Subdiferenciál $\partial F(x)$ je neprázdná konvexní kompaktní množina taková, že $\|g\| \le L \ \forall g \in \partial F(x)$.

(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) Existuje číslo $\varepsilon > 0$ takové, že pro libovolný vektor $g \in \partial F(x)$ platí

$$F(x + h) - F(x) \ge g^T h \quad \forall h \in B(0,\varepsilon).$$

Důkaz (a) Podle věty 226 (b) je funkce $F'(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (tvrzení 10) 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 78 a poznámky 322 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 78 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

$$F'(x,\overline{h}) > \max\left\{g^T \overline{h} : g \in \partial F(x)\right\}. \tag{530}$$

Uvažujme lineární funkci $l(\lambda\overline{h}) \stackrel{\Delta}{=} \lambda F'(x,\overline{h})$ definovanou na jednorozměrném podprostoru $\{\lambda\overline{h} : \lambda \in R\} \subset R^n$. Jelikož je $F'(x,\cdot)$ 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 (530).
(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 226 (c) je funkce $F'(\cdot,.)$ shora polospojitá, takže $g^T h \leq F'(x,h)$.
(d) Nechť funkce F je konvexní v $B(x,\varepsilon)$ a $g \in \partial F(x)$. Podle definice 78 a věty 226 (a) platí

$$g^T h \leq F'(x,h) \leq \frac{F(x+th) - F(x)}{t}$$

pro $0 < t \leq 1$ a $h \in B(0,\varepsilon)$. Zvolíme-li $t = 1$, dostaneme dokazovanou nerovnost. □

Poznámka 323 Porovnáme-li větu 227 (b) s poznámkou 316, vidíme, že směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$F'(x,h) = \delta_{\partial F(x)}(h).$$

Věta 228 *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 $x \in R^n$, platí

$$F'(x,h) = (\nabla F(x))^T h.$$

Nechť $g \in \partial F(x)$. Pak podle definice 78 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, že $(\nabla F(x))^T h > g^T h$, neboť by muselo platit $(\nabla F(x))^T(-h) < g^T(-h)$, což je nemožné. Tedy $(\nabla F(x))^T h = g^T h \ \forall h \in R^n$, neboli $g = \nabla F(x)$. □

Věta 229 *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, jestliže*

$$0 \in \partial F(x).$$

Důkaz Podle věty 226 (a) má funkce $F : R^n \to R$ v bodě $x \in R^n$ lokální minimum právě tehdy, jestliže $F'(x,h) \geq 0$, $\forall h \in R^n$. Podle definice 78 tedy platí $0 \in \partial F(x)$. Jestliže $0 \in \partial F(x)$, existuje podle věty 227 (d) číslo $\varepsilon > 0$ takové, že $F(x+h) - F(x) \geq 0 \ \forall h \in B(x,\varepsilon)$, takže F má v bodě x lokální minimum. □

Některé další vlastnosti subdiferenciálů konvexních funkcí budou v obecnější podobě uvedeny v následujícím oddílu. Ukážeme ještě, jak lze vlastnosti konvexních funkcí použít k vyšetřování konvexních množin.

Věta 230 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak platí*

$$T_C(x) = \{y \in R^n : d'_C(x,y) = 0\}$$

($d'_C(x,y)$ je směrová derivace funkce $d_C(x)$ ve směru $y \in R^n$).

Důkaz (a) Označme $K = \{y \in R^n : d'_C(x,y) = 0\}$. Předpokládejme nejprve, že $y \in T_C(x)$. Pak existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in C$. Jelikož $d'_C(x,y) \geq 0$ (plyne to z toho, že $d_C(x) = 0$ a $d_C(z) \geq 0 \ \forall z \in R^n$), stačí dokázat, že $d'_C(x,y) \leq 0$. Platí

$$d'_C(x,y) = \lim_{t \downarrow 0} \frac{d_C(x+ty) - d_C(x)}{t} = \lim_{t \downarrow 0} \frac{\min_{z \in C} \|x + ty - z\|}{t} \leq \lim_{t \downarrow 0} \frac{\min_{z \in C} \|x + ty_i - z\| + t\|y - y_i\|}{t}$$

$\forall i \in N$. Ale

$$\min_{z \in C} \|x + ty_i - z\| = \min_{z \in C} \left\| \left(1 - \frac{t}{t_i}\right) x + \frac{t}{t_i}(x + t_i y_i) - z \right\| = 0,$$

pokud $t \leq t_i$, neboť v tomto případě platí $0 \leq t/t_i \leq 1$, takže

$$\left(1 - \frac{t}{t_i}\right) x + \frac{t}{t_i}(x + t_i y_i) \in C.$$

Můžeme tedy psát

$$d'_C(x,y) \leq \lim_{t \downarrow 0} \frac{t\|y - y_i\|}{t} = \|y - y_i\| \quad \forall i \in N$$

a jelikož $y_i \to y$, dostaneme $d'_C(x,y) \leq 0$. Tedy $d'_C(x,y) = 0$, čili $y \in K$. Platí tedy $T_C(x) \subset K$.

(b) Nechť $y \in K$ a $t_i \downarrow 0$. Z definice množiny K plyne, že

$$d'_C(x,y) = \lim_{i\to\infty} \frac{d_C(x+t_iy)}{t_i} = 0.$$

Nechť body $z_i \in C$, $i \in N$, jsou zvoleny tak, že

$$\|x + t_iy - z_i\| \le d_C(x+t_iy) + \frac{t_i}{i}$$

(což je možné vzhledem k definici vzdálenosti $d_C(x+t_iy)$). Položme $y_i = (z_i - x)/t_i$, $i \in N$. Pak platí

$$x + t_iy_i = x + (z_i - x) = z_i \in C$$

a

$$\|y - y_i\| = \|y - \frac{z_i - x}{t_i}\| = \frac{1}{t_i}\|x + t_iy - z_i\| \le \frac{d_C(x+t_iy)}{t_i} + \frac{1}{i},$$

takže

$$\lim_{i\to\infty} \|y - y_i\| = d'_C(x,y) + \lim_{i\to\infty} \frac{1}{i} = 0.$$

Tedy $y_i \to y$, $t_i \downarrow 0$ a $x + t_iy_i \in C$, takže $y \in T_C(x)$. Platí tedy $K \subset T_C(x)$. □

Věta 231 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak platí*

$$N_C(x) = \overline{\bigcup_{\lambda\ge 0} \lambda\partial d_C(x)}$$

Důkaz (a) Předpokládejme, že $z \in \partial d_C(x)$. Pak podle definice 78 platí

$$d'_C(x,y) \ge z^T y \quad \forall y \in R^n.$$

Jestliže $y \in T_C(x)$, platí podle věty 230 $d'_C(x,y) = 0$, takže

$$z^T y \le 0 \quad \forall y \in T_C(x),$$

což podle definic 72 a 75 dává $z \in N_C(x)$. Jelikož $N_C(x)$ je uzavřený konvexní kužel, platí

$$\overline{\bigcup_{\lambda\ge 0} \lambda\partial d_C(x)} \subset N_C(x)$$

(b) Nechť $z \in N_C(x)$. Pak podle definic 72, 75 a věty 230 platí

$$z^T y \le 0 = d'_C(x,y) = \lambda(y) d'_C(x,y) \quad \forall y \in T_C(x),$$

kde $\lambda(y) = 1$ $\forall y \in T_C(x)$. Zbývá dokázat podobnou nerovnost i pro $y \notin T_C(x)$ (kde obecně $\lambda(y) \ne 1$). Nechť $y \notin T_C(x)$. Jelikož $d'_C(x,y) > 0$ pro $y \notin T_C(x)$ ($d'_C(x,y) \ge 0$ a $d'_C(x,y) \ne 0$ pro $y \notin T_C(x)$ podle věty 230), platí

$$\lambda(y) \triangleq \frac{\|z\|\|y\|}{d'_C(x,y)} \ge 0.$$

Použitím Schwarzovy nerovnosti dostaneme

$$z^T y \le \|z\|\|y\| = \lambda(y) d'_C(x,y).$$

Dokázali jsme tedy, že pro libovolný vektor $y \in R^n$ existuje $\lambda(y) \ge 0$ tak, že $z^T y \le \lambda(y) d'_C(x,y)$. Odtud plyne, že $z \in \overline{\bigcup_{\lambda\ge 0} \lambda\partial d_C(x)}$, takže

$$N_C(x) \subset \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)}$$

. □

15.3 Lipschitzovské funkce

Definice 79 *Řekneme, že funkce $F : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ (s konstantou L), jestliže existuje $\varepsilon > 0$ tak, že platí*

$$|F(x_2) - F(x_1)| \leq L\|x_2 - x_1\|, \tag{531}$$

pokud $x_1 \in B(x, \varepsilon)$ a $x_2 \in B(x, \varepsilon)$. 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 80 *Zobecněnou (Clarkovu) směrovou derivaci funkce $F : R^n \to R$ v bodě $x \in R^n$ ve směru $h \in R^n$ definujeme předpisem*

$$F^0(x, h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + th) - F(y)}{t}. \tag{532}$$

Poznámka 324 Je-li $F^0(x, h)$ zobecněnou směrovou derivací funkce F ve smyslu Definice 80, 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 232 *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

$$F(y + th_2) \leq F(y + th_1) + Lt\|h_2 - h_1\|$$

(viz (531)), takže

$$\begin{aligned}
F^0(x, h_2) &= \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + th_2) - F(y)}{t} \\
&\leq \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{F(y + th_1) - F(y)}{t} + L\|h_2 - h_1\| \\
&= F^0(x, h_1) + L\|h_2 - h_1\|,
\end{aligned}$$

neboli

$$F^0(x, h_2) - F^0(x, h_1) \leq L\|h_2 - h_1\|.$$

Protože nezáleží na pořadí vektorů h_1 a h_2, platí

$$|F^0(x, h_2) - F^0(x, h_1)| \leq L\|h_2 - h_1\|.$$

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

$$\begin{aligned}
F^0(x_i, h_i) &\leq \frac{F(y_i + t_i h_i) - F(y_i)}{t_i} + \frac{1}{i} \\
&= \frac{F(y_i + t_i h) - F(y_i)}{t_i} + \frac{F(y_i + t_i h_i) - F(y_i + t_i h)}{t_i} + \frac{1}{i}.
\end{aligned}$$

Z lipschitzovské spojitosti funkce F plyne

$$\left\| \frac{F(y_i + t_i h_i) - F(y_i + t_i h)}{t_i} \right\| \leq L\|h_i - h\|$$

pro dostatečně velké indexy i, takže

$$\limsup_{i \to \infty} F^0(x_i, h_i) \leq F^0(x, h) + \lim_{i \to \infty} \left(L\|h_i - h\| + \frac{1}{i} \right) = F^0(x, h).$$

(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 325 Podle definice 80 platí $F^0(x,0) = 0$, takže podle věty 232 (a) dostaneme

$$|F^0(x,h)| = |F^0(x,h) - F^0(x,0)| \le L\|h\|.$$

Definice 81 *Nechť funkce $F : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$. Pak množinu*

$$\partial F(x) = \{g \in R^n : F^0(x,h) \ge g^T h \quad \forall h \in R^n\}$$

nazveme subdiferenciálem funkce F v bodě x. Elementy $g \in \partial F(x)$ budeme nazývat subgradienty funkce F v bodě x.

Věta 233 *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 $\|g\| \le L$ $\forall g \in \partial F(x)$.

(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 232 (a) je funkce $F^0(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (tvrzení 10) existuje vektor $g \in R^n$ takový, že

$$g^T h \le F^0(x,h) \quad \forall h \in R^n,$$

takže subdiferenciál $\partial F(x)$ je neprázdný. Nechť $g_1 \in \partial F(x)$, $g_2 \in \partial F(x)$ a $\lambda_1 \ge 0$, $\lambda_2 \ge 0$, $\lambda_1 + \lambda_2 = 1$. Pak pro libovolný vektor $h \in R^n$ platí

$$(\lambda_1 g_1 + \lambda_2 g_2)^T h = \lambda_1 g_1^T h + \lambda_2 g_2^T h \le \lambda_1 F^0(x,h) + \lambda_2 F^0(x,h) = F^0(x,h),$$

takže subdiferenciál $\partial F(x)$ je konvexní. Nechť $g \in \partial F(x)$. Pak podle definice 81 a poznámky 325 platí

$$\|g\|^2 = g^T g \le F^0(x,g) \le L\|g\|,$$

čili $\|g\| \le L$, takže subdiferenciál $\partial F(x)$ je omezený. Nechť $g_i \in \partial F(x)$ a $g_i \to g$. Pak platí

$$g^T h = \lim_{i\to\infty} g_i^T h \le F^0(x,h),$$

takže $g \in \partial F(x)$ a subdiferenciál $\partial F(x)$ je uzavřený.
(b) Podle definice platí

$$F^0(x,h) \ge \max\left\{g^T h : g \in \partial F(x)\right\}.$$

Předpokládejme, že pro nějaký vektor $\overline{h} \in R^n$ platí

$$F^0(x, \overline{h}) > \max\left\{g^T \overline{h} : g \in \partial F(x)\right\}. \tag{533}$$

Uvažujme lineární funkci $l(\lambda\overline{h}) = \lambda f^0(x, \overline{h})$ definovanou na jednorozměrném podprostoru $\{\lambda\overline{h} : \lambda \in R\} \subset R^n$. Jelikož je $f^0(x, \cdot)$ 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 $F'(x, \overline{h}) = \overline{g}^T\overline{h}$, což je ve sporu s předpokladem (533).
(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 232 (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 81 právě tehdy, jestliže $(-F)^0(x, h) \geq g^T h \ \forall h \in R^n$, což je podle věty 232 (c) ekvivalentní $F^0(x, -h) \geq g^T h \ \forall h \in R^n$, což podle definice 81 znamená $-g \in \partial F(x)$. Tedy $\partial(-F)(x) = -\partial F(x)$. □

Poznámka 326 Porovnáme-li větu 233 (b) s poznámkou 316 vidíme, že zobecněná směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$F^0(x, h) = \delta_{\partial F(x)}(h).$$

Věta 234 *Nechť funkce $F : R^n \to R$ je spojitě diferencovatelná v bodě $x \in R^n$. Pak F je lipschitzovská v okolí bodu x a platí*

$$\partial F(x) = \{\nabla F(x)\}. \tag{534}$$

Důkaz Je-li F spojitě diferencovatelná v bodě x, pak gradient $\nabla F(x)$ existuje a je omezený v okolí bodu x. Existují tedy čísla $\varepsilon > 0$ a $L > 0$ tak, že $\|\nabla F(y)\| \leq L \ \forall y \in B(x, \varepsilon)$. Nechť $x_1 \in B(x, \varepsilon)$ a $x_2 \in B(x, \varepsilon)$. Pak podle věty o střední hodnotě platí

$$F(x_2) - F(x_1) = (\nabla F(y))^T (x_2 - x_1),$$

kde $y \in (x_1, x_2) \subset B(x, \varepsilon)$. 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 $B(x, \varepsilon)$. Ze spojité diferencovatelnosti funkce F v bodě x plyne, že $F'(y, h) = (\nabla F(y))^T h$ pokud $y \in B(x, \varepsilon)$. Předpokládejme, že $x_i \in B(x, \varepsilon)$ 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 $F^0(x, h) = (\nabla F(x))^T h \ \forall h \in R^n$, takže $\nabla F(x) \in \partial F(x)$. Předpokládejme, 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 327 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 $F^0(x,h) \geq F'(x,h) = (\nabla F(x))^T h \ \forall h \in R^n$). Rovnost (534) lze dokázat pouze v případě spojité diferencovatelnosti.

Věta 235 *Nechť funkce $F : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak platí*

(a) $F^0(x,h) = F'(x,h) \ \forall h \in R^n$.

(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 81. Abychom dokázali (a), stačí dokázat, že $F^0(x,h) \leq F'(x,h)$, neboť obrácenou nerovnost dostaneme ihned z definice 80 (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 324). 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 225 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 72 (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 328 Věta 235 říká, že v případě konvexních funkcí je zobecněná směrová derivace totožná s obyčejnou směrovou derivací a subdiferenciál podle definice 81 splývá se subdiferenciálem podle definice 78.

Rovnost $F^0(x,h) = F'(x,h)$ není obecně splněna, ani když $F'(x,h)$ existuje (příkladem jsou nehladké konkávní funkce). Tato rovnost však přináší teoretické výhody, takže je účelné vyšetřovat funkce, pro než platí.

Definice 82 *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 236 *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 234 (neboť $F^0(x,h) = \max_{g\in\partial F(x)} g^T h = (\nabla F(x))^T h = F'(x,h)\ \forall h \in R^n$). Konvexní funkce je regulární podle věty 235.
(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 226 (b) a věty 232 (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 77 plyne, že $(F_1+F_2)'$ existuje a platí $(F_1+F_2)' = F_1' + F_2'$. Podle definice 80 platí $(F_1+F_2)^0 \geq (F_1+F_2)'$. Z druhé strany

$$\begin{aligned}
(F_1+F_2)^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{(F_1+F_2)(y+th)-(F_1+F_2)(y)}{t}\\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_1(y+th)+F_2(y+th)-F_1(y)-F_2(y)}{t}\\
&\leq \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_1(y+th)-F_1(y)}{t} + \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F_2(y+th)-F_2(y)}{t}\\
&= F_1^0(x,h) + F_2^0(x,h),
\end{aligned}$$

takže

$$(F_1+F_2)' = F_1' + F_2' = F_1^0 + F_2^0 \geq (F_1+F_2)^0,$$

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 80 platí $F^0(x,h) \geq F'(x,h)$. Z druhé strany

$$\begin{aligned}
F^0(x,h) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{F(y+th)-F(y)}{t}\\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{\max(F_1(y+th),F_2(y+th))-\max(F_1(y),F_2(y))}{t}\\
&\leq \limsup_{\substack{y\to x\\ t\downarrow 0}} \max\left(\frac{F_1(y+th)-F_1(y)}{t}, \frac{F_2(y+th)-F_2(y)}{t}\right)\\
&\leq \max(F_1^0(x,h), F_2^0(x,h)).
\end{aligned}$$

Platí tedy

$$F'(x,h) = \max(F_1'(x,h), F_2'(x,h)) = \max(F_1^0(x,h), F_2^0(x,h)) \geq F^0(x,h),$$

což dohromady s předchozí nerovností dává $F^0(x,h) = F'(x,h)$. □

Věta 237 *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 81 platí $\partial(\lambda_1 F_1)(x) = \lambda_1 \partial F_1(x)$. V opačném případě s použitím věty 233 (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 236 (a)). Použijeme-li poznámku 326 a větu 207, dostaneme

$$\delta_{\partial(F_1+F_2)(x)}(h) \leq \delta_{\partial F_1(x)}(h) + \delta_{\partial F_2(x)}(h) = \delta_{\partial F_1(x)+\partial F_2(x)}(h) \tag{535}$$

$\forall h \in R^n$, takže podle věty 206 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 236 (a) platí $(F_1 + F_2)^0 = (F_1 + F_2)' = F_1' + F_2' = F_1^0 + F_2^0$, takže v (535) a tedy i v (b) nastane rovnost. Je-li funkce F_1 spojitě diferencovatelná v bodě x, pak podle definice 80 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 329 *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 a koeficienty λ_i nezáporné nebo jsou-li všechny funkce F_i až na jednu spojitě diferencovatelné.

Věta 238 *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 81 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 233 (d) platí $0 \in \partial F(x)$. □

Pro další analýzu nehladkých funkcí je důležitá věta o střední hodnotě. Abychom zjednodušili symboliku, budeme pro libovolný vektor $v \in R^n$ používat označení

$$(\partial F(z))^T v = \left\{ g^T v : g \in \partial F(z) \right\}.$$

Věta 239 *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*

$$F(y) - F(x) \in (\partial F(z))^T (y-x).$$

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

$$\partial\varphi(\lambda) \subset (\partial F(x + \lambda(y-x)))^T (y-x). \tag{536}$$

Podle věty 233 (a) jsou množiny na obou stranách této inkluze intervaly. Podle věty 204 a 206 stačí dokázat, že

$$\delta_{\partial\varphi(\lambda)}(\beta) \leq \delta_{(\partial F(x+\lambda(y-x)))^T(y-x)}(\beta) \tag{537}$$

pro $\beta = 1$ a $\beta = -1$. Podle definice 80 a věty 233 (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 316 a poznámky 326 dává (537) a tedy i (536). 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 238 dává $0 \in \partial\psi(\lambda^*)$. Použijeme-li větu 237 a vztah (536), 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 $F(y)-F(x) \in (\partial F(z))^T(y-x)$ pro $z = x + \lambda^*(y-x) \in (x,y)$. □

Je-li funkce $F : R^n \to R$ lokálně lipschitzovská v otevřené množině $\Omega \subset R^n$, je podle Rademacherovy věty (tvrzení 11) 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 240 *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

$$F^0(x,h) \le \limsup_{\substack{y\to x \\ y\notin\Omega_F}} \nabla^T F(y)h. \tag{538}$$

Zvolme $h \in R^n$, $\varepsilon > 0$ libovolně a označme α pravou stranu v (538). Z definice horní limity (limes superior) plyne existence čísla $\delta > 0$ takového, že $\nabla^T F(y)h \le \alpha + \varepsilon$ pokud $y \in B(x,\delta)$ a $y \notin \Omega_F$. Bez újmy na obecnosti můžeme předpokládat, že F je lipschitzovská v $B(x,\delta)$, takže podle Rademacherovy věty má $B(x,\delta) \cap \Omega_F$ Lebesgueovu míru nula. Označme

$$L_y = \{y + th : 0 < t < \delta/(2\|h\|)\},$$

takže $L_y \subset B(x,\delta)$, pokud $y \in B(x,\delta/2)$. Z teorie Lebesgueovy míry plyne, že pro skoro všechny body $y \in B(x,\delta/2)$ má množina $L_y \cap \Omega_F$ Lebesgueovu míru nula. Pro skoro všechny body $y \in B(x,\delta/2)$ 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 $\nabla^T F(y+\vartheta h)h \le \alpha + \varepsilon$ kdykoliv $\nabla F(y+\vartheta h)$ existuje, můžeme tento integrál majorizovat, takže

$$F(y+th) - F(y) \le t(\alpha+\varepsilon). \tag{539}$$

Tato nerovnost platí pro skoro všechny body $y \in B(x,\delta/2)$ a pro všechna $t \in (0,\delta/(2\|h\|))$. Jelikož funkce F je spojitá, musí (539) platit pro všechny body $y \in B(x,\delta/2)$ a pro všechna $t \in (0,\delta/(2\|h\|))$, což podle definice 80 dává

$$F^0(x,h) \le \alpha + \varepsilon.$$

Jelikož $\varepsilon > 0$ je libovolné, dostáváme (538).
(b) Protože Ω_F má Lebesgueovu míru nula, existuje alespoň jedna posloupnost $y_i \to x$, $y_i \notin \Omega_F$. Podle poznámky 327 platí $\nabla F(y_i) \in \partial F(y_i)$, takže podle věty 233 (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 233 (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 326 a nerovnost (538), 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 206 platí $\partial F(x) \subset$ conv $\partial_B F(x)$. □

15.4 Lipschitzovská zobrazení

Přístup použitý ve větě 240 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

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f_1(x)}{\partial x_1}, & \dots, & \frac{\partial f_1(x)}{\partial x_n} \\ \dots & \dots & \dots \\ \frac{\partial f_m(x)}{\partial x_1}, & \dots, & \frac{\partial f_m(x)}{\partial x_n} \end{bmatrix},$$

která má opět Lebesgueovu míru nula.

Definice 83 *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak množinu*

$$\partial f(x) = \text{conv } \partial_B f(x),$$

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 330 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 331 Je-li zobrazení $f : R^n \to R^m$ diferencovatelné v bodě $x \in R^n$, pak přímo z definice 83 plyne, že

$$\nabla f(x) \in \partial f(x)$$

(stačí zvolit posloupnost $x_i = x \to x \notin \Omega_f$).

Věta 241 *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 $\partial f_i(x)$, $1 \leq i \leq m$, 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 $\|J\| \leq L \ \forall J \in \partial f(x)$.

(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 240.
(b) Kompaktnost plyne bezprostředně z (a) a z věty 233 (a). Konvexita plyne přímo z definice 83. 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 240). Nerovnost $\|J\| \leq L$ plyne z definice 83 a z toho, že $\|\nabla f(x_i)\| \leq L$ 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 $x_i \in B(x, 1/(2i))$ (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

$$J_i \notin \partial f(x) + B(0, \varepsilon).$$

Protože množina $\partial f(x) + B(0, \varepsilon)$ je konvexní, nemůže platit $\partial_B f(x_i) \subset \partial f(x) + B(0, \varepsilon)$ (v opačném případě by muselo platit $J_i \in \operatorname{conv} \partial_B f(x_i) \subset \partial f(x) + B(0, \varepsilon)$). Existuje tedy matice $\overline{J}_i \in \partial_B f(x_i)$ taková, že $\overline{J}_i \notin \partial f(x) + B(0, \varepsilon)$. Podle definice 83 musí existovat bod $y_i \in B(x_i, 1/(2i)) \subset B(x, 1/i)$ takový, že $\|\nabla f(y_i) - \overline{J}_i\| < \varepsilon/2$, takže

$$\nabla f(y_i) \notin \partial f(x) + B(0, \varepsilon/2). \tag{540}$$

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 $y_i \in B(x, 1/i)$), $y_i \notin \Omega_f$ (neboť $\nabla f(y_i)$ existuje) a $\nabla f(y_i) \to \overline{J}$. Podle definice 83 tedy platí $\overline{J} \in \partial f(x)$, což je ve sporu s (540). □

Lemma 73 *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 234 a definici 79). 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 241 (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

$$\partial f([x, y]) = \operatorname{conv} \bigcup_{z \in [x,y]} \partial f(z). \tag{541}$$

Lemma 74 *Množina $\partial f([x, y])$ je konvexní a kompaktní.*

Důkaz Konvexita plyne bezprostředně z (541). Abychom dokázali kompaktnost, stačí podle věty 195 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 241 (c) dostaneme $J \in \partial f(z) \subset \bigcup_{z\in[x,y]} \partial f(z)$. □

Věta 242 *Nechť zobrazení $f : R^n \to R^m$ je lokálně lipschitzovské na otevřené množině Ω obsahující úsečku $[x,y]$. Pak platí*

$$f(y) - f(x) \in \partial f\left([x,y]\right)(y-x). \tag{542}$$

Důkaz Podle lemmatu 73 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 239, podle které pro libovolný vektor $v \in R^m$ existuje bod $z \in (x,y)$ takový, že

$$v^T(f(y) - f(x)) \in \partial(v^T f)(z)(y-x) = v^T \partial f(z)(y-x). \tag{543}$$

Vztah (542) dokážeme sporem. Předpokládejme, že $f(y) - f(x) \notin \partial f\left([x,y]\right)(y-x)$. Jelikož množina na pravé straně je podle lemmatu 74 konvexní a kompaktní, musí podle věty 200 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 (543), neboť podle (543) existuje prvek $J \in \partial f(z) \subset \partial f\left([x,y]\right)$ takový, že $v^T(f(y) - f(x)) = v^T J(y-x)$. □

Věta 243 *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*

$$\partial F(x) \subset \operatorname{conv}\left(\partial f(x)\right)^T \partial\varphi(f(x)) \stackrel{\Delta}{=} \operatorname{conv}\left\{J^T v : J \in \partial f(x), v \in \partial\varphi(f(x))\right\}, \tag{544}$$

přičemž rovnost nastává zejména v těchto případech

(a) Funkce φ je spojitě diferencovatelná v bodě $f(x)$. V tomto případě platí

$$\partial F(x) = \left(\partial f(x)\right)^T \nabla\varphi(f(x)). \tag{545}$$

(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

$$\partial F(x) = \left(\nabla f(x)\right)^T \partial\varphi(f(x)). \tag{546}$$

(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 79). Označme S množinu na pravé straně (544). Abychom dokázali inkluzi $\partial F(x) \subset S$, použijeme větu 206 a poznámku 326. Jelikož podle věty 204 pro libovolný vektor $h \in R^n$ platí

$$\delta_S(h) = \max\left\{v^T J h : J \in \partial f(x), v \in \partial\varphi(f(x))\right\},$$

stačí podle věty 206 a poznámky 326 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

$$\delta_{\partial F}(h) = F^0(x,h) \leq v^T J h. \tag{547}$$

Podle poznámky 324 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 239 (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 242 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 (541) a podle věty 194 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 $J_i^k \in \partial f(y_i^k)$, $y_i^k \in [x_i, x_i + t_i h]$, $\lambda_i^k \geq 0$, $k \in [1, m+1]$, $\lambda_i^1 + \ldots + \lambda_i^{m+1} = 1$. Z tohoto důvodu musí alespoň pro jeden index $k \in [1, m+1]$ platit

$$\frac{F(x_i + t_i h) - F(x_i)}{t_i} \leq v_i^T J_i^k h. \tag{548}$$

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 233 (c) a věty 241 (c) platí $v \in \partial\varphi(f(x))$ a $J \in \partial f(x)$, takže z (548) plyne (547). Nyní vyšetříme speciální případy:

(a) Tento případ je tvrzením lemmatu 73.

(b) Je-li zobrazení f spojitě diferencovatelné, můžeme množinu S zapsat ve tvaru $S = (\nabla f(x))^T \partial\varphi(f(x))$ (protože množina $\partial f(x) = \{\nabla f(x)\}$ je jednoprvková nemusíme používat její konvexní obal). Použijeme-li definici 69, poznámku 326 a regularitu funkce φ (definice 82), můžeme psát

$$\begin{aligned}
\delta_S(h) &= \max_{v\in\partial\varphi(f(x))} v^T \nabla f(x) h = \max_{v\in\partial\varphi(f(x))} v^T f'(x,h) \\
&= \varphi^0(f(x), f'(x,h)) = \varphi'(f(x), f'(x,h)) \\
&= \lim_{t\downarrow 0} \frac{\varphi(f(x) + t f'(x,h)) - \varphi(f(x))}{t} = \lim_{t\downarrow 0} \left(\frac{\varphi(f(x+th)) - \varphi(f(x))}{t} + T(t) \right),
\end{aligned}$$

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 $\delta_S(h) = F'(x,h) \le F^0(x,h)$, což podle věty 206 dává $S \subset \partial F(x)$, takže z (544) plyne $\partial F(x) = S$. Z nerovnosti $F^0(x,h) \le \delta_S(h) = F'(x,h) \le F^0(x,h)$ pak plyne regularita funkce F v bodě x.

(c) Označme

$$S' = \operatorname{conv} \left\{ \sum_{i=1}^m u_i v_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)) \right\}.$$

Podle (544) platí $\partial F(x) \subset S$ a podle věty 241 (a) platí $S \subset S'$, takže $\partial F(x) \subset S'$. 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

$$\begin{aligned}
\delta_{S'}(h) &= \max \left\{ \sum_{i=1}^m v_i u_i^T h : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)) \right\} \\
&\le \max \left\{ \sum_{i=1}^m v_i \max_{u_i \in \partial f_i(x)} u_i^T h : v \in \partial\varphi(f(x)) \right\} \\
&= \max \left\{ \sum_{i=1}^m v_i f_i^0(x,h) : v \in \partial\varphi(f(x)) \right\} \\
&= \max \left\{ \sum_{i=1}^m v_i f_i'(x,h) : v \in \partial\varphi(f(x)) \right\} \\
&= \varphi^0(f(x), f'(x,h)) = \varphi'(f(x), f'(x,h)).
\end{aligned}$$

Konec důkazu je již stejný jako konec důkazu tvrzení (b). Dostaneme $\delta_{S'}(h) = F'(x,h) \le F^0(x,h)$, což podle věty 206 dává $S' \subset \partial F(x)$, takže z $\partial F(x) \subset S \subset S'$ plyne $\partial F(x) = S = S'$. Z nerovnosti $F^0(x,h) \le \delta_S(h) \le \delta_{S'}(h) = F'(x,h) \le F^0(x,h)$ pak plyne regularita funkce F v bodě x. □

Důsledek 26 *Jsou-li splněny předpoklady věty 243, platí*

$$\partial F(x) \subset \operatorname{conv} \left\{ \sum_{i=1}^m u_i v_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x)), \right\} \tag{549}$$

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 243 a některé úvahy (například $S \subset S'$) z jejího důkazu. □

Důsledek 27 *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 $f_1(x) \geq 0$, $f_2(x) \geq 0$. 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 234) 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 243 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 26 (c). □

Důsledek 28 *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*

$$\partial F(x) = \frac{1}{2} \partial(f^T f)(x) = (\partial(f(x))^T f(x) = \left\{ J^T f(x) : J \in \partial(f(x)) \right\}. \tag{550}$$

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 234) lipschitzovská v okolí libovolného bodu $u \in R^m$, přičemž platí $\nabla \varphi(u) = u$. Podle věty 243 (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 244 *Nechť funkce $f_i : R^n \to R$, $1 \leq i \leq m$, jsou lipschitzovské v okolí bodu $x \in R^n$. Pak funkce*

$$F(x) = \max_{1 \leq i \leq m} f_i(x)$$

je lipschitzovská v okolí bodu x a platí

$$\partial F(x) \subset \operatorname{conv} \left\{ \partial f_i(x) : i \in I(x) \right\}, \tag{551}$$

kde $I(x) = \{i \in \{1, \ldots, m\} : f_i(x) = F(x)\}$. Jsou-li funkce f_i, $1 \leq i \leq m$, regulární v bodě x, je funkce F regulární v bodě x a v (551) platí rovnost.

Důkaz Definujme funkci $\varphi : R^m \to R$ předpisem $\varphi(u) = \max(u_1, \ldots, u_m)$. 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 225. Nechť $I(u) = \{i \in \{1, \ldots, m\} : u_i = \varphi(u)\}$. 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 81 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 $v_i \ge 0\ \forall i \in I(u)$ 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 $v_i \ge 0\ \forall i \in I(u)$ 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

$$\partial\varphi(u) = \left\{v \in R^n : v_i \ge 0, \sum_{i\in I(u)} v_i = 1, \sum_{i\notin I(u)} v_i = 0\right\}.$$

Podle důsledku 26 pak platí

$$\begin{aligned}\partial F(x) &\subset \operatorname{conv}\left\{\sum_{i=1}^m v_i u_i : u_i \in \partial f_i(x), v \in \partial\varphi(f(x))\right\}\\ &= \operatorname{conv}\left\{\sum_{i\in I(u)} v_i \partial f_i(x) : v_i \ge 0, \sum_{i\in I(u)} v_i = 1\right\}\\ &= \operatorname{conv}\left\{\partial f_i(x), i \in I(u)\right\}.\end{aligned}$$

Funkce φ je konvexní, takže je podle věty 236 regulární. Jsou-li funkce f_i, $1 \le i \le m$, regulární, je podle věty 236 i funkce F regulární a jelikož $v_i \ge 0$, $1 \le i \le m$, platí v (551) rovnost. □

15.5 Polohladká zobrazení

Definice 84 *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*

$$\lim_{\substack{J\in\partial f(x+th)\\ t\downarrow 0}} Jh \tag{552}$$

(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

$$\lim_{\substack{J \in \partial f(x+th') \\ h' \to h, t \downarrow 0}} Jh' \tag{553}$$

(nezávislá na volbě $J \in \partial f(x+th')$), řekneme, že zobrazení f je polohladké v bodě x.

Poznámka 332 Jelikož $\partial f(x)$ je množinové zobrazení, mohlo by se zdát, že existence limity (553) je výjimečná. V dalším textu však ukážeme (poznámka 336), že polohladkost je vlastnost převážné většiny zajímavých lokálně lipschitzovských zobrazení.

Poznámka 333 Z definice 84 plyne, že každé polohladké zobrazení je slabě polohladké. Slabá polohladkost se však nezachovává při skládání funkcí a také věta 250 vyžaduje platnost vztahu (553).

Věta 245 *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 $[x, x+t_i h]$. Použijeme-li větu 242, dostaneme

$$\frac{f(x+t_i h) - f(x)}{t_i} \in \partial f\left([x, x+t_i h]\right) h = \left(\operatorname{conv} \bigcup_{t \in [0,t_i]} \partial f(x+th) \right) h = \operatorname{conv} \left(\bigcup_{t \in [0,t_i]} \partial f(x+th) h \right) \subset R^m$$

Podle věty 194 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

$$\frac{f(x+t_i h) - f(x)}{t_i} = \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

kde $0 \le \lambda_i^k \le 1$ a $\lambda_1^k + \ldots + \lambda_{m+1}^k = 1$. 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 (552) platí

$$\begin{aligned} \lim_{i \to \infty} \frac{f(x+t_i h) - f(x)}{t_i} &= \lim_{i \to \infty} \left(\sum_{k=1}^{m+1} \lambda_i^k J_i^k h \right) = \sum_{k=1}^{m+1} \left(\lim_{i \to \infty} \lambda_i^k \right) \left(\lim_{i \to \infty} J_i^k h \right) \\ &= \left(\sum_{k=1}^{m+1} \lambda^k \right) \lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh = \lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh, \end{aligned}$$

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 334 Zobrazení $f'(x,\cdot) : R^n \to R^m$ (směrová derivace) vystupující ve větě 245 je pozitivně homogení a lipschitzovské (poznámka 321). Nemusí však být subaditivní jako v případě konvexních funkcí.

Poznámka 335 Podle věty 245, 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 330.

Věta 246 *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 (553) existuje právě tehdy, existují-li pro $1 \le i \le m$ limity

$$\lim_{\substack{J \in \partial f(x+th') \\ h' \to h, t \downarrow 0}} e_i^T J h'.$$

(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 247 *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 248 *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 $B(x,\varepsilon)$, $x+th' \in B(x,\varepsilon)$ a $g \in \partial F(x+th')$. Pak podle věty 227 (d) platí

$$F(x) - F(x+th') \ge g^T(x - (x+th')),$$

neboli

$$\frac{F(x+th') - F(x)}{t} \le g^T h'.$$

Z druhé strany podle definice 78 platí

$$g^T h' \le F'(x+th', h').$$

Jelikož funkce konvexní v okolí bodu $x \in R^n$ je v okolí tohoto bodu lipschitzovská s nějakou konstantou L (věta 225), můžeme psát

$$\lim_{h' \to h,\ t \downarrow 0} \frac{\|F(x+th') - F(x+th)\|}{t} \le \lim_{h' \to h} L\|h' - h\| = 0$$

a jelikož podle věty 226 (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

$$\begin{aligned} F'(x,h) &= \lim_{\substack{h'\to h\\ t\downarrow 0}} \frac{F(x+th')-F(x)}{t} \le \liminf_{\substack{g\in\partial F(x+th')\\ h'\to h,\ t\downarrow 0}} g^T h' \le \limsup_{\substack{g\in\partial F(x+th')\\ h'\to h,\ t\downarrow 0}} g^T h' \\ &\le \limsup_{h'\to h,\ t\downarrow 0} F'(x+th',h') \le F'(x,h), \end{aligned}$$

(poslední nerovnost plyne z věty 226 (c)). Tím je dokázána existence limity (553) (s g^T místo J). □

Věta 249 *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 243 platí $\partial F(x_k) \subset S_k$, kde symbol $S_k \subset R^n$ označuje kompaktní množinu na pravé straně výrazu (544) (s x_k místo x). Nechť

$$\begin{aligned} w_k^- &= (J_k^-)^T v_k^- = \arg\min_{w\in S_k} w^T h, \quad v_k^- \in \partial\varphi(f(x_k)), \quad J_k^- \in \partial f(x_k), \\ w_k^+ &= (J_k^+)^T v_k^+ = \arg\max_{w\in S_k} w^T h, \quad v_k^+ \in \partial\varphi(f(x_k)), \quad J_k^+ \in \partial f(x_k). \end{aligned}$$

Pak pro libovolný vektor $w_k \in \partial F(x_k) \subset S_k$ platí

$$(w_k^-)^T h \le w_k^T h \le (w_k^+)^T h. \tag{554}$$

Jelikož všechny veličiny v těchto vzorcích jsou podle věty 233 (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 233 (c)). Jelikož zobrazení f je polohladké, platí $J^- h = J^+ h = f'(x,h)$, takže s použitím (554) dostaneme

$$(v^-)^T f'(x,h) \le \liminf_{k\to\infty} w_k^T h \le \limsup_{k\to\infty} w_k^T h \le (v^+)^T f'(x,h).$$

Jelikož funkce φ je polohladká a podle poznámky 335 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 84, 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 29 *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 247, věty 248 a věty 249. □

Důsledek 30 *Nechť funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou polohladké v bodě $x \in R^n$ a $\lambda_i \in R$, $1 \le i \le m$. 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 246 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 29 jsou funkce $F_1 = \varphi_1 \circ f$ a $F_2 = \varphi_2 \circ f$ polohladké v bodě x. □

Důsledek 31 *Nechť funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou polohladké v bodě $x \in R^n$ a $f = [f_1, \dots, f_m]^T$. Pak funkce $F = \|f\|$, kde $\|\cdot\|$ je libovolná norma v R^m, je polohladká v bodě x. Speciálně funkce $F_1 = \max_{1\le i\le m}(|f_i|)$ (maximum absolutních hodnot) a $F_2 = \sum_{i=1}^m |f_i|$ (součet absolutních hodnot) jsou polohladké v bodě x. Dále funkce $F_3 = \max_{1\le i\le m}(f_i)$ (bodové maximum) je polohladká v bodě x.*

Důkaz Podle věty 246 je zobrazení $f = [f_1, \dots, f_m]^T$ polohladké v bodě x. Funkce $\varphi(u) = \|u\|$ je konvexní, neboť z vlastností vektorové normy plyne, že pro $0 \le \lambda \le 1$ platí

$$\varphi(\lambda u + (1-\lambda)v) = \|\lambda u + (1-\lambda)v\| \le \lambda\|u\| + (1-\lambda)\|v\|.$$

Funkce $F = \varphi \circ f$ je tedy podle důsledku 29 polohladká. Také funkce $\varphi_3(u) = \max_{1\le i\le m}(u_i)$ je konvexní (důkaz věty 244), takže funkce $F_3 = \varphi_3 \circ f$ je podle důsledku 29 polohladká. □

Důsledek 32 *(Obrácení věty 246). Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ přičemž $f = [f_1, \dots, f_m]^T$. Pak funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou polohladké v bodě x.*

Důkaz Zřejmě $f_i = \varphi_i \circ f$, $1 \le i \le m$, kde funkce $\varphi_i : R^m \to R$, definované předpisem $\varphi_i(u) = e_i^T u = u_i$, jsou spojitě diferencovatelné. Polohladkost funkcí $f_i : R^n \to R$, $1 \le i \le m$, tedy plyne z důsledku 29. □

Důsledek 33 *Lineární kombinace polohladkých zobrazení je polohladké zobrazení. Skalární součin polohladkých zobrazení je polohladká funkce.*

Důkaz Podle důsledku 32 jsou složky polohladkých zobrazení polohladkými funkcemi. Podle důsledku 30 je lineární kombinace polohladkých funkcí polohladkou funkcí, takže podle věty 246 je lineární kombinace polohladkých zobrazení polohladkým zobrazením. Polohladkost skalárního součinu plyne z důsledku 32, důsledku 30 a věty 246. □

Poznámka 336 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 $h_i \to 0$, $h_i \ne 0$ platí $o(\|h_i\|)/\|h_i\| \to 0$.

Věta 250 *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*

$$Jh - f'(x,h) = o(\|h\|) \tag{555}$$

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

$$\lim_{i\to\infty} \frac{J_i h_i - f'(x,h_i)}{\|h_i\|} = 0. \tag{556}$$

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

$$\frac{\|J_i h_i - f'(x,h_i)\|}{\|h_i\|} = \|J_i h_i' - f'(x,h_i')\| \ge \varepsilon \quad \forall i \in N,$$

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 245 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 321 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

$$\|J_i h_i' - f'(x,h)\| \geq 2\varepsilon \quad \forall i \in N \tag{557}$$

(v opačném případě by existovala limita (553) rovnající se $f'(x,h)$, takže zobrazení f by bylo podle definice 84 polohladké). Jelikož směrová derivace je podle poznámky 321 lipschitzovská, platí pro dostatečně velké indexy $\|f'(x,h_i') - f'(x,h)\| \leq \varepsilon$, což spolu s (557) 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 (556) a tudíž ani (555). □

Poznámka 337 Vzhledem k platnosti věty 250 se polohladké zobrazení často definuje jako lokálně lipschitzovské zobrazení, které vyhovuje podmínce (555).

Definice 85 *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*

$$f(x+h) - f(x) - f'(x,h) = o(\|h\|), \tag{558}$$

pokud $h \to 0$ (to znamená, že zobrazení $f'(x,.)$ má stejné aproximační vlastnosti jako Frechetova derivace).

Věta 251 *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

$$\frac{|f(x+h_i) - f(x) - f'(x,h_i)|}{\|h_i\|} = \left| \frac{f(x+t_i h_i') - f(x)}{t_i} - f'(x,h_i') \right| \geq \varepsilon \quad \forall i. \tag{559}$$

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 245 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 (559), neboť funkce $f'(x,\cdot)$ je podle poznámky 321 spojitá. □

Důsledek 34 *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$. Pak platí*

$$f(x+h) - f(x) - Jh = o(\|h\|), \tag{560}$$

pokud $h \to 0$ a $J \in \partial f(x+h)$.

Důkaz Tvrzení plyne bezprostředně z věty 250 a věty 251. □

16 Metody pro řešení soustav nehladkých rovnic

16.1 Newtonova metoda

Nyní se budeme zabývat řešením soustavy rovnic

$$f(x) = 0, \tag{561}$$

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

$$x_{k+1} = x_k + d_k, \tag{562}$$

kde vektor d_k se vybírá tak, aby platilo

$$\omega_k = \frac{\|A_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega \tag{563}$$

a matice A_k se vybírá tak, aby platilo

$$\Delta_k = \|A_k - J_k\| \leq \Delta \tag{564}$$

pro nějaký prvek $J_k \in \partial_B f(x_k)$. Přitom $\omega \geq 0$, $\Delta \geq 0$ a normy v (563) a (564) jsou euklidovské.

Definice 86 *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 83).*

Poznámka 338 V iteračním procesu (562)-(564) 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 252 *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 $y \in B(x, \delta)$.*

Důkaz Nejprve dokážeme existenci čísla $\delta > 0$ a konstanty $c \geq 0$ tak, že všechny Jacobiho matice $\nabla f(z)$ jsou regulární a platí

$$\|(\nabla f(z))^{-1}\| \leq c, \tag{565}$$

pokud $z \in B(x, \delta) \backslash \Omega_f$ (množina Ω_f je uvedena v definici 83). Předpokládejme, že (565) neplatí. Pak musí existovat posloupnost $x_i \to x$, $x_i \in B(x, \delta) \backslash \Omega_f$ 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 241 (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 83 platí $J \in \partial_B f(x)$, což je v rozporu s definicí 86. Nechť nyní $y \in B(x, \delta) \cap \Omega_f$ a $J \in \partial_B f(y)$. Pak existuje číslo $0 < \delta' < \delta$ tak, že $B(y, \delta') \subset B(x, \delta)$ a (565) platí pokud $z \in B(y, \delta') \backslash \Omega_f$. Jelikož podle definice 83 platí

$$J = \lim_{i\to\infty} \nabla f(y_i)$$

pro nějakou posloupnost $y_i \to y$, $y_i \in B(y,\delta')\backslash\Omega_f$, dostaneme z (565) a ze spojité závislosti vlastních čísel na koeficientech matice nerovnost $\|J^{-1}\| \le c$. □

Věta 253 *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 $\Delta > 0$ tak, že pokud $x_1 \in B(x^*,\varepsilon)$, je iterační proces (562)-(564) dobře definován (matice A_k jsou regulární) a posloupnost $\{x_k\}$ konverguje k bodu x^* Q-lineárně. Jestliže navíc platí $\omega_k \to 0$ 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 252. Položme $\Delta = 1/(5c)$ a zvolme $\varepsilon \le \delta$ tak, aby zobrazení f bylo lipschitzovské (s nějakou konstantou L) v $B(x^*,\varepsilon)$ a aby platilo

$$\|f(x) - f(x^*) - J(x-x^*)\| \le \frac{\Delta}{2}\|x-x^*\| \quad \forall J \in \partial_B f(x), \tag{566}$$

pokud $x \in B(x^*,\varepsilon)$ (to je možné vzhledem k (560)). Dále položme $\omega = \Delta/(2L)$. Předpokládejme, že $x_k \in B(x^*,\varepsilon)$ (platí to pro $k = 1$). Pak podle věty 252 platí $\|J_k^{-1}\| \le c$. 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

$$\|A_k^{-1}\| - \|J_k^{-1}\|\|A_k - J_k\|\|A_k^{-1}\| \le \|J_k^{-1}\|,$$

neboli

$$\|A_k^{-1}\| \le \frac{\|J_k^{-1}\|}{1 - \|J_k^{-1}\|\|A_k - J_k\|} \le \frac{c}{1-c\Delta} = \frac{5}{4}c,$$

což podle (562)-(564) a (566) (s využitím vztahu $f(x^*) = 0$) dává

$$\begin{aligned}
\|x_{k+1} - x^*\| &= \|x_k + d_k - x^*\| = \|x_k + A_k^{-1}(A_k d_k + f(x_k) - f(x_k)) - x^*\| \\
&= \|A_k^{-1}(A_k d_k + f(x_k) - (f(x_k) - J_k(x_k - x^*)) + (A_k - J_k)(x_k - x^*))\| \\
&\le \|A_k^{-1}\| \left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \|A_k - J_k\|\|x_k - x^*\| + \omega_k\|f(x_k) - f(x^*)\|\right) \\
&\le \frac{5}{4}c \left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&\le \frac{5}{4}c\left(\frac{1}{2}\Delta + \Delta + \frac{1}{2}\Delta\right)\|x_k - x^*\| = \frac{1}{2}\|x_k - x^*\|.
\end{aligned} \tag{567}$$

Odtud plyne, že $x_{k+1} \in B(x^*,\varepsilon)$, takže můžeme pokračovat stejným způsobem dále. Dokázali jsme tak indukcí, že ve všech iteračních krocích platí $x_{k+1} \in B(x^*,\varepsilon)$ a $\|x_{k+1} - x^*\| \le (1/2)\|x_k - x^*\|$ čili, že posloupnost $\{x_k\}$ konverguje k bodu x^* Q-lineárně. Nechť nyní $\omega_k \to 0$ a $\Delta_k \to 0$. Pak podle (560) a (567) platí

$$\begin{aligned}
\|x_{k+1} - x^*\| &\le \frac{5}{4}c\left(\|f(x_k) - f(x^*) - J_k(x_k - x^*)\|\right. \\
&\quad \left. + \Delta_k\|x_k - x^*\| + \omega_k L\|x_k - x^*\|\right) \\
&= \frac{5}{4}c\left(o\left(\|x_k - x^*\|\right) + o\left(\|x_k - x^*\|\right) + o\left(\|x_k - x^*\|\right)\right) \\
&= o\left(\|x_k - x^*\|\right)
\end{aligned} \tag{568}$$

a posloupnost $\{x_k\}$ konverguje k bodu x^* Q-superlineárně. Jelikož $f(x^*) = 0$, můžeme podle (568) psát

$$\lim_{k\to\infty} \frac{\|f(x_{k+1})\|}{\|x_k - x^*\|} \leq L \lim_{k\to\infty} \frac{\|x_{k+1} - x^*\|}{\|x_k - x^*\|} = 0. \tag{569}$$

S použitím (562)-(564) a (567) 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 (569) 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 253 říká, že nepřesná Newtonova metoda (562)-(564) 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 (562) 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

$$\omega_k = \frac{\|J_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega. \tag{570}$$

Krok 3 (Délka kroku). Nechť $t_k = \varrho^{i_k}$, kde i_k je nejmenší nezáporné celé číslo i vyhovující podmínce

$$F(x_k + \varrho^i d_k) - F(x_k) \leq -2\sigma\varrho^i F(x_k). \tag{571}$$

Krok 4 (Aktualizace). Položíme $x_{k+1} := x_k + t_k d_k$ a $k := k+1$. Přejdeme na Krok 2.

Věta 254 *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 (561).

(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 (570) platí

$$\|d_k\| = \|J_k^{-1}(J_k d_k + f(x_k)) - J_k^{-1} f(x_k)\| \leq (1+\omega)\|J_k^{-1}\|\|f_k\| \leq c(1+\omega)\sqrt{2F(x_1)}. \tag{572}$$

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 28 a podle (570) platí

$$\begin{aligned} F'(x_k, d_k) &= (\nabla F(x_k))^T d_k = f^T(x_k) J_k d_k \\ &= f^T(x_k) f(x_k) + f^T(x_k) J_k d_k - f^T(x_k) f(x_k) \\ &\leq \|f(x_k)\|\|f(x_k) + J_k d_k\| - \|f(x_k)\|^2 \\ &\leq (\omega - 1)\|f(x_k)\|^2 = -2(1-\omega)F(x_k). \end{aligned} \tag{573}$$

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 K množiny všech indexů taková, že $x_k \overset{K}{\to} x^*$. Vyšetříme nyní dva případy.
(1) Předpokládejme nejprve, že $t_k \geq \tau > 0 \ \forall k \in K$. Pak podle (571) platí

$$\begin{aligned} F(x_1) &\geq F(x_1) - \lim_{k\to\infty} F(x_k) = \sum_{k=1}^{\infty}(F(x_k) - F(x_{k+1})) \\ &\geq \sum_{k=1}^{\infty} 2\sigma t_k F(x_k) \geq 2\tau\sigma \sum_{k\in K} F(x_k), \end{aligned}$$

takže nutně $F(x_k) \overset{K}{\to} 0$, což spolu s $x_k \overset{K}{\to} x^*$ dává $F(x^*) = 0$ (neboť funkce F je spojitá).
(2) Předpokládejme nyní, že $t_k \overset{K_1}{\to} 0$ pro nějakou podmnožinu $K_1 \subset K$. Odtud plyne, že $i_k \overset{K_1}{\to} \infty$, takže pro dostatečně velké indexy $k \in K_1$ platí $i_k > 0$ a jelikož (571) 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 $\{\|d_k\|\}_{K_1}$ je podle (572) omezená, má tato posloupnost alespoň jeden hromadný bod d^*. Existuje tedy podmnožina $K_2 \subset K_1$ taková, že $d_k \overset{K_2}{\to} d^*$, což spolu s $x_k \overset{K_2}{\to} x^*$ a $t_k \overset{K_2}{\to} 0$ (takže $x'_k \overset{K_2}{\to} x^*$) v limitě dává

$$(\nabla F(x^*))^T d^* \geq -2\sigma F(x^*).$$

Z druhé strany podle (573) 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 K je indexová množina použitá v části (a) důkazu. Našim cílem je ukázat, že pro dostatečně velké indexy $k \in K$ platí $x_{k+1} = x_k + d_k$, a pak použít indukční postup z důkazu věty 253. Jelikož $x_k \overset{K}{\to} x^*$, $\omega_k \overset{K}{\to} 0$ (a $\Delta_k = 0$), jsou pro dostatečně velké indexy $k \in K$ splněny předpoklady použité v důkazu věty 253 ($x_k \in B(x^*, \varepsilon)$ a $\omega_k \leq 1/(10cL)$), takže pro bod $x_k + d_k$ platí (567) (s $x_k + d_k$ místo x_{k+1}) a

$$\lim_{k \overset{K}{\to} \infty} \frac{\|x_k + d_k - x^*\|}{\|x_k - x^*\|} = 0,$$
$$\lim_{k \overset{K}{\to} \infty} \frac{\|f(x_k + d_k)\|}{\|f(x_k)\|} = 0.$$

Jelikož $\sigma < 1/2$, existuje index $\overline{k} \in K$ takový, že $\|f(x_k + d_k)\| \leq (1-2\sigma)\|f(x_k)\|$, pokud $k \in K$ 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 (571) je splněna s $i_k = 0$. Platí tedy $x_{k+1} = x_k + d_k$ a vzhledem k (567) můžeme množinu K formálně doplnit o index $k+1$. Pokračujeme-li takto pro další hodnoty indexu, vidíme (tak jako v důkazu věty 253), že $x_k \to x^*$ superlineárně. □

Poznámka 339 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 340 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 83 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 32 polohladké, podle důsledku 34 platí

$$\begin{gathered} f_1(x+h) - f_1(x) - e_1^T J h = o(\|h\|), \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ f_n(x+h) - f_1(x) - e_n^T J h = o(\|h\|) \end{gathered}$$

a n je konečné, zůstává klíčový vztah (560) v platnosti i pro $J \in \partial_b f(x)$ a v důkazech věty 253 a věty 254 se v podstatě nic nezmění.

16.2 Aplikace nehladkých rovnic

Definice 87 *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*

$$x_i^* \geq 0, \quad p_i(x^*) \geq 0, \quad x_i^* p_i(x^*) = 0 \tag{574}$$

pro libovolný index $1 \leq i \leq n$.

Úlohu nelineární komplementarity lze snadno převést na řešení ekvivalentní soustavy polohladkých rovnic $f(x) = 0$, kde

$$f(x) = \begin{bmatrix} \psi(x_1, p_1(x)) \\ \ldots\ldots\ldots\ldots \\ \psi(x_n, p_n(x)) \end{bmatrix} \tag{575}$$

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 $u_1 u_2 = 0$. Tuto vlastnost má například Pangova funkce

$$\psi(u) = \min(u_1, u_2),$$

která je polohladká podle důsledku 31 (neboť $\min(u_1, u_2) = -\max(-u_1, -u_2)$). Nevýhodou Pangovy funkce je to, že není zaručena spojitá diferencovatelnost zobrazení $F = (1/2) f^T f$, které se používá při výběru délky kroku. Výhodnější vlastnosti má Fischerova-Burmeisterova funkce

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2). \tag{576}$$

Lemma 75 *Funkce $\psi : R^2 \to R$ definovaná vztahem (576) je spojitě diferencovatelná v $R^2 \backslash \{0\}$ a polohladká v bodě 0, přičemž $\partial_B \psi(0) = S(-e, 1)$ a $\partial \psi(0) = \overline{B(-e, 1)}$, kde $e = [1, 1]^T$ ($S(u, \varepsilon)$ je kružnice a $\overline{B(u, \varepsilon)} =$ conv $S(u, \varepsilon)$ kruh se středem u a poloměrem ε). 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

$$\nabla \psi(u) = \begin{bmatrix} \frac{u_1}{\sqrt{u_1^2 + u_2^2}} - 1 \\ \frac{u_2}{\sqrt{u_1^2 + u_2^2}} - 1 \end{bmatrix}. \tag{577}$$

Polohladkost funkce ψ v bodě 0 plyne z věty 248, 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

$$\partial \psi(0) = \text{conv } \partial_B \psi(0) = \text{conv } S(-e, 1) = \overline{B(-e, 1)}.$$

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

$$\psi(u) = \sqrt{u_1^2 + u_2^2} - (u_1 + u_2) < \sqrt{u_1^2 + 2u_1 u_2 + u_2^2} - (u_1 + u_2) = 0.$$

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

$$\begin{aligned}\left|\frac{\partial\overline{\psi}(u)}{\partial u_1}\right| &= \left|\frac{u_1}{\sqrt{u_1^2 + u_2^2}}(u_1 + u_2) + \sqrt{u_1^2 + u_2^2}\right| \\ &\leq \frac{|u_1|}{\sqrt{u_1^2 + u_2^2}}|u_1 + u_2| + \sqrt{u_1^2 + u_2^2} \leq |u_1 + u_2| + \sqrt{u_1^2 + u_2^2}\end{aligned}$$

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 255 *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 (575), kde $\psi : R^2 \to R$ je funkce definovaná předpisem (576). 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

$$\partial_B f_i(x) = \nabla f_i(x) = \left(\frac{x_i}{\sqrt{x_i^2 + p_i^2(x)}} - 1\right) e_i + \left(\frac{p_i(x)}{\sqrt{x_i^2 + p_i^2(x)}} - 1\right) \nabla p_i(x), \tag{578}$$

pokud $x_i^2 + p_i^2(x) \neq 0$ a

$$\partial_B f_i(x) = \bigcup_{\varphi\in[0,2\pi]} [(\cos\varphi - 1)e_i + (\sin\varphi - 1)\nabla p_i(x)], \tag{579}$$

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 246 a věty 249, neboť $f_i(x) = \psi(x_i, p_i(x))$, funkce ψ je polohladká podle lemmatu 75 a zobrazení p je spojitě diferencovatelné.
(b) Podle lemmatu 75 je funkce $\psi(x_i, p_i)$ spojitě diferencovatelná, pokud $x_i^2 + p_i^2 \neq 0$. Vztah (578) plyne z (577) s použitím pravidla pro derivování složené funkce. V případě, že $x_i^2 + p_i^2 = 0$, můžeme použít stejný limitní proces jako v lemmatu 75, 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 (579).
(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 75 je druhá mocnina funkce ψ spojitě diferencovatelná, takže i funkce F je spojitě diferencovatelná. □

Věta 255 naznačuje jednu z možností jak řešit úlohy nelineární komplementarity. Úloha nelineární komplementarity se převede na ekvivalentní soustavu nehladkých rovnic (575), které se řeší pomocí Algoritmu 4.1. Podle poznámky 340 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 (578)-(579). 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

$$\nabla F(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0, \tag{580}$$

$$\left.\begin{array}{ll} -c_i(x) \geq 0, & \lambda_i \geq 0, \\ \lambda_i c_i(x) = 0, & 1 \leq i \leq m \end{array}\right\} \tag{581}$$

(tvrzení 4). Podmínky (581) jsou v podstatě podmínkami nelineární komplementarity (574). Můžeme tedy sestavit soustavu $n + m$ nehladkých rovnic

$$f(x, \lambda) \triangleq \left[\begin{array}{l} \nabla F(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) \\ \psi(\lambda_1, -c_1(x)) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \psi(\lambda_m, -c_m(x)) \end{array}\right] = 0, \tag{582}$$

kde ψ je Fischerova-Burmeisterova funkce (576). 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 (582) 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

$$F_j = F(y_j), \quad g_j \in \partial F(y_j), \tag{583}$$

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

$$F_j^k = F_j + g_j^T(x_k - y_j), \tag{584}$$
$$\alpha_j^k = F(x_k) - F_j^k \tag{585}$$

pro $j \in \mathcal{J}_k$. Tato po částech lineární funkce je v konvexním případě majorizována funkcí F.

Věta 256 *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 227 (d) $F(x) \geq F_j + g_j^T(x - y_j)$ $\forall j \in \mathcal{J}_k$, takže podle (584) dostaneme $F(x_k) \geq F_j^k$, což podle (585) 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 256 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

$$\|x_k - y_j\| \leq \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \stackrel{\Delta}{=} s_j^k \tag{586}$$

a čísla α_j^k se určují podle vzorce

$$\alpha_j^k = \max\left(|F(x_k) - F_j^k|, \gamma(s_j^k)^\nu\right), \quad j \in \mathcal{J}_k. \tag{587}$$

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. Podrobnou diskusi o těchto metodách je možné nalézt v pracích [**?**], [**?**], [**?**]. 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. Podrobnější diskusi o konstrukci svazků lze nalézt v práci [**?**].

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 (594), (600), (601) a transformační vztahy (605)). 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

$$F_Q^k(x) = \frac{1}{2}(x-x_k)^T G_k (x-x_k) + \max_{j\in\mathcal{J}_k}(F_L^k(x), F(x_k) + (x-x_k)^T g_a^k - \alpha_a^k),$$

kde

$$\alpha_a^k = \max\left(|F(x_k) - F_a^k|, \gamma(s_a^k)^\nu\right). \tag{588}$$

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

$$\frac{1}{2}d^T G_k d + v \tag{589}$$

na množině určené omezeními

$$\begin{aligned} -\alpha_j^k + d^T g_j &\leq v, \quad j \in \mathcal{J}_k, &(590)\\ -\alpha_a^k + d^T g_a^k &\leq v, &(591)\end{aligned}$$

(minimalizuje se přes všechny dvojice $(d,v) \in R^{n+1}$ vyhovující nerovnostem (590), (591)).

Věta 257 *Řešení úlohy (589)-(591) lze vyjádřit ve tvaru*

$$\begin{aligned} d_k &= -G_k^{-1}\tilde{g}_a^k, &(592)\\ v_k &= -d_k^T G_k d_k - \tilde{\alpha}_a^k, &(593)\end{aligned}$$

kde

$$\begin{aligned} \tilde{g}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k g_j + \lambda_a^k g_a^k, &(594)\\ \tilde{\alpha}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k \alpha_j^k + \lambda_a^k \alpha_a^k &(595)\end{aligned}$$

a kde Lagrangeovy multiplikátory λ_j^k, $j \in \mathcal{J}_k$, λ_a^k, jsou řešením duální úlohy kvadratického programování: Minimalizovat funkci

$$\frac{1}{2}\left(\sum_{j\in\mathcal{J}_k}\lambda_j g_j + \lambda_a g_a^k\right)^T G_k^{-1}\left(\sum_{j\in\mathcal{J}_k}\lambda_j g_j + \lambda_a g_a^k\right) + \sum_{j\in\mathcal{J}_k}\lambda_j\alpha_j^k + \lambda_a\alpha_a^k \tag{596}$$

na množině určené omezeními

$$\left.\begin{array}{l}\lambda_j \geq 0, \quad j\in\mathcal{J}_k, \quad \lambda_a \geq 0,\\ \sum_{j\in\mathcal{J}_k}\lambda_j + \lambda_a = 1.\end{array}\right\} \tag{597}$$

Minimální hodnota funkce (596), odpovídající řešení úlohy (596)-(597), je

$$w_k = \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1} \tilde{g}_a^k + \tilde{\alpha}_a^k = -v_k - \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1} \tilde{g}_a^k. \tag{598}$$

Důkaz Jelikož matice G_k je positivně definitní, je funkce (589) konvexní. Omezení (590)-(591) jsou lineární a tudíž také konvexní, takže pár $(d_k, v_k) \in R^{n+1}$ je podle věty **??** řešením úlohy (589)-(591) právě tehdy, existují-li Lagrangeovy multiplikátory $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, takové, že

$$\begin{bmatrix} G_k d_k \\ 1 \end{bmatrix} + \sum_{j \in \mathcal{J}_k} \lambda_j^k \begin{bmatrix} g_j \\ -1 \end{bmatrix} + \lambda_a^k \begin{bmatrix} g_a^k \\ -1 \end{bmatrix} = 0, \tag{599}$$

přičemž

$$\begin{aligned} \lambda_j^k > 0 \quad &\Rightarrow \quad -\alpha_j^k + d_k^T g_j = v_k, \\ \lambda_a^k > 0 \quad &\Rightarrow \quad -\alpha_a^k + d_k^T g_a^k = v_k \end{aligned}$$

(podmínky komplementarity). Z poslední rovnice soustavy (599) dostaneme

$$\sum_{j \in \mathcal{J}_k} \lambda_j^k + \lambda_a^k = 1.$$

Platí tedy (592) (594) a (597). Použijeme-li označení (594)-(595) a podmínky komplementarity, můžeme psát

$$-\tilde{\alpha}_a^k + d_k^T \tilde{g}_a^k = v_k,$$

což spolu s (592) dává (593). 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 (596)-(597). Tato úloha je opět konvexní, takže čísla $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, jsou podle věty **??** jejím řešením právě tehdy, existují-li Langrangeovy multiplikátory v_k (odpovídající rovnosti v (597)) a $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$ (odpovídající nerovnostem v (597)) tak, že

$$\begin{aligned} -(g_j)^T d_k + \alpha_j^k + v_k - \mu_j^k &= 0, \quad j \in \mathcal{J}_k, \\ -(g_a^k)^T d_k + \alpha_a^k + v_k - \mu_a^k &= 0, \end{aligned}$$

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 (592) a (594)). Poslední rovnosti však nejsou nic jiného než nerovnosti (590), (591), 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 (589)-(591). □

Poznámka 341 Poznamenejme, že omezení (591) není třeba používat pokud $\mathcal{J}_k = \{1, \ldots, k\}$, neboť je v tomto případě lineární kombinací omezení (590). Pak ale $\lambda_a^k = 0$ v (594)-(595).

Poznámka 342 Kromě agregovaných gradientů (594) se pomocí Lagrangeových multiplikátorů $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$ definují agregované hodnoty

$$\tilde{F}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k F_j^k + \lambda_a^k F_a^k, \tag{600}$$

$$\tilde{s}_a^k = \sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k. \tag{601}$$

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 88 a definici 89. V obou definicích používáme označení

$$\beta_{k+1} = \max\left(|F(x_k) - F_{k+1} - (x_k - y_{k+1})^T g_{k+1}|, \gamma|x_k - y_{k+1}|^\nu\right) \tag{602}$$

a konstanty $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $0 < \tau < 1$ a $D > 0$.

Definice 88 *(Spádový krok) Spádovým krokem nazveme krok, ve kterém platí $t_R^k = t_L^k > 0$,*

$$F(x_{k+1}) \leq F(x_k) - \sigma_L t_L^k w_k \tag{603}$$

a buď $t_L^k \geq \tau$ nebo $\beta_{k+1} > \sigma_A w_k$.

Definice 89 *(Nulový krok) Nulovým krokem nazveme krok, ve kterém platí $t_R^k > t_L^k = 0$,*

$$d_k^T g_{k+1} \geq \beta_{k+1} - \sigma_R w_k \tag{604}$$

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

$$\left.\begin{array}{ll} F_j^{k+1} = F_j^k + (x_{k+1} - x_k)^T g_j, & j \in J_k \\ F_a^{k+1} = \tilde{F}_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k & \\ F_{k+1}^{k+1} = F_{k+1} + (x_{k+1} - y_{k+1}) g_{k+1} & \\ g_a^{k+1} = \tilde{g}_a^k & \\ s_j^{k+1} = s_j^k + \|x_{k+1} - x_k\|, & j \in J_k \\ s_a^{k+1} = \tilde{s}_a^k + \|x_{k+1} - x_k\| & \\ s_{k+1}^{k+1} = \|x_{k+1} - y_{k+1}\| & \end{array}\right\} \tag{605}$$

Zbývá uvést podmínky, které by měly splňovat matice G_k. Abychom zaručili globální konvergenci svazkové metody, použijeme tento předpoklad.

Předpoklad 22 *Matice 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 (589)-(591) (omezení (591) 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 257). 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 88) nebo nulový krok (definice 89). 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 (605) a určíme matici G_{k+1} tak, aby vyhovovala Předpokladu 22. 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 343 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 258 *Nechť funkce $F : R^n \to R$ je lokálně lipschitzovská a nechť pro libovolnou posloupnost $t^i \downarrow 0$ platí*

$$\limsup_{\substack{g^i \in \partial F(x+t^i d) \\ i \to \infty}} d^T g^i \geq \liminf_{i \to \infty} \frac{F(x + t^i d) - F(x)}{t^i}. \tag{606}$$

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 $S = \{t \geq 0 : F(x + td) \leq F - \sigma_T t w\}$. Protože $\{t^i_A\} \subset S$, $t^i_A \uparrow t^*$ a funkce F je spojitá, musí platit

$$F(x + t^* d) \leq F - \sigma_T t^* w, \tag{607}$$

takže $t^* \in S$. Nechť $I = \{i : t^i \notin S\}$. Ukážeme nejprve, že množina I je nekonečná. Pokud by existoval index $\bar{i} \in I$ takový, že $t^i \in S$ $\forall i > \bar{i}$, muselo by platit $t^{\bar{i}}_U = t^i_U \downarrow t^*$ $\forall i > \bar{i}$, neboli $t^* = t^{\bar{i}}_U \notin S$, což je ve sporu s $t^* \in S$. 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 (607) 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 (606), dostaneme

$$-\sigma_T w \leq \liminf_{i \overset{I}{\to} \infty} \frac{F(x + t^* d + (t^i - t^*)d) - F(x + t^* d)}{t^i - t^*} \leq \limsup_{i \overset{I}{\to} \infty} d^T g^i. \tag{608}$$

Vyšetříme nyní dva případy.
(a) Nechť $t^* > 0$. Podle (607) 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 (608).
(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 233 (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

$$\limsup_{i \overset{I}{\to} \infty} d^T g^i \leq -\sigma_R w < -\sigma_T w,$$

což je opět ve sporu s (608). □

Poznámka 344 Podle věty 245 splňuje podmínku (606) každá slabě polohladká funkce, neboť výraz na pravé straně (606) je v tomto případě směrovou derivací (která existuje) a výraz na levé straně je roven limitě (552).

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 23 *Funkce $F : R^n \to R$ je lokálně lipschitzovská na množině $\mathcal{D}_F(F_1) + \overline{B(0,D)}$, kde množina $\mathcal{D}_F(F_1) = \{x \in \mathcal{D} : F(x) \leq F(x_1)\}$ je kompaktní, a je splněna podmínka (606) (například, když F je slabě polohladká).*

Poznámka 345 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 258 platí $\|y_k - z_k\| \leq D$, kde $z_k \in X$, můžeme psát $y_k \in \mathcal{D}_F(F_1) + \overline{B(0,D)}$. Množina $\mathcal{D}_F(F_1) + \overline{B(0,D)}$ je kompaktní, takže posloupnost $\{y_k\}$ je omezená. Z lokální lipschitzovskosti funkce F na $\mathcal{D}_F(F_1) + \overline{B(0,D)}$ plyne omezenost posloupnosti $\{g_k\}$. Podle (609) je i posloupnost $\{\tilde{g}_a^k\}$ omezená. Z (592) a Předpokladu 22 pak plyne omezenost posloupnosti $\{d_k\}$.

Lemma 76 *Existují čísla $\tilde{\lambda}_i^k \geq 0$, $1 \leq i \leq k$, $\tilde{\lambda}_1^k + \ldots + \tilde{\lambda}_k^k = 1$ taková, že hodnoty $\tilde{F}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$ získané v Kroku 2 Algoritmu 5.1 vyhovují vztahům*

$$\left(\tilde{F}_a^k, \tilde{g}_a^k, \tilde{s}_a^k\right) = \sum_{i=1}^{k} \tilde{\lambda}_i^k (F_i^k, g_i, s_i^k) \tag{609}$$

(závorky v (609) 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 (609) (platí to zřejmě pokud $\mathcal{J}_k = \{1, \ldots, k\}$, kdy $\lambda_a^k = 0$, takže vztahy (594), (600), (601) implikují (609) 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 (589)-(591) (nebo úlohy (596)-(597)), 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 (597) 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 (605), (594), (600), (601) 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 77 *Jestliže posloupnost $\{x_k\}$ generovaná Algoritmem 5.1 má hromadný bod $x^* \in R^n$ a existuje podposloupnost $\{x_k\}_K \subset \{x_k\}$ taková, že $x_k \overset{K}{\to} x^*$ a $w_k \overset{K}{\to} 0$, pak bod x^* je stacionárním bodem funkce F (platí $0 \in \partial F(x^*)$).*

Důkaz Podle lemmatu 76 platí (609). Podle věty 194 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

$$(\tilde{g}_a^k, \tilde{s}_a^k) = \sum_{i=1}^{n+2} \lambda^{k,i} \left(g^{k,i}, s^{k,i}\right), \tag{610}$$

kde $\lambda^{k,i} \geq 0$, $1 \leq i \leq n+2$, $\lambda^{k,1} + \ldots + \lambda^{k,n+2} = 1$. Podle poznámky 345 jsou vektory $y^{k,i}$, $g^{k,i}$, $1 \leq i \leq n+2$, omezené, takže existuje podmnožina $\overline{K} \subset K$ taková, že $y^{k,i} \overset{\overline{K}}{\to} y_i^*$, $g^{k,i} \overset{\overline{K}}{\to} g_i^*$, $\lambda^{k,i} \overset{\overline{K}}{\to} \lambda_i^*$, $1 \leq i \leq n+2$. Podle věty 233 (c) platí $g_i^* \in \partial F(y_i^*)$, $1 \leq i \leq n+2$. Z (610) pak plyne $(\tilde{g}_a^k, \tilde{s}_a^k) \to (\tilde{g}_a^*, \tilde{s}_a^*)$, kde

$$(\tilde{g}_a^*, \tilde{s}_a^*) = \sum_{i=1}^{n+2} \lambda_i^* (g_i^*, s_i^*) \tag{611}$$

a $\lambda_i^* \geq 0$, $1 \leq i \leq n+2$, $\lambda_1^* + \ldots + \lambda_{n+2}^* = 1$. Navíc (586) implikuje $s^{k,i} \geq \|x_k - y^{k,i}\|$, což spolu s $x_k \overset{\overline{K}}{\to} x^*$, $y^{k,i} \overset{\overline{K}}{\to} y_i^*$ a $s^{k,i} \overset{\overline{K}}{\to} s_i^*$ dává

$$s_i^* \geq \|x^* - y_i^*\| \tag{612}$$

pro $1 \leq i \leq n+2$. Jelikož $w_k \overset{\overline{K}}{\to} 0$, matice G_k jsou stejnoměrně pozitivně definitní a $\tilde{\alpha}_a^k \geq 0$, musí podle (598) platit $\tilde{g}_a^k \overset{\overline{K}}{\to} 0$, $\tilde{\alpha}_a^k \overset{\overline{K}}{\to} 0$. Podle (587), (588) a (595) dostaneme

$$\begin{aligned}
\tilde{\alpha}_a^k &= \sum_{j \in \mathcal{J}_k} \lambda_j^k \max\left(|F(x_k) - F_j^k|, \gamma (s_j^k)^\nu\right) + \lambda_a^k \max\left(|F(x_k) - F_a^k|, \gamma (s_a^k)^\nu\right) \\
&\geq \max\left(\sum_{j \in \mathcal{J}_k} \lambda_j^k |F(x_k) - F_j^k| + \lambda_a^k |F(x_k) - F_a^k|, \gamma \left(\sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k \right)^\nu \right) \\
&\geq \max\left(\left|F(x_k) - \tilde{F}_a^k\right|, \gamma \left(\tilde{s}_a^k\right)^\nu \right),
\end{aligned} \tag{613}$$

neboť funkce $\max(\cdot, \cdot)$ a $|\cdot|^\nu$, $\nu \geq 1$, jsou konvexní. Platí tedy $\tilde{g}_a^k \overset{\overline{K}}{\to} 0$, $\tilde{s}_a^k \overset{\overline{K}}{\to} 0$, což s použitím (611) a (612) 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 $0 = \lambda_1^* g_1^* + \ldots + \lambda_{n+2}^* g_{n+2}^* \in \partial F(x^*)$. □

Poznámka 346 *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 77. Platí $\tilde{g}_a^k = 0$ a $\tilde{s}_a^k = 0$, což jako v důkazu lemmatu 77 dává $0 \in \partial F(x_k)$.*

Lemma 78 *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 (605) a (588) 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 (613) 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 257 minimem funkce (596) (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 22). Dále poznamenejme, že pro $k > l$ z (592) a (604) 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 (602) 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 345) a ze stejnoměrné pozitivní definitnosti matic G_k (Předpoklad 22). 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

$$w_{k+1} \leq w_k - \frac{(1-\sigma_R)^2 w_k^2}{4M}. \tag{614}$$

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 (614) nerostoucí pro $k > l$). Pak bychom z (614) 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 77. □

Věta 259 *Nechť funkce $F : R^n \to R$ splňuje Předpoklad 23. 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 78 právě jeden hromadný bod posloupnosti $\{x_k\}$, který je stacionárním bodem funkce F. Předpokládejme, že $x_k \overset{K}{\to} x^*$ (množina K a bod x^* existují, protože posloupnost $\{x_k\}$ je omezená). Utvořme nekonečnou množinu

$$\overline{K} = \left\{k = k(i) : k(i) \geq i, i \in K, x_i = \ldots = x_{k(i)} \neq x_{k(i)+1}\right\},$$

takže krok s indexem $k \in \overline{K}$ je spádový a $x_k \overset{\overline{K}}{\to} x^*$. Jelikož posloupnost $\{F(x_k)\}$ je nerostoucí a zdola omezená (protože F je lokálně lipschitzovská na kompaktní množině), musí mít limitu a tudíž $F(x_k) - F(x_{k+1}) \overset{\overline{K}}{\to} 0$. Jelikož pro $k \in \overline{K}$ platí (603), můžeme psát

$$0 \leq \sigma_L t_L^k w_k \leq F(x_k) - F(x_{k+1}),$$

takže $t_L^k w_k \overset{\overline{K}}{\to} 0$. Podle věty 258 platí $\overline{K} = K_1 \cup K_2$, kde $K_1 = \{k \in \overline{K} : t_L^k \geq \tau\}$ a $K_2 = \{k \in \overline{K} : \beta_{k+1} > \sigma_A w_k\}$. Je-li množina K_1 nekonečná, pak z $t_L^k w_k \overset{K_1}{\to} 0$ plyne $w_k \overset{K_1}{\to} 0$ a podle lemmatu 77 je bod x^* stacionárním bodem funkce F. Je-li množina K_1 konečná, musí být množina K_2 nekonečná. Předpokládejme, že existuje číslo δ takové, že množina $K_3 = \{k \in K_2, w_k > \delta\}$ je nekonečná. Pak z $t_L^k w_k \overset{K_3}{\to} 0$ plyne $t_L^k \overset{K_3}{\to} 0$. Z Předpokladu 22 a z omezenosti směrových vektorů (poznámka 345) 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 $t_L^k \overset{K_3}{\to} 0$ implikuje $\|x_{k+1} - x_k\| \overset{K_3}{\to} 0$. Protože ve spádových krocích platí $y_{k+1} = x_{k+1}$, dostaneme $\|y_{k+1} - x_k\| \overset{K_3}{\to} 0$. To po dosazení do (602) a využití spojitosti funkce F dává $\beta_{k+1} \overset{K_3}{\to} 0$. Jelikož $K_3 \subset K_2$, platí $0 \leq \sigma_A w_k < \beta_{k+1}$, takže $w_k \overset{K_3}{\to} 0$, což je ve sporu s definicí množiny K_3. Platí tedy $w_k \overset{K_2}{\to} 0$ a podle lemmatu 77 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 22). 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 22. 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 (589)-(591).