Numerické optimalizační metody

Lukšan, Ladislav
2005

Institute of Computer Science
Academy of Sciences of the Czech Republic

Numerické optimalizační metody

L.Lukšan

Technical report No. 930

Prosinec 2005

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

L.Lukšan [1]

Technical report No. 930

Prosinec 2005

Abstract:

Tato zpráva popisuje teoretické i praktické vlastnosti numerických metod pro nepodmíněnou optimalizaci. Kromě metod pro standardní úlohy středních rozměrů jsou studovány i metody pro rozsáhlé řídké a strukturované úlohy.

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

[1]This work was supported by the Grant Agency of the Czech Academy of Sciences, project code IAA1030405. L.Lukšan is also from Technical University of Liberec, Hálkova 6, 461 17 Liberec.

Contents

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 $F : R^n \to R$ a

$$g(x) = [\partial F/\partial x_1, \ldots, \partial F/\partial x_n]^T,$$

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

Zde $F(x)$ je účelová funkce, $g(x)$ 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 $F : R^n \to R$ je dvakrát spojitě diferencovatelná. V tomto případě budeme psát $F \in \mathcal{C}^2$ nebo $F \in \mathcal{C}^2 : R^n \to R$. Spojitost druhých parciálních derivací implikuje symetrii matice $G(x)$. Při vyšetřování konvergence optimalizačních metod budeme často používat předpoklady (F1)-(F5):

Definice 1 *Řekneme, že funkce $F : R^n \to R$ je zdola omezená, jestliže existuje konstanta $\underline{F}$ taková, že platí*

$$F(x) \geq \underline{F} \quad \forall x \in R^n. \tag{F1}$$

Definice 2 *Řekneme, že funkce $F : R^n \to R$ má kompaktní hladiny, jestliže množina*

$$\mathcal{L}(\overline{F}) = \{x \in R^n : F(x) \leq \overline{F}\} \tag{F2}$$

je kompaktní $\forall \overline{F} \in R$ (prázdná množina se předpokládá kompaktní).

Definice 3 *Řekneme, že funkce $F \in \mathcal{C}^2 : R^n \to R$ má omezené druhé derivace, jestliže existuje konstanta $\overline{G} > 0$ taková, že platí*

$$|d^T G(x) d| \leq \overline{G} \|d\|^2 \quad \forall x \in R^n \quad \forall d \in R^n. \tag{F3}$$

Podmínka (F3) je ekvivalentní podmínce $\|G(x)\| \leq \overline{G}$ $\forall x \in R^n$.

Poznámka 1 Místo omezenosti druhých derivací stačí obvykle předpokládat lipschitzovskost prvních derivací:

$$\|g(x+d) - g(x)\| \leq \overline{G} \|d\| \quad \forall x \in R^n \quad \forall d \in R^n.$$

Jelikož v praktických případech není lipschitzovskost prvních derivací o mnoho slabším předpokladem než existence a omezenost druhých derivací, budeme pro zjednodušení důkazů používat podmínku (F3).

Definice 4 *Řekneme, že funkce $F \in \mathcal{C}^2 : R^n \to R$ je stejnoměrně konvexní, jestliže existuje konstanta $\underline{G} > 0$ taková, že platí*

$$d^T G(x) d \geq \underline{G} \|d\|^2 \quad \forall x \in R^n \quad \forall d \in R^n. \tag{F4}$$

Definice 5 *Řekneme, že funkce $F \in \mathcal{C}^2 : R^n \to R$ má lipschitzovské druhé derivace, jestliže existuje konstanta $\overline{L} > 0$ taková, že platí*

$$\|G(x+d) - G(x)\| \leq \overline{L} \|d\| \quad \forall x \in R^n \quad \forall d \in R^n. \tag{F5}$$

Podmínky (F4) a (F5) jsou často zbytečně silné. Studujeme-li asymptotické chování iteračního procesu v okolí minima $x^* \in R^n$, stačí předpokládat pozitivní definitnost a lokální lipschitzovskost matice $G(x)$ v bodě $x^* \in R^n$. Při definování lokálních vlastností funkce $F : R^n \to R$ budeme používat označení $\mathcal{B}(x^*, \varepsilon) = \{x \in R^n : \| x - x^* \| < \varepsilon\}$ pro ε–kouli, která je okolím bodu x^*.

Definice 6 *Řekneme, že funkce $F \in \mathcal{C}^2 : R^n \to R$ je ryze konvexní v bodě $x^* \in R^n$, je-li matice $G(x^*)$ pozitivně definitní. Pak pro libovolnou konstantu $0 < \underline{G} < \underline{\lambda}(G(x^*))$ existuje číslo $\varepsilon > 0$ takové, že*

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

Definice 7 *Řekneme, že funkce $F \in \mathcal{C}^2 : R^n \to R$ má v bodě $x^* \in R^n$ lokálně lipschitzovské druhé derivace, jestliže existuje konstanta $\overline{L}$ a číslo $\varepsilon > 0$ takové, že platí*

$$\|G(x) - G(x^*)\| \leq \overline{L} \|x - x^*\| \quad \forall x \in \mathcal{B}(x^*, \varepsilon) \quad \forall d \in R^n. \tag{$\overline{\text{F5}}$}$$

Poznámka 2 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 ($\overline{\text{F4}}$) implikuje (F4) a ($\overline{\text{F5}}$) implikuje (F5) $\forall i \geq k$

V konvergenčních důkazech budeme často používat věty o střední hodnotě známé z úvodních kurzů matematické analýzy:

Tvrzení 1 *Nechť $F \in \mathcal{C}^2 : R^n \to R$, $x \in R^n$ a $d \in R^n$. Pak platí*

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

kde $\tilde{x} = x + \tilde{\lambda} d$ a $0 \leq \tilde{\lambda} \leq 1$.

Použijeme-li tvrzení 1 a (F3), dostaneme

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

Použijeme-li tvrzení 1 a (F4), dostaneme

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

Tvrzení 2 *Nechť $F \in \mathcal{C}^2 : R^n \to R$, $x \in R^n$ a $d \in R^n$. Pak platí*

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

Použijeme-li tvrzení 2 a (F3), dostaneme

$$\|g(x+d)-g(x)\| \le \overline{G}\|d\|, \tag{3}$$

$$d^T(g(x+d)-g(x)) \le \overline{G}\|d\|^2. \tag{4}$$

Použijeme-li tvrzení 2 a (F4), dostaneme

$$\|g(x+d)-g(x)\| \ge \underline{G}\|d\|, \tag{5}$$

$$d^T(g(x+d)-g(x)) \ge \underline{G}\|d\|^2. \tag{6}$$

Důkaz posledních dvou nerovností:

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

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

1.2 Podmínky optimality

Definice 8 *Řekneme, že bod $x^* \in R^n$ je lokálním minimem funkce $F : R^n \to R$, 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 $x^* \neq x$, řekneme, že bod $x^* \in R^n$ je ostrým lokálním minimem funkce $F : R^n \to R$.

Věta 1 *Nechť bod $x^* \in R^n$ je lokálním minimem funkce $F : R^n \to R$ 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 positivně semidefinitní).

Důkaz Nechť $F \in C^1$ na $\mathcal{B}(x^*, \varepsilon)$. Předpokládejme, že $g^* = g(x^*) \neq 0$. Jelikož $F \in C^1$ na $\mathcal{B}(x^*, \varepsilon)$, existuje číslo $\overline{\alpha} > 0$, takové, že $x^* - \alpha g^* \in \mathcal{B}(x^*, \varepsilon)$ a $(g^*)^T g(x^* - \alpha g^*) \ge (g^*)^T g^*/2$, pokud $0 \le \alpha \le \overline{\alpha}$ (plyne to ze spojitosti gradientu $g(x^* - \alpha g^*)$). Nechť $0 < \alpha \le \overline{\alpha}$. Pak podle věty o střední hodnotě platí $F(x^* - \alpha g^*) = F(x^*) - \alpha (g^*)^T g(x^* - \tilde{\alpha} g^*)$, kde $0 \le \tilde{\alpha} \le \alpha \le \overline{\alpha}$, takže

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

což je ve sporu s definicí 8. Nechť navíc $F \in C^2$ na $\mathcal{B}(x^*, \varepsilon)$. Předpokládejme, že matice $G(x^*)$ není positivně 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 $F \in C^2$ na $\mathcal{B}(x^*, \varepsilon)$, existuje číslo $\overline{\alpha} > 0$, takové, že $x^* + \alpha v^* \in \mathcal{B}(x^*, \varepsilon)$ a $(v^*)^T G(x^* + \alpha v^*) v^* \le \lambda^* (v^*)^T v^*/2$, pokud $0 \le \alpha \le \overline{\alpha}$ (plyne to ze spojitosti Hessovy matice $G(x^* + \alpha v^*)$). Nechť $0 < \alpha \le \overline{\alpha}$. 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 $0 \le \tilde{\alpha} \le \alpha \le \overline{\alpha}$, takže

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

což je ve sporu s definicí 8.

Věta 2 *Nechť $F \in \mathcal{C}^2 : R^n \to R$ 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 ostrým lokálním minimem funkce $F : R^n \to R$.

Důkaz Jelikož matice $G(x^*)$ je positivně definitní, platí $\lambda^* > 0$, kde λ^* je nejmenší vlastní číslo matice $G(x^*)$. Nechť $v \in R^n$. Jelikož $F \in C^2$ 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

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

1.3 Základní pojmy z teorie konvergence

Nyní se budeme zabývat vlastnostmi konvergentních posloupností.

Definice 9 *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 3 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 1 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 2 o střední hodnotě, dostaneme

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

kde

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

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 (F5), 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 (F5). 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 10 *Jestliže existují index $k \in N$ a čísla $M_k > 0$ a $0 < q < 1$, tak že*

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

$\forall i \geq k$, řekneme, že posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) R-lineárně.

Věta 5 *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) R-lineárně právě tehdy jestliže*

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

Důkaz Z definice 10 plyne

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

$\forall i \geq k$, kde $q < 1$, takže

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

Z druhé strany nechť $\hat{q} < 1$. Pak pro libovolné číslo $\hat{q} < q < 1$ existuje index $k \in N$ tak, že platí

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

$\forall i \geq k$, 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 4 Číslo $\hat{q}$ použité ve větě 5 nezávisí na posunu indexů. Pro libovolné číslo $k \in N$ platí

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

Definice 11 *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* R-superlineárně, jestliže*

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

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

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

Definice 13 *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 volit $M_k = 1$). 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

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

Protože číslo ε je libovolné, musí platit

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

Poznámka 5 Q-lineární (Q-superlineární) konvergence implikuje monotonnost posloupnosti $\|x_i - x^*\|$, $i \in N$ (počínaje vhodným indexem $k \in N$).

Definice 14 *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 6 m-kroková Q-superlineární konvergence implikuje cyklickou m-krokově Q-superlineárnĭ konvergenci.

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

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

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

(používáme nerovnost mezi geometrickým a aritmetickým průměrem a vztah $\lim_{k\to\infty} \frac{1}{k} \sum_{j=1}^{k} o(1) = 0$). Můžeme tedy psát

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

Poznámka 7 Předpoklady věty 7 jsou splněny pro cyklicky přerušovanou metodu sdružených gradientů s asymptoticky přesným výběrem délky kroku (věta 25).

Definice 15 *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) kvadraticky, jestliže existuje index $k \in N$ a konstanta $0 < M_k < \infty$ tak, že*

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

Definice 16 *Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* (alespoň) m-krokově kvadraticky, jestliže existuje index $k \in N$, číslo $m \in N$ a konstanta $0 < M_k < \infty$ tak, že*

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

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

Definice 17 *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,

  $O(n)$ - operací na iteraci.

Nevýhody:

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

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

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

  $O(n^3)$ - 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

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

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 18 *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 8 Definice dostatečné spádovosti směrových vektorů se může zdát dosti umělá. Nicméně je tato definice často velmi užitečná pro důkazy globální konvergence (věta 83). 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 14).

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

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

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

Poznámka 9 Podle věty 29 platí stejný odhad i tehdy, 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 19 *Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje buď silnou Wolfeho podmínku nebo slabou Wolfeho podmínku nebo Goldsteinovu podmínku nebo Armijovu podmínku, jestliže existují čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) taková, že*

$$F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \tag{S2}$$

a buď

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

nebo

$$s_i^T g_{i+1} \geq \varepsilon_2 s_i^T g_i \tag{S3b}$$

nebo

$$F_{i+1} - F_i \geq \varepsilon_2 \alpha_i s_i^T g_i \tag{S3c}$$

nebo $\alpha_i > 0$ je první člen vyhovující podmínce (S2) 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

$$\underline{\beta} \alpha_i^j \leq \alpha_i^{j+1} \leq \overline{\beta} \alpha_i^j \quad \forall j \in N, \tag{S3d}$$

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

Poznámka 10 Při vyšetřování globální konvergence vystačíme s nerovnostmi $0 < \varepsilon_1 < \varepsilon_2 < 1$. Pro zaručení superlineární konvergence (věta 17) je třeba, aby platilo $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$ (v případě podmínky (S2c) navíc $1/2 < \varepsilon_2 < 1$).

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

Definice 20 *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 (S2) a některou z podmínek (S3a)-(S3d). 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 11 Při realizaci metod sdružených gradientů odvozených z metody největšího spádu se používá podmínka (S3a) s $0 < \varepsilon_1 < \varepsilon_2 < 1/2$. Při realizaci metod s proměnnou metrikou odvozených z Newtonovy metody (kde $\alpha_i \to 1$ pro $i \to \infty$) se používá podmínka (S3b) s $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$. Při realizaci metod založených na numerickém výpočtu gradientů se používá podmínka (S3c) s $0 < \varepsilon_1 < 1/2 < \varepsilon_2 < 1$ (obvykle $\varepsilon_2 = 1 - \varepsilon_1$). Při realizaci metod pro nehladké úlohy se používá podmínka (S3d) s $0 < \varepsilon_1 < 1/2$.

Poznámka 12 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 1 *(Konzistence) Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ splňuje podmínky (F1) a (F3) a směrový vektor $s_i \in R^n$ splňuje podmínku* (S1a). *Pak jak silná Wolfeho podmínka tak slabá Wolfeho podmínka tak Goldsteinova podmínka tak Armijova podmínka je konzistentní v tom myslu, že existuje délka kroku $\alpha_i > 0$, která této podmínce vyhovuje.*

Důkaz Nechť $0 < \varepsilon_1 < \varepsilon_2 < 1$. Označme

$$\mathcal{M}_i = \{\alpha \geq 0 : F(x_i + \beta s_i) - F_i \leq \varepsilon_1 \beta s_i^T g_i \quad \forall 0 \leq \beta \leq \alpha\}.$$

Nechť $\tilde{\alpha}_i = \sup \mathcal{M}_i$. 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 definice množiny $\mathcal{M}_i$ dává $\tilde{\alpha}_i \leq (\underline{F} - F_i)/\varepsilon_1 s_i^T g_i < \infty$. Ukážeme nejprve, že

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

a

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

takže délka kroku $\alpha_i = \tilde{\alpha}_i$ splňuje podmínky (S2), (S3b), (S3c) (neboť $0 < \varepsilon_1 < \varepsilon_2 < 1$) a tedy jak slabá Wolfeho podmínka tak Goldsteinova podmínka jsou konzistentní. Rovnost plyne ze spojitosti funkce $F : R \to R^n$, 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 $\varepsilon > \varepsilon_1$. Potom podle (F3) platí

$$\begin{aligned} F(x_i + \alpha s_i) - F_i &\leq F(x_i + \tilde{\alpha}_i s_i) - F_i + s_i^T g(x_i + \tilde{\alpha}_i s_i)(\alpha - \tilde{\alpha}_i) + \frac{1}{2}\overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2 \\ &= \varepsilon_1 \tilde{\alpha}_i s_i^T g_i + \varepsilon(\alpha - \tilde{\alpha}_i) s_i^T g_i + \frac{1}{2}\overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2 \\ &= \varepsilon_1 \alpha s_i^T g_i - (\varepsilon_1 - \varepsilon)(\alpha - \tilde{\alpha}_i) s_i^T g_i + \frac{1}{2}\overline{G}\|s_i\|^2(\alpha - \tilde{\alpha}_i)^2 \end{aligned}$$

a pro $\alpha = \tilde{\alpha}_i + (\varepsilon_1 - \varepsilon) s_i^T g_i / \overline{G}\|s_i\|^2 > \tilde{\alpha}_i$ dostaneme

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

což je spor neboť $\tilde{\alpha}_i = \sup \mathcal{M}_i$. Pokud $s_i^T g(x_i + \tilde{\alpha}_i s_i) < 0$, splňuje délka kroku $\alpha_i = \tilde{\alpha}_i$ podmínky (S2) a (S3a). V opačném případě z nerovnosti $s_i^T g_i < 0$ a ze spojitosti derivací funkce $F : R \to R^n$ plyne existence čísla $0 < \alpha_i < \tilde{\alpha}_i$ takového, že $s_i^T g(x_i + \alpha_i s_i) = 0$. Zřejmě $\alpha_i \in \mathcal{M}_i$, takže platí (S2) a (S3a).

Silná Wolfeho podmínka je tedy konzistentní. 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 Armijovy podmínky.

2.2 Globální konvergence

Nyní budeme studovat globální konvergenci metod spádových směrů. Nejprve dokážeme pomocnou větu, která zdůvodňuje použití podmínek (S3a)-(S3d).

Lemma 2 *Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ splňuje podmínky (F1) a (F3) a délka kroku se vybírá tak, aby byla splněna podmínka (S2) a některá z podmínek (S3a)-(S3d). Pak existuje konstanta $\varepsilon_3 > 0$ taková, že pro libovolný index $i \in N$ platí*

$$\alpha_i \geq -\frac{\varepsilon_3 s_i^T g_i}{\overline{G}\|s_i\|^2} = \frac{\varepsilon_3 \cos\theta_i \|g_i\|}{\overline{G}\|s_i\|} \tag{a}$$

a

$$F_{i+1} - F_i \leq -\frac{\varepsilon_1\varepsilon_3(s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_3}{\overline{G}}\cos^2\theta_i\|g_i\|^2. \tag{b}$$

Důkaz Nerovnost (b) plyne bezprostředně z nerovnosti (a), neboť podle (S2) platí

$$F_{i+1} - F_i \leq \varepsilon_1\alpha_i s_i^T g_i \leq -\frac{\varepsilon_1\varepsilon_3(s_i^T g_i)^2}{\overline{G}\|s_i\|^2} = -\frac{\varepsilon_1\varepsilon_3\cos^2\theta_i}{\overline{G}}\|g_i\|^2.$$

Zbývá tedy dokázat nerovnost (a). Zřejmě (S3a) implikuje (S3b). Platí-li (S3b), můžeme s použitím odhadu (4) 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 (a) s $\varepsilon_3 = (1 - \varepsilon_2) > 0$. Platí-li (S3c), můžeme s použitím odhadu (1) psát

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

neboli

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

takže platí (a) s $\varepsilon_3 = 2(1 - \varepsilon_2) > 0$. Platí-li (S3d), pak buď $\alpha_i = \alpha_i^1$, takže platí (a) s $\varepsilon_3 = \underline{\alpha}\overline{G} > 0$, nebo $\alpha_i = \alpha_i^j \geq \underline{\beta}\alpha_i^{j-1}$, kde

$$F(x_i + \alpha_i^{j-1}s_i) - F(x_i) \geq \varepsilon_1\alpha_i^{j-1}s_i^T g_i.$$

Použijeme-li odhad (1), dostaneme

$$F(x_i + \alpha_i^{j-1}s_i) - F(x_i) \leq \alpha_i^{j-1}s_i^T g_i + \frac{1}{2}(\alpha_i^{j-1})^2\overline{G}\|s_i\|^2,$$

což spolu s předchozí nerovností dává

$$s_i^T g_i + \frac{1}{2}\alpha_i^{j-1}\overline{G}\|s_i\|^2 \geq \varepsilon_1 s_i^T g_i$$

a jelikož $\alpha_i = \alpha_i^j \geq \underline{\beta}\alpha_i^{j-1}$, můžeme psát

$$\alpha_i \geq \underline{\beta}\frac{2(\varepsilon_1 - 1)s_i^T g_i}{\overline{G}\|s_i\|^2} \geq \underline{\beta}\frac{2(1-\varepsilon_1)\cos\theta_i\|g_i\|}{\overline{G}\|s_i\|},$$

takže platí (a) s $\varepsilon_3 = 2\underline{\beta}(1-\varepsilon_1) > 0$.

Poznámka 13 Jsou-li splněny předpoklady lemmatu 2, platí

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

To plyne bezprostředně z (b), neboť

$$F_1 - \underline{F} \geq \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \sum_{i=1}^{\infty}\frac{\varepsilon_1\varepsilon_3(s_i^T g_i)^2}{\overline{G}\|s_i\|^2}$$

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

Věta 9 *(Globální konvergence) Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ splňuje podmínky (F1) a (F3). Pak metoda spádových směrů, pro kterou platí*

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

je globálně konvergentní.

Důkaz Použijeme-li (b), můžeme psát

$$F_{i+1} = F_1 + \sum_{j=1}^{i}(F_{j+1} - F_j) \leq F_1 - \frac{\varepsilon_1\varepsilon_3}{\overline{G}}\sum_{j=1}^{i}\cos^2\theta_j\|g_j\|^2.$$

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

$$\underline{F} \leq \lim_{i\to\infty} F_i \leq F_1 - (\varepsilon_1\varepsilon_3/\overline{G})\sum_{i=1}^{\infty}\cos^2\theta_i\|g_i\|^2,$$

takže

$$\sum_{i=1}^{\infty}\cos^2\theta_i\|g_i\|^2 \leq \frac{(F_1 - \underline{F})\overline{G}}{\varepsilon_1\varepsilon_3} < \infty.$$

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 14 Podmínka použitá ve větě 9 je splněna například tehdy, jestliže existuje konstanta $\underline{c}$ taková, že

$$\sum_{i=1}^{k} \cos^2 \theta_i \geq \underline{c}\, k \quad \forall k \in N.$$

Tuto nerovnost lze dokázat pro některé typy metod s proměnnou metrikou.

Poznámka 15 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 $i \to \infty$.

Důsledek 1 *Metoda největšího spádu je globálně konvergentní, přičemž* $\|g_i\| \to 0$ *pro* $i \to \infty$.

Poznámka 16 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

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

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

Poznámka 17 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.

Poznámka 18 Metoda spádových směrů je globálně konvergentní například tehdy, existuje-li konstanta $\overline{\kappa} > 0$ taková, že

$$\frac{1}{\cos \theta_{i+1}^2} \leq \frac{1}{\cos^2 \theta_i} + \overline{\kappa} \quad \forall i \in N.$$

Pak platí $1/\cos^2 \theta_i \leq 1/\cos^2 \theta_1 + i\overline{\kappa} \leq i(1/\cos^2 \theta_1 + \overline{\kappa})$, což dává

$$\sum_{i=1}^{\infty} \cos^2 \theta_i \geq \frac{\cos^2 \theta_1}{1 + \overline{\kappa} \cos^2 \theta_1} \sum_{i=1}^{\infty} \frac{1}{i} > \infty,$$

neboť harmonická řada je divergentní. Podobně metoda spádových směrů používající směrové vektory $s_i \in R^n$, $i \in N$, určené řešením soustav lineárních rovnic $B_i s_i = -g_i$ je globálně konvergentní například tehdy, existuje-li konstanta $\overline{\kappa} > 0$ taková, že

$$\kappa_{i+1} \leq \kappa_i + \overline{\kappa}.$$

Ukážeme, že metoda dostatečně spádových směrů je globálně konvergentní

Věta 10 *Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ 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), (S2), definici čísla $\cos\theta_i$ a definici dostatečné spádovosti (nerovnost z posnámky 8), 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 $0 < \underline{\omega} < 1$ takové že

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

$\forall k \in N$, neboli

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

což spolu s předpoklady věty dává $\cos\theta_k \geq 1/C_k \geq \underline{\omega}/C_1\ \forall k \in N$. To je však spor, neboť stejnoměrná metoda spádových směrů je podle poznámky 15 globálně konvergentní.

Ukážeme ještě jeden způsob, jak lze konstruovat globálně konvergentní metody pomocí korekcí směrových vektorů, což je obvykle šetrnější než způsob uvedený v poznámce 17.

Věta 11 *Uvažujme metodu spádových směrů, která používá směrové vektory*

$$s_i = -H_i g_i - \sigma\|H_i g_i\|\, g_i, \quad i \in N,$$

kde H_i, $i \in N$, jsou positivně semidefinitní matice takové, že $H_i g_i \neq 0$, a $\sigma > 0$ je číslo, které nezávisí na indexu $i \in N$. Splňuje-li funkce $F \in \mathcal{C}^2 : R^n \to R$ podmínky (F1) a (F3), je tato metoda globálně konvergentní.

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

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

a

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

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

$$\frac{-s_i^T g_i}{\|s_i\| \|g_i\|} \geq \frac{\sigma \|H_i g_i\| \|g_i\|^2}{(1+\sigma\|g_i\|)\|H_i g_i\| \|g_i\|} = \frac{\sigma \|g_i\|}{1+\sigma\|g_i\|} \geq \frac{\sigma\underline{\varepsilon}}{1+\sigma\underline{\varepsilon}} \triangleq \varepsilon_0,$$

neboť funkce $x/(1+x)$ je rostoucí. Směrové vektory s_i, $i \in N$, jsou tedy stejnoměrně spádové, což podle poznámky 15 implikuje $\|g_i\| \to 0$. To je však ve sporu s předpokladem, že $\|g_i\| \geq \underline{\varepsilon}$ $\forall i \in N$.

Poznámka 19 Metody s proměnnou metrikou používají směrové vektory $s_i = -H_i g_i$, $i \in N$. Věta 47 tvrdí, že metody s proměnnou metrikou jsou globálné konvergentní, splňuje-li funkce $F \in \mathcal{C}^2 : R^n \to R$ podmínky (F1), (F3) a (F4). Bez požadavku stejnoměrné konvexity (F4) 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 $\sigma = 10^{-12}$.

Zatím jsme se zabývali pouze metodami, kde posloupnost F_i, $i \in N$, byla nerostoucí. Někdy je výhodné (zejména v souvislosti s Newtonovou metodou) používat nemonotonní metody spádových směrů.

Definice 21 *Řekneme, že délka kroku $\alpha_i > 0$, $i \in N$, splňuje buď silnou nebo slabou nemonotonní Wolfeho podmínku, jestliže existují čísla $0 < \varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) tak, že*

$$F_{i+1} \leq \overline{F}_i + \varepsilon_1 \alpha_i s_i^T g_i \tag{$\overline{\text{S2}}$}$$

*a platí buď (*S3a*) nebo (*S3b*), přičemž $\overline{n}_1 = 1$, $\overline{F}_1 = F_1$ a*

$$\overline{n}_{i+1} = \lambda_i \overline{n}_i + 1 \tag{$\overline{\text{S4a}}$}$$

$$\overline{F}_{i+1} = \frac{\lambda_i \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \tag{$\overline{\text{S4b}}$}$$

pro $i \in N$ a $0 \leq \lambda_i \leq 1$.

Poznámka 20 Pokud $\lambda_i = 0$ $\forall i \in N$, platí $\overline{n}_i = 1$ a $\overline{F}_i = F_i$ pro $i \in N$. Pokud $\lambda_i = 1$ $\forall i \in N$, platí $\overline{n}_i = i$ a

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

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

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

pro $i \in N$, neboť z ($\overline{\text{S2}}$) a ($\overline{\text{S4}}$) plyne

$$F_{i+1} = \frac{\lambda_i \overline{n}_i + 1}{\overline{n}_{i+1}} F_{i+1} \leq \frac{\lambda_i \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \leq \frac{\lambda_i \overline{n}_i + 1}{\overline{n}_{i+1}} \overline{F}_i = \overline{F}_i$$

a funkce

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

je pro $F_{i+1} \leq \overline{F}_i$ neklesající.

Definice 22 *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 ($\overline{\text{S2}}$) (s $\overline{F}_i$ podle ($\overline{\text{S4}}$)) a některou z podmínek (S3a) nebo (S3b).*

Věta 12 *(Globální konvergence) Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ splňuje podmínky (F1) a (F3). Pak nemonotonní metoda spádových směrů, pro kterou platí*

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

je globálně konvergentní.

Důkaz Předně poznamenejme, že z (S3a) nebo (S3b) plyne nerovnost (a) uvedená v lemmatu 2. Použijeme-li tuto nerovnost spolu s ($\overline{\text{S2}}$), dostaneme

$$F_{i+1} \le \overline{F}_i + \varepsilon_1 \alpha_i s_i^T g_i \le \overline{F}_i - \frac{\varepsilon_1 \varepsilon_3 \cos^2 \theta_i}{\overline{G}} \|g_i\|^2,$$

což spolu s ($\overline{\text{S4}}$) dává

$$\overline{F}_{i+1} = \frac{\lambda_i \overline{n}_i \overline{F}_i + F_{i+1}}{\overline{n}_{i+1}} \le \frac{\lambda_i \overline{n}_i + 1}{\overline{n}_{i+1}} \overline{F}_i - \frac{\varepsilon_1 \varepsilon_3 \cos^2 \theta_i}{\overline{n}_{i+1} \overline{G}} \|g_i\|^2 = \overline{F}_i - \frac{\varepsilon_1 \varepsilon_3 \cos^2 \theta_i}{\overline{n}_{i+1} \overline{G}} \|g_i\|^2.$$

Jelikož podle (F1) a poznámky 20 platí $\overline{F}_i \ge F_i \ge \underline{F}$, můžeme tak jako v důkazu věty 9 psát

$$F_1 - \underline{F} \ge \frac{\varepsilon_1 \varepsilon_3}{\overline{G}} \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{\overline{n}_{i+1}} \|g_i\|^2,$$

neboli

$$\frac{1}{2} \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{i} \|g_i\|^2 \le \sum_{i=1}^{\infty} \frac{\cos^2 \theta_i}{\overline{n}_{i+1}} \|g_i\|^2 \le \frac{(F_1 - \underline{F})\overline{G}}{\varepsilon_1 \varepsilon_3} < \infty,$$

neboť podle poznámky 20 platí $\overline{n}_{i+1} \le i + 1 \le 2i$. Z poslední nerovnosti dostaneme dokazované tvrzení postupem uvedeným v důkazu věty 9.

Poznámka 21 Podmínka použitá ve větě 12 je mnohem silnější než podmínka vystupující ve větě 9. Je však splněna pro nemonotonní metody stejnoměrně spádových směrů, kdy $\cos\theta_i \ge \varepsilon_0$ $\forall i \in N$. Jestliže kromě $\cos\theta_i \ge \varepsilon_0$ platí $0 \le \lambda_i < 1$ $\forall i \in N$, dá se dokázat, že $\lim_{i\to\infty} \|g_i\| = 0$.

Je-li metoda spádových směru globálně konvergentní (ve smyslu definice 17), nemusí ještě platit $x_i \to x^*$. Splňuje-li funkce $F : R^n \to R$ 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 $x_i \in R^n$, $i \in N$, postačujícím podmínkám pro lokální minimum (tvrzení 2), pak za jistých předpokladů platí $x_i \to x^*$.

Věta 13 *Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ splňuje podmínky (F1)-(F3) a nechť $x^* \in R^n$ je hromadným bodem posloupnosti $x_i \in R^n$, $i \in N$, generované metodou spádových směrů takovou, že $\alpha_i \le \overline{\alpha}\|g_i\|/\|s_i\|$ $\forall i \in N$. Pak, vyhovuje-li bod $x^* \in R^n$ předpokladům věty 2, platí $x_i \to x^*$.*

Důkaz Protože bod $x^* \in R^n$ vyhovuje předpokladům tvrzení 2, platí $g(x^*) = 0$ a $\underline{\lambda}(G(x^*)) > 0$, kde $\underline{\lambda}(G(x^*))$ je nejmenší vlastní číslo matice $G(x^*)$. Nechť $\underline{G} < \underline{\lambda}(G(x^*))$. Ze spojitosti Hessovy matice $G(x)$ plyne existence čísla ε takového, že

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

pokud $x \in \mathcal{B}(x^*, \varepsilon)$. Jelikož je navíc splněna podmínka (F3), dostaneme s použitím odhadů (1)-(3) a (2)-(5) nerovnosti

$$F - F^* \leq \frac{1}{2}\overline{G}\|x - x^*\|^2, \tag{7}$$

$$F - F^* \geq \frac{1}{2}\underline{G}\|x - x^*\|^2, \tag{8}$$

$$\|g\| \leq \overline{G}\|x - x^*\|, \tag{9}$$

$$\|g\| \geq \underline{G}\|x - x^*\|, \tag{10}$$

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

$$F_i - F^* < \frac{\underline{G}\varepsilon^2}{2(1 + \overline{\alpha}\overline{G})^2} \qquad \forall i \geq l.$$

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 $x_k \in \mathcal{B}(x^*, \varepsilon)$. Použijeme-li podmínku $\alpha_k \leq \overline{\alpha}\|g_k\|/\|s_k\|$ a (9), dostaneme

$$\|x_{k+1} - x^*\| \leq \|x_k - x^*\| + \overline{\alpha}\|g_k\| \leq (1 + \overline{\alpha}\overline{G})\|x_k - x^*\| \tag{11}$$

a podle (8) platí

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

neboli

$$\|x_k - x^*\| < \frac{\varepsilon}{1 + \overline{\alpha}\overline{G}},$$

což po dosazení do (11) 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

$$\frac{\underline{G}}{2}\|x_i - x^*\|^2 \leq F_i - F^* < \frac{\underline{G}\varepsilon^2}{2(1 + \overline{\alpha}\overline{G})^2}$$

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

Poznámka 22 Podmínka $\alpha_i \leq \overline{\alpha}\|g_i\|/\|s_i\|$ není příliš omezující. Tuto podmínku splňuje Armijův výběr délky kroku a také pravidla (S3a), (S3b), (S3c) 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 silnější podmínkou (F4). 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 14 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou dostatečně spádových směrů. 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 (F3) a (F4). Pak směrové vektory jsou stejnoměrně spádové (existuje číslo $\varepsilon_0 > 0$ takové že platí (S1b)).*

Důkaz Použijeme-li (S2), definici čísla $\cos\theta_i$ a definici dostatečné spádovosti (nerovnost z posnámky 8), můžeme (podobně jako v důkazu věty 10) psát

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

Z druhé strany (7) a (10) implikuje

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

(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 $\varepsilon_0 = \underline{\omega}/C_1 > 0$ takové, že $\cos\theta_i \geq \varepsilon_0 \ \forall i \in N$.

Jsou li splněny předpoklady věty 14, je metoda stejmoměrně 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 14.

Věta 15 *(Lineární konvergence) Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů takovou, že*

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

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 (F3) a (F4). Pak platí

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

kde $q = \sqrt{1 - \underline{c}\varepsilon_1\varepsilon_3\underline{G}/\overline{G}}$.

Důkaz Podle (2) 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 (10), dostaneme

$$\|g\|^2 \geq \underline{G}(F - F^*). \tag{12}$$

Platí tedy

$$F_{i+1} - F^* \leq F_i - F^* - \frac{\cos^2\theta_i \varepsilon_1 \varepsilon_3}{\overline{G}} \|g_i\|^2 \leq \left(1 - \frac{\cos^2\theta_i \varepsilon_1 \varepsilon_3 \underline{G}}{\overline{G}}\right)(F_i - F^*) = (1 - \frac{\cos^2\theta_i}{\overline{c}})(F_i - F^*)$$

$\forall i \in N$, kde $\overline{c} = \overline{G}/(\varepsilon_1\varepsilon_3\underline{G})$ a $\varepsilon_3 > 0$ je číslo z lemmatu 2. 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), což s použitím (7) a (8) 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 $q = \sqrt{1-\underline{c}/\overline{c}} = \sqrt{1-\underline{c}\varepsilon_1\varepsilon_3\underline{G}/\overline{G}}$.

Poznámka 23 Z monotonie posloupnosti F_i, $i \in N$, a z nerovností (7), (8) 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 24 Podle věty 15 a poznámky 23 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 3.

Definice 23 *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 3 *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 $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje podmínkám (F3) a (F4). 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 (F3) a (F4) dává $g_i \sim e_i$, takže podle poznámky 23 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 (F3) a (F4) 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 4 *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 16 *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 $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje podmínkám (F3) a (F4). 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 (F3) a (F4) 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 3 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 4 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 14, 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 25 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 26 Používáme-li směrové vektory $s_i = -H_i g_i$, platí

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

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 16 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 4 dostaneme

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

Poznámka 27 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 přitom vyžadovat, aby konstanty ε_1 a ε_2 v podmínkách (S2) a (S3) vyhovovaly nerovnostem uvedeným v poznámce 10, tedy aby platilo $0 < \varepsilon_1 < 1/2$ a v případě podmínky (S2c) též $1/2 < \varepsilon_2 < 1$.

Věta 17 *(Superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spádových směrů taková, že $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje podmínkám (F3) a (F4). Nechť*

$$\lim_{i\to\infty} \frac{\|B_i s_i + g_i\|}{\|g_i\|} = 0 \tag{13}$$

a

$$\lim_{i\to\infty} \frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0 \tag{14}$$

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

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

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

$\forall i \geq k_1$. Označme $\omega_i = (B_i s_i + g_i)/\|g_i\|$ a $\vartheta_i = (B_i - G_i)s_i/\|s_i\|$. Pak platí

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

takže

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

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

a jelikož $\|\vartheta_i\| \to 0$ a $\|\omega_i\| \to 0$ (podle (13) a (14)) 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 ω_i a ϑ_i a z (a) plyne

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

a jelikož $\|\omega_i\| \to 0$ a $\|\vartheta_i\| \to 0$ (podle (13) a (14)) 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) a (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 (13), (14) 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 (S2c) též $1/2 < \varepsilon_2 < 1$), existuje index $k \geq k_2$ takový, že (S2) a (S3a), (S3b), (S3c) (s $\alpha_i = 1$) platí $\forall i \geq k$.
(d) Superlineární konvergence. Použijeme větu 3, podle které

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

Použijeme-li 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 (13), (14) a (a) platí

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

Poznámka 28 Platí také v jistém smyslu obrácená věta: 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 $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje podmínkám (F3) a (F4), přičemž 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 $x^* \in R^n$. Pak (13) $\Rightarrow$ (14) a (14) $\Rightarrow$ (13). Jinými slovy, platí-li některá z podmínek (13) nebo (14), platí i ta druhá. Důkaz tohoto tvrzení se provádí tak, že se nejprve pomocí vztahu $\|g_{i+1}\|/\|g_i\| \to 0$ a věty 3 dokáže platnost tvrzení (a) z důkazu věty 17. Pak se podobnou argumentací jako v části (d) důkazu věty 17 ukáže, že (13) $\Rightarrow$ (14) a (14) $\Rightarrow$ (13).

2.4 Výběr délky kroku

Nyní se budeme zabývat implementací metod spádových směrů. Popíšeme nejprve algoritmus pro výběr délky kroku.

Algoritmus 1 (S3b) 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 (S2) a (S3b), ukončíme výpočet s $\alpha_i = \alpha$. Není-li splněna podmínka (S2), 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 (S2) a (S3b), ukončíme výpočet s $\alpha_i = \alpha$. Není-li splněna podmínka (S2) položíme $\overline{\alpha} = \alpha$ a přejdeme na krok 4. V opačném případě položíme $\underline{\alpha} = \alpha$ a přejdeme na krok 4.

Poznámka 29 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. Počáteční délku kroku budeme značit α_0 a další hodnoty $\underline{\alpha}_j \le \alpha_j \le \overline{\alpha}_j$, $j \in N$. 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 (S2) 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 1) plyne $\underline{\alpha}_j \le (\underline{F}-\varphi(0))/(\varepsilon_1\varphi'(0))$. Jelikož pro $j \ge 2$ platí $\underline{\alpha}_j \ge \gamma_1^{j-2}\alpha_0$ a $\gamma_1 > 1$, dostaneme po konečném počtu extrapolací číslo které je větší než uvedená mez. První fáze algoritmu tedy obsahuje konečný počet kroků.
(b) Ve druhé fázi algoritmu platí $\varphi(\underline{\alpha}_j) - \varphi(0) \le \varepsilon_1\underline{\alpha}_j\varphi'(0)$ a $\varphi'(\underline{\alpha}_j) < \varepsilon_2\varphi'(0)$. Označme

$$\mathcal{M}_j = \{\alpha \ge \underline{\alpha}_j : \varphi(\beta) - \varphi(0) \le \varepsilon_1\beta\varphi'(0) \quad \forall\, 0 \le \beta \le \alpha\}.$$

Nechť $\tilde{\alpha_j} = \sup \mathcal{M}_j$. Pak podobně jako v důkazu lemmatu 1 platí $\varphi(\tilde{\alpha}_j) - \varphi(0) = \varepsilon_1 \tilde{\alpha}_j \varphi'(0)$ a $\varphi'(\tilde{\alpha}_j) \geq \varepsilon_1 \varphi'(0)$. Použijeme-li tyto nerovnosti a (F1), dostaneme

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

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

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

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 (S1) (stejnoměrná spádovost) a vyhovuje-li funkce F podmínkám (F3) a (F4), můžeme předchozí tvrzení podstatně zesílit (budeme to potřebovat pro důkaz asymptotické přesnosti výběru délky kroku).

Lemma 5 *Uvažujme algoritmus 1 s počáteční délkou kroku $\delta_1\|g\|/\|s\| \leq \alpha_0 \leq \delta_2\|g\|/\|s\|$, kde konstanty δ_1 a δ_2 nezávisí na vnějším indexu. Nechť je splněna podmínka (S1) a nechť funkce F vyhovuje podmínkám (F3) a (F4). 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 lze nalézt 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 $\overline{\alpha}_j = \alpha_0$, takže lze položit $c_1 = \delta_1$ a $c_2 = \delta_2$. V dalších fázích algoritmu použijeme nerovnosti uvedené v důkazu věty 18.
(a) V první fázi algoritmu využijeme toho, že vzhledem k (F4) můžeme $\underline{F}$ nahradit F^*, což s použitím (12) (důkaz věty 15) a nerovnosti (S1) (zapsané ve tvaru $-\varphi(0) \geq \varepsilon_1\|s\|\|g\|$) dává

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

S druhé strany víme že $\underline{\alpha}_j \geq \gamma_1^{j-2}\alpha_0$ a $\overline{\alpha}_j - \underline{\alpha}_j \geq (\gamma_1 - 1)\underline{\alpha}_j$. Platí tedy

$$\overline{\alpha}_j \geq \overline{\alpha}_j - \underline{\alpha}_j \geq (\gamma_1 - 1)\gamma_1^{j-2}\alpha_0 \geq (\gamma_1 - 1)\delta_1\|g\|/\|s\|.$$

(b) Ve druhé fázi algoritmu se již $\overline{\alpha}_j$ nezvětšuje, takže můžeme psát

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

(c) Jelikož $\alpha_0 \sim \|g\|/\|s\|$ a stejnou vlastnost mají všechny meze uvedené v (a) a (b), lze počet kroků algoritmu 1 omezit číslem, které nezávisí na vnějším indexu.

Poznámka 30 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):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{2(1-A)}. \tag{15}$$

Kvadratická interpolace (dvě derivace):

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{1-B}. \tag{16}$$

Kubická interpolace:

$$\alpha - \underline{\alpha} = \frac{\overline{\alpha} - \underline{\alpha}}{D + \sqrt{D^2 - 3C}}, \tag{17}$$

kde

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

Věta 19 *Nechť jsou splněny předpoklady lemmatu 5. Pak je-li délka kroku v algoritmu 1 spočtena podle (15) nebo (16) nebo (17), 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 3. Jelikož z (F3) a (F4) plyne $s_i^T G^* s_i \sim \|s_i\|^2$ a z (S1) 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 5 platí $\overline{\alpha}_i \sim \|g_i\|/\|s_i\|$ a $\overline{\alpha}_i - \underline{\alpha}_i \sim \|g_i\|/\|s_i\|$, takže $\alpha_i^* - \underline{\alpha}_i = O(\|g_i\|/\|s_i\|)$ 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 $\|g_i\| \sim \|e_i\|$ ((F3), (F4) 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 30

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

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 $s_i^T g(x_i + \alpha_i^*)/s_i^T g_i \to 0$). Nyní se omezíme na vzorec (17) (důkaz pro (15) a (16) 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 (17), 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 31 Je-li kromě podmínek (F3) a (F4) splněna i podmínka (F5), 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 32 Počáteční výběr délky kroku. Pokud $s_i \sim g_i$, což je případ většiny efektivních metod (Newtonova metoda, metody s proměnnou metrikou, přerušované metody sdružených gradientů), 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 $\alpha_0 = \min(1, 2(\underline{F} - F_i)/s_i^T g_i)$.

Poznámka 33 Shrnutí. Pro metody spádových směrů platí tyto implikace:

- (F1) - (F3) $\Rightarrow$ globální konvergence, pokud

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

- (F3) - (F4) $\Rightarrow$ lineární konvergence, pokud

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

- (F3) - (F4) $\Rightarrow$ asymptotický odhad

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

  jsou-li směrové vektory stejnoměrně spádové a používáme-li asymptoticky přesný výběr délky kroku.

- (F3) - (F4) $\Rightarrow$ superlineární konvergence, pokud

$$\frac{\|B_i s_i + g_i\|}{\|g_i\|} \to 0 \quad \text{a} \quad \frac{\|(G^* - B_i)s_i\|}{\|g_i\|} \to 0.$$

3 Metody sdružených gradientů

3.1 Základní vlastnosti metod sdružených gradientů

Definice 24 *Řekneme, že základní optimalizační metoda je metodou sdružených gradientů jestliže*

$$s_1 = -g_1$$

a

$$s_{i+1} = -g_{i+1} + \beta_i s_i, \tag{CG}$$

pro $i \in N$, kde

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

(Hestenes, Stiefel), nebo

$$\beta_i = \beta_i^{PR} = \frac{y_i^T g_{i+1}}{g_i^T g_i} \tag{CGb}$$

(Polak, Ribiere), nebo

$$\beta_i = \beta_i^{FR} = \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i} \tag{CGc}$$

(Fletcher, Reeves), nebo

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

(Dai, Yuan). Přitom $y_i = g_{i+1} - g_i$.

Poznámka 34 Metoda (CGa) je nejméně závislá na nepřesném výběru délky kroku a spolu s metodou (CGb) dává nejlepší praktické výsledky. Metoda (CGc) je nejjednodušší a je globálně konvergentní bez přerušování iteračního procesu. Metoda (CGd) je také globálně konvergentní (dokonce za slabších předpokladů než metoda (CGc)).

Poznámka 35 Důvod proč se používají vzorce (CGa)–(CGd) plyne z věty 20 a poznámky 36. Odtud také plyne, že lze používat i vzorce

$$\beta_i = \frac{g_{i+1}^T y_i}{|g_i^T s_i|}$$

a

$$\beta_i = \frac{g_{i+1}^T g_{i+1}}{|g_i^T s_i|}.$$

První z těchto metod má podobné vlastnosti jako metody (CGa)–(CGb) a je velmi efektivní pro praktické použití. Druhá generuje směrové vektory, které jsou vždy spádové, nelze však dokázat, že je globálně konvergentní.

Významnou vlastností metod sdružených gradientů je nalezení minima kvadratické funkce po konečném počtu kroků, v případě, že výběr délky kroku je přesný.

Věta 20 *(Kvadratické ukončení) Nechť $Q : R^n \to R$ je ryze konvexní kvadratická funkce, $Q(x) = \frac{1}{2}(x - x^*)^T G(x - x^*)$. Nechť x_i, $i \in N$, je posloupnost generovaná metodou sdružených gradientů 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 (Pro CGa). Předpokládejme, že $g_i \neq 0 \; \forall 1 \leq i \leq n$. Dokážeme indukcí, že $s_i \neq 0$ a $\alpha_i \neq 0$ $\forall 1 \leq i \leq n$ a že platí

$$s_j^T g_i = 0, \tag{α}$$

$$g_j^T g_i = 0, \tag{β}$$

$$s_j^T G s_i = 0, \tag{γ}$$

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

$\forall 1 \leq j < i \leq n+1$. Rovnosti (δ) a (γ) jsou ekvivalentní, neboť pro kvadratickou funkci $Q(x)$ platí $y_i = g_{i+1} - g_i = G(x_{i+1} - x_i) = G d_i = \alpha_i s_i$ a $\alpha_i \neq 0$ podle indukčního předpokladu. Z (β) plyne, že nenulové gradienty g_i, $1 \leq i \leq n$, jsou vzájemně ortogonální, tudíž lineárně nezávislé, takže nutně $g_{n+1} = 0$. Zřejmě $s_1^T g_1 = -g_1^T g_1 < 0$ 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 (α) a (δ) platí:

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

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

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

takže podle (a) platí

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

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

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

$\forall 1 \leq j < i$ neboť podle předpokladu (δ) platí $y_j^T s_i = 0 \; \forall 1 \leq j < i$. Dále podle (CGa) 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 $s_j^T G s_{i+1} = 0 \; \forall 1 \leq j \leq i$.

Poznámka 36 Důkaz byl proveden pro (CGa). Věta 20 platí i pro ostatní metody sdružených gradientů neboť podle (β) platí

$$y_i^T g_{i+1} = g_{i+1}^T g_{i+1} - g_i^T g_{i+1} = g_{i+1}^T g_{i+1}$$

a z (α) plyne

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

Poznámka 37 Nechť H je symetrická pozitivně definitní matice. Položme $\tilde{x} = H^{-1/2}x$ a $\tilde{F}(\tilde{x}) = F(x)$, takže $\tilde{g}(\tilde{x}) = H^{1/2}g(x)$ a $\tilde{G}(\tilde{x}) = H^{1/2}G(x)H^{1/2}$. Aplikujeme-li metodu sdružených gradientů na funkci $\tilde{F}(\tilde{x})$ a vrátíme-li se k původním proměnným, dostaneme

$$s_1 = -Hg_1$$

a

$$s_{i+1} = -Hg_{i+1} + \beta_i s_i, \tag{PCG}$$

pro $i \in N$, kde

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

nebo

$$\beta_i = \frac{y_i^T H g_{i+1}}{g_i^T H g_i}, \tag{PCGb}$$

nebo

$$\beta_i = \frac{g_{i+1}^T H g_{i+1}}{g_i^T H g_i}, \tag{PCGc}$$

nebo

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

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 (F4) - (F5), 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 (α), (γ), (δ) zůstanou beze změny, ale místo (β) platí

$$g_j^T H g_i = 0, \tag{β'}$$

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

3.2 Globální konvergence

Jak již bylo poznaménáno (poznámka 34, jsou metody (CGc) a (CGd) za jistých předpokladů globálně konvergentní. Nejprve dokážeme globální konvergenci metody (CGd). Větu zformulujeme tak, aby zahrnovala poněkud širší třídu metod sdružených gradientů.

Věta 21 *(Globální konvergence metody (CGd)). Nechť funkce $F : R^n \to R$ splňuje podmínky (F1) a (F3). Pak metoda sdružených gradientů (CG) s výběrem délky kroku splňujícím slabou Wolfeho podmínku (S2) a (S3b) je globálně konvergentní, pokud*

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

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

Důkaz (a) Dokážeme nejprve, že

$$|\beta_i| \le \frac{g_{i+1}^T s_{i+1}}{g_i^T s_i} \qquad (*)$$

$\forall i \in N$. Použijeme-li (CG) 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

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

neboť $y_i^T s_i > 0$ a pro $-(1-\varepsilon_2)/(1+\varepsilon_2) \le \lambda_i \le 1$ platí $(1-|\lambda_i| - (1-\lambda_i)\varepsilon_2) \ge 0$. Z $(*)$ 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 $(*)$ $g_{i+1}^T s_{i+1} \le |\beta_i| g_i^T s_i < 0$, pokud $\beta_i \ne 0$. Jestliže $\beta_i = 0$, dostaneme podle (CG) $g_{i+1}^T s_{i+1} = -g_{i+1}^T g_{i+1} < 0$.
(b) Zapíšeme-li (CG) ve tvaru $s_{i+1} + g_{i+1} = \beta_i s_i$, dostaneme umocněním, převedením dvou členů na pravou stranu a použitím nerovnosti $(*)$ vztah

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

neboli

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

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 uvedenou v poznámce 13.

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

Věta 22 *(Globální konvergence metody (CGc)). Nechť funkce $F : R^n \to R$ splňuje podmínky (F1) a (F3). Pak metoda sdružených gradientů (CG) s výběrem délky kroku splňujícím silnou Wolfeho podmínku (S2) a (S3a), kde $0 < \varepsilon_1 < \varepsilon_2 < 1/2$, je globálně konvergentní, pokud*

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

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

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

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

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 (CG) 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.
(b) Podle lemmatu 2 (vztah (b)) platí

$$F_i - F_{i+1} \geq \frac{\varepsilon_1 \varepsilon_3}{\overline{G}} \frac{(s_i^T g_i)^2}{\|s_i\|^2} = \frac{\varepsilon_1 \varepsilon_3}{\overline{G}} \frac{(s_i^T g_i)^2}{\|g_i\|^4} \frac{\|g_i\|^4}{\|s_i\|^2}$$

a použitím pravé části Al-Baaliho nerovnosti dostaneme

$$-\frac{s_i^T g_i}{\|g_i\|^2} \geq 1 - \frac{\varepsilon_2}{1-\varepsilon_2} = \frac{1-2\varepsilon_2}{1-\varepsilon_2} > 0,$$

.
neboť $0 < \varepsilon_2 < 1/2$. Platí tedy

$$F_1 - \underline{F} \geq \lim_{i\to\infty}(F_1 - F_{i+1}) = \sum_{i=1}^{\infty}(F_i - F_{i+1}) \geq \frac{\varepsilon_1\varepsilon_3(1-2\varepsilon_2)^2}{\overline{G}(1-\varepsilon_2)^2}\sum_{i=1}^{\infty}\frac{\|g_i\|^4}{\|s_i\|^2},$$

takže

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

(c) Z (S3a) plyne (S3b). Použijeme-li (S3b) a levou část Al-Baaliho nerovnosti, 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 (CG), můžeme psát

$$\begin{aligned}
\|s_{i+1}\|^2 &\leq \|g_{i+1}\|^2 + 2|\beta_i||s_i^T g_{i+1}| + \beta_i^2\|s_i\|^2 \\
&\leq \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2}|\beta_i|\|g_i\|^2 + \beta_i^2\|s_i\|^2 \\
&= \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2}|\lambda_i|\|g_{i+1}\|^2 + \lambda_i^2\frac{\|g_{i+1}\|^4}{\|g_i\|^4}\|s_i\|^2 \\
&\leq \|g_{i+1}\|^2 + \frac{2\varepsilon_2}{1-\varepsilon_2}\|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{\|g_i\|^4}\|s_i\|^2 \\
&= \frac{1+\varepsilon_2}{1-\varepsilon_2}\|g_{i+1}\|^2 + \frac{\|g_{i+1}\|^4}{\|g_i\|^4}\|s_i\|^2,
\end{aligned}$$

neboť $|\lambda_i| \leq 1$. Předpokládejme, že neplatí

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

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

$$\frac{\|s_{i+1}\|^2}{\|g_{i+1}\|^4} \leq \frac{1+\varepsilon_2}{1-\varepsilon_2}\frac{1}{\underline{\varepsilon}^2} + \frac{\|s_i\|^2}{\|g_i\|^4} \leq \frac{1+\varepsilon_2}{1-\varepsilon_2}\frac{i+1}{\underline{\varepsilon}^2}$$

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

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

což je spor, neboť podle (b) je tento součet konečný.

Poznámka 38 Věta 22 vyžaduje silnější předpoklady než věta 21. Je třeba, aby byla splněna silná Wolfeho podmínka a aby navíc platilo $\varepsilon_2 < 1/2$. Samotnou Al-Baaliho nerovnost však můžeme použít i za poněkud slabších předpokladů. Stačí aby platilo $|\lambda_i| \leq \overline{\varepsilon}_2/\varepsilon_2$, kde $0 < \varepsilon_2 < \overline{\varepsilon}_2 < 1/2$. V tomto případě můžeme psát

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

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

Poznámka 39 Jak již bylo zmíněno v poznámce 34, dávají metody (CGa) a (CGb) lepší praktické výsledky než metody (CGc) a (CGd). Vlastnosti metod (CGa) a (CGb) lze ještě zlepšit tím, že pokládáme $\beta = 0$ (restart) pokud neplatí

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

kde $0 < \varepsilon_2 < \overline{\varepsilon}_2 < 1/2$. Pak je podle poznámky 38 splněna Al-Baaliho nerovnost a výsledná metoda je metodou spádových směrů.

3.3 Přerušované metody sdružených gradientů

Poznámka 40 Metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce po nejvýše n krocích (věta 20). Neplatí to však jestliže:

- Výběr délky kroku není přesný.
- Funkce není kvadratická.
- Hessova matice je špatně podmíněná a projevují se zaokrouhlovací chyby.

Pak je třeba pokračovat ve výpočtu. Aby byly i nadále splněny předpoklady věty 20, je třeba iterační proces přerušit ($s_{n+1} = -g_{n+1}$). V dalších úvahách se budeme zabývat cyklicky přerušovanými metodami sdružených gradientů.

Definice 25 *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 (GCa)–(GCd) pro $i \notin M$, kde $M = \{l \in N : l = nk + 1, k \in N\}$.*

Poznámka 41 Definice 25 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 39, je metodou stejnoměrně spádových směrů a platí $s_i \sim g_i$.

Věta 23 *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 (S2) a (S3a), přičemž platí*

$$0 < \beta_i \leq \frac{\overline{\varepsilon}_2}{\varepsilon_2} \frac{\|g_{i+1}\|^2}{\|g_i\|^2}, \qquad (*)$$

kde $0 < \varepsilon_2 < \overline{\varepsilon}_2 < 1/2$. Nechť $x_i \to x^*$, kde $x^* \in R^n$ je stacionárním bodem funkce $F : R^n \to R$ vyhovující podmínkám (F3) a (F4). 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 $(*)$, můžeme použít Al-Baaliho nerovnost ve tvaru uvedeném v poznámce 38.
(a) Zřejmě $\|e_i\| = O(\|e_{i-1}\|)$ (poznámka 23) 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 $(*)$ 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

$$\frac{\|s_i\|}{\|g_i\|} \le 1 + \frac{\overline{\varepsilon}_2}{\varepsilon_2}\frac{\|g_i\|}{\|g_{i-1}\|}\frac{\|s_{i-1}\|}{\|g_{i-1}\|} \le 1 + c\frac{\|s_{i-1}\|}{\|g_{i-1}\|}.$$

Nechť $k = \sup\{j \in M, j \le i\}$. Protože $s_k = -g_k$, platí $\|s_k\|/\|g_k\| = 1$, takže rekurentním použitím poslední nerovnosti dostaneme

$$\frac{\|s_i\|}{\|g_i\|} \le \sum_{j=0}^{i-k} c^j \le \sum_{j=0}^{n} c^j \triangleq \overline{c}.$$

(b) Použijeme-li Al-Baaliho nerovnost (levou část) dostaneme

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

což spolu s (a) dává

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

takže $-s_i^T g_i \ge \varepsilon_0 \|s_i\|\|g_i\|$ kde $\varepsilon_0 = (1-2\overline{\varepsilon}_2)/(\overline{c}(1-\overline{\varepsilon}_2)) > 0$.
(c) Použitím Al-Baaliho nerovnosti a Schwartzovy nerovnosti dostaneme

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

což dává $\|s_i\| \ge \underline{c}\|g_i\|$, kde $\underline{c} = (1-2\overline{\varepsilon}_2)/(1-\overline{\varepsilon}_2) > 0$. Jelikož z (a) plyne $\|s_i\| \le \overline{c}\|g_i\|$, 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 $e_i \ne 0$ a $g_i \ne 0$ $\forall i \in N$, neboť v opačném případě iterační proces končí v minimu. Dále budeme předpokládat, že

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

Pokud pro nějaký index $l \le i < l+n$ neplatí $\|e_i\| \sim \|e_l\|$, pak nutně $\|e_i\| = o(\|e_l\|)$ (jelikož podle poznámky 23 je $\|e_i\| = O(\|e_l\|)$), takže rychlost konvergence je vyšší než lineární (tato úvaha je precizována ve větě 25).

Věta 24 *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 $F \in C^2 : R^n \to R$ vyhovující podmínkám (F3) a (F4). Nechť $\|e_i\| \sim \|e_l\|$ $\forall l \in M$, $\forall l \le i < l+n$. Pak pro $i \in N$ platí*

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

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

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

Důkaz (Pro CGa). Poznamenejme, že předpokládáme, že $e_i \ne 0$ a $g_i \ne 0$ pro $l \le i < l+n$. Důkaz věty provedeme indukcí. Dokážeme navíc, že pro $l \le j < i < l+n$ platí

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

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

$$s_j^T G^* s_i = o(\|e_l\|^2), \qquad (\tilde{\gamma})$$

$$s_j^T y_i = y_j^T s_i = o(\|e_l\|^2), \qquad (\tilde{\delta})$$

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 3), 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 $l \le i < l + n - 1$.
(a) Podle indukčních předpokladů $(\tilde{\alpha})$ a $(\tilde{\delta})$ platí

$$s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = o(\|e_l\|^2)$$

pro $l \le j < i$. Z asymptotické přesnosti výběru délky kroku a z $(\tilde{\tau})$ plyne, že

$$s_i^T g_{i+1} = s_i^T g_i o(1) = o(\|g_i\|^2) = o(\|e_i\|^2) = o(\|e_l\|^2).$$

Platí tedy $s_j^T g_{i+1} = o(\|e_l\|^2)$ pro $l \le j \le i$.
(b) Zřejmě

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

takže podle (a) a $(\tilde{\rho})$ platí

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

(c) 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 $(\tilde{\tau})$ 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 $(\tilde{\rho})$ a $(\tilde{\sigma})$ platí

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

a z asymptotické přesnosti výběru délky kroku a z $(\tilde{\rho})$ a $(\tilde{\tau})$ plyne, že

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

(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 3), 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 $(\tilde{\rho})$ a $(\tilde{\tau})$ 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 $(\tilde{\rho})$, $(\tilde{\delta})$ 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 (CGa) 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 $(\tilde{\sigma})$ 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 $(\tilde{\sigma})$ platí $s_j \sim g_j \sim e_l$ a podle (d) je $\alpha_{i+1} \sim 1$.

Poznámka 42 Je-li kromě podmínek (F3) a (F4) splněna i podmínka (F5), 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 43 Podle věty 24 (vztah $(\tilde{\rho})$) existuje pro cyklicky přerušovanou metodu sdružených gradientů s asymptoticky přesným výběrem délky kroku index $\underline{l} \in M$ takový, že nerovnost uvedená v poznámce 39 je splněna pro $i \geq \underline{l}$. Pak již nedochází k přerušování iteračního procesu vlivem porušení této nerovnosti a platí beze zbytku definice 25. 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 26 *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 6 *Nechť jsou splněny předpoklady věty 24. Nechť $\bar{x}_i \in R^n$, $i \in N$, je referenční posloupnost z definice 26 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 24. 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 3, můžeme psát

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

neboť $\bar{\alpha}_l = -\bar{s}_l^T \bar{g}_l / \bar{s}_l^T G^* \bar{s}_l$ 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 24 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 (F3) a (F4) plyne

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

Můžeme tedy psát

$$\bar{\beta}_i = \frac{\bar{y}_i^T \bar{g}_{i+1}}{\bar{s}_i^T \bar{y}_i} = \frac{y_i^T g_{i+1}(1 + o(1))^2}{s_i^T y_i(1 + o(1))^2} = \frac{y_i^T g_{i+1}}{s_i^T y_i}(1 + o(1)) = \beta_i(1 + o(1)).$$

(c) Podle (b) a indukčních předpokladů platí

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

(d) Podle lemmatu 3 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 $l \leq i < l + n$. Pak podle indukčních předpokladů platí

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

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 44 Tvrzení lemmatu 6 platí, pokud $\|e_i\| \sim \|e_l\|$ pro $l \leq i < l+n$. Jestliže $\|e_i\| \sim \|e_l\|$ pouze pro $l \leq i < l+m$, kde $m < n$, můžeme psát $e_i - \bar{e}_i = o(\|e_l\|)$ pro $l \leq i \leq l+m$ (plyne to z induktivní povahy důkazu).

Věta 25 *(n-kroková superlineární konvergence) Nechť jsou splněny předpoklady věty 24. Pak platí*

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

Důkaz Ve větě 25 mlčky předpokládáme, že k přerušování iteračního procesu dochází vždy po n krocích (poznámka 43). Podle věty 20 víme, že metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce $Q(x)$ po $m \leq n$ 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 23), 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 6 (které podle poznámky 44 platí pro $l \leq i < l+m$), dostaneme

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

Jelikož $e_{l+n} = O(\|e_{l+m}\|)$ (poznámka 23), v obou případech platí $\|e_{l+n}\| = o(\|e_l\|)$, což dává tvrzení věty.

Poznámka 45 Podle věty 7 a poznámky 7 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 7 *Nechť jsou splněny předpoklady věty 20. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

$$\frac{Q(x_{i+1}) - Q(x^*)}{Q(x_1) - Q(x^*)} \leq \max_{1 \leq k \leq n} \left(\overline{P}_i(\lambda_k)\right)^2,$$

kde $\overline{P}_i(\lambda)$ je libovolný polynom stupně i takový, že $\overline{P}_i(0) = 1$, a λ_k, $1 \leq k \leq n$, jsou vlastní čísla matice G.

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

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

je Krylovův podprostor stupně j generovaný maticí G a vektorem g_1. Pro $j = 1$ je to zřejmé. Předpokládejme, že to platí pro $j = i-1$. Protože z $x_i = x_{i-1} + \alpha_{i-1}s_{i-1}$ plyne $g_i = g_{i-1} + \alpha_{i-1}Gs_{i-1}$ (vlastnost kvadratické funkce (Q)) a protože platí $g_{i-1} \in \mathcal{K}_{i-1}$ a $Gs_{i-1} \in \mathrm{span}(Gg_1, G^2g_1, \ldots, G^{i-1}g_1) \subset \mathcal{K}_i$ (indukční předpoklad), dostaneme $g_i \in \mathcal{K}_i$. Dále protože $s_i = -g_i + \beta_{i-1}s_{i-1}$ (CG) a protože platí $s_{i-1} \in \mathcal{K}_{i-1} \subset \mathcal{K}_i$ (indukční předpoklad) a $g_i \in \mathcal{K}_i$ (dokázaná inkluze), dostaneme $s_i \in \mathcal{K}_i$

(b) Podle (a) platí

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

kde $P^*_{i-1}(G)$ je určitý polynom stupně $i-1$ v G. Označme $\overline{P}^*_i = I + GP^*_{i-1}$ (takže $\overline{P}^*_i$ je stupně i a $\overline{P}^*_i(0) = 1$). Jelikož z důkazu věty 20 plyne, že $s_j^T g_{i+1} = 0 \ \forall 1 \leq j \leq i$, je

$$x_{i+1} = x^* + \overline{P}_i^*(G)(x_1 - x^*) = \arg \min_{x=x^*+\overline{P}_i(G)(x_1-x^*)} Q(x),$$

takže

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

pro libovolný polynom $\overline{P}_i$ stupně i takový, že $\overline{P}_i(0) = 0$. 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

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

Po vydělení dostaneme

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

Věta 26 *Nechť jsou splněny předpoklady věty 20. Nechť $g_i \neq 0$ pro nějaký index $1 \leq i \leq n$. Pak platí*

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

Důkaz Podle lemmatu 7 platí

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

pro libovolný polynom $\overline{P}_i(\lambda)$ stupně nanejvýš i takový, že $\overline{P}_i(0) = 1$. Zvolíme polynom $\overline{P}_i(\lambda)$ tak, aby minimalizoval hodnotu

$$\max_{\lambda_1\leq\lambda\leq\lambda_n} |\overline{P}_i(\lambda)|.$$

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

$$\overline{P}_i(\lambda) = \frac{T_i\left(\frac{\lambda_n+\lambda_1-2\lambda}{\lambda_n-\lambda_1}\right)}{T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right)},$$

kde $T_i(\xi) = \cos(i \arccos \xi)$ pro $|\xi| \leq 1$ a $T_i(\xi) = ((\xi + \sqrt{\xi^2-1})^i + (\xi - \sqrt{\xi^2-1})^i)/2$ pro $|\xi| \geq 1$. Jelikož $|T_i(\xi)| \leq 1$ pro $|\xi| \leq 1$, platí

$$\max_{\lambda_1 \leq \lambda \leq \lambda_n} |\overline{P}_i(\lambda)| \leq 1/T_i\left(\frac{\lambda_n+\lambda_1}{\lambda_n-\lambda_1}\right).$$

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 hodnotu ξ, 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

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

Poznámka 46 Použijeme-li odhad z věty 26 spolu s (c) a (d) z důkazu věty 13, 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 47 Větu 26 lze snadno zobecnit tak, aby platila pro předpodmíněnou metodu sdružených gradientů. Podle poznámky 37 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 27 *(Asymptotický odhad) Nechť jsou splněny předpoklady věty 24. 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, $l \le i < l+n$, 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 $i \ge l$ docházelo k přerušování iteračního procesu vždy po n krocích. Nechť $M = 2\sqrt{\kappa(G^*)}$ a q je kvocient uvedený ve větě 27. Pak podle poznámky 46 pro $l \le i \le l+n$ platí

$$\|\bar{x}_i - x^*\| \le Mq^{i-l}\|\bar{x}_l - x^*\| = Mq^{i-l}\|x_l - x^*\|.$$

Použijeme-li lemma 6, můžeme pro $l \le i \le l+n$ psát

$$\|x_i - \bar{x}_i\| = o(\|x_l - x^*\|) = \|x_l - x^*\|o(1).$$

Platí tedy

$$\frac{\|x_i - x^*\|}{\|x_l - x^*\|} \le \frac{\|\bar{x}_i - x^*\|}{\|x_l - x^*\|} + \frac{\|x_i - \bar{x}_i\|}{\|x_l - x^*\|} \le Mq^{i-l} + o(1).$$

Poznámka 48 Věta 27 se týká pouze vnitřních iterací každého cyklu. Celkově je cyklicky přerušovaná metoda sdružených gradientů s přesným výběrem délky kroku R-superlineárně konvergentní (poznámka 45).

Poznámka 49 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 Implementace metod sdružených gradientů

Poznámka 50 Uvedeme několik poznámek k implementaci metod sdružených gradientů.

- Z tabulky uvedené za algoritmem 2 je zřejmé, že je účelné používat metody (CGa) a (CGb), které jsou mnohem efektivnější než metody (CGc) a (CGd). Pro rozsáhlé úlohy ($n > 500$) je výhodné tyto metody cyklicky přerušovat vždy po n iteračních krocích.
- Používáme-li metody (CGa) a (CGb), je výhodné zaručit platnost nerovnosti (*) z věty 23. V tomto případě se iterační proces přeruší, pokud neplatí $0 < \beta_i < \eta_1\|g_{i+1}\|^2/\|g_i\|^2$, kde $\eta_1 \le 1/(2\varepsilon_2)$.
- Používáme-li metody (CGc) a (CGd), je třeba testovat sdruženost směrů. V tomto případě se iterační proces přeruší, pokud
$$s_{i+1}^T y_i \ge \eta_2\|s_{i+1}\|\|y_i\|,$$
kde $\eta_2 \approx 0.04 - 0.05$. Také je možné testovat ortogonalitu gradientů. V tomto případě se iterační proces se přeruší, pokud
$$g_{i+1}^T g_i \ge \eta_3\|g_{i+1}\|\|g_i\|.$$
kde $\eta_3 \approx 0.4 - 0.5$.

- Je výhodné volit počáteční délku kroku podle vzorce

$$\alpha = \min\left(1, \frac{2(F_i - F_{i-1})}{s_i^T g_i}, \frac{2(\underline{F} - F_i)}{s_i^T g_i}\right),$$

kde $\underline{F}$ je dolní odhad pro minimální hodnotu funkce F.

- Místo slabé Wolfeho podmínky (S3b) je třeba použít silnou Wolfeho podmínku (S3a), kde $\varepsilon_2 = 10^{-1}$. Algoritmus 1 je třeba pozměnit tak, že v něm ponecháme podmínku (S3b) ale k podmínce (S2) přidáme podmínku

$$s_i^T g_{i+1} \leq \varepsilon_2 |s_i^T g_i|,$$

která je částí podmínky (S3a).

- často je výhodné metodu sdružených gradientů škálovat (nejjednodušší předpodmínění). Místo

$$-s_{i+1} = g_{i+1} - \beta_i s_i$$

se použije vzorec

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

kde

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

a ($y_i = g_{i+1} - g_i, d_i = x_{i+1} - x_i = \alpha_i s_i$). V tomto případě však musíme vzorce pro parametr β_i upravit tak, že

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

$$\beta_i = \frac{1}{\gamma_i} \frac{y_i^T g_{i+1}}{g_i^T g_i}, \tag{CGb}$$

$$\beta_i = \frac{1}{\gamma_i} \frac{g_{i+1}^T g_{i+1}}{g_i^T g_i}, \tag{CGc}$$

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

Parametr γ_{i+1}, je nutné udržovat v určitých mezích ($0.005 \leq \gamma_{i+1} \leq 200$).

Algoritmus metody sdružených gradientů lze popsat zhruba takto:

Algoritmus 2 (CG) Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-1}$, $\eta_1 = 1.2$, $\eta_2 = 0.05$, $\eta_3 = 0.46$, $\underline{\varepsilon} > 0$.

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 koeficient β_{i-1} podle některé z metod (CGa)–(CGd) a rozhodneme o přerušení iteračního procesu podle některé ze strategií uvedených v poznámce 50. Určíme škálovací koeficient γ_i (obvykle $\gamma_i = 1$) a určíme směrový vektor s_i.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1 upraveného podle poznámky 50. Položíme $x_{i+1} = x_i + \alpha_i s_i$, vypočteme $F_{i+1} = F(x_{i+1})$, $g_{i+1} = g(x_{i+1})$, zvětšíme i o 1 a přejdeme na krok 2.

Následující tabulka ukazuje srovnání jednotlivých metod sdružených gradientů při minimalizaci 22 testovacích funkcí se 1000 proměnnými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i celkový čas výpočtu). Označení η_1 nebo η_2 značí, že byla použita podmínka s η_1 nebo η_2, uvedená v poznámce 50.

Metoda	NIT-NFV	selhání	čas
CGa	20389 - 42871	-	22.65
CGa + $\eta 1$	19342 - 38730	-	30.99
CGb	20759 - 41836	-	22.20
CGb + $\eta 1$	19012 - 38125	-	21.93
CGc	38917 - 77918	2	22.34
CGc + $\eta 2$	24762 - 49842	-	22.55
CGd	38487 - 77044	2	36.28
CGd + $\eta 2$	24887 - 50122	-	22.26

3.6 Předpodmíněná metoda sdružených gradientů pro řešení soustav lineárních rovnic

Podle vět 20 a 26 je metoda sdružených gradientů zvláště vhodná k hledání minima ryze konvexní kvadratické funkce nebo, což je totéž, pro řešení soustavy lineárních rovnic se symetrickou positivně definitní maticí. Nyní budeme uvažovat kvadratickou funkci

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

která se používá k určení směrového vektoru v metodách spádových směrů (i v metodách s lokálně omezeným krokem popsaných v další kapitole). V poznámce 37 jsme ukázali, že metodu sdružených gradientů lze předpodmínit tak, že se místo kvadratické funkce $Q(s)$ minimalizuje kvadratická funkce

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

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 37 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 26 zaručuje rychlou konvergenci metody.

Algoritmus metody sdružených gradientů pro řešení soustavy lineárních rovnic s positivně definitní maticí B a s předpodmiňovačem C používá rekurentní vztahy

$$s_1 = 0, \qquad g_1 = g, \qquad p_1 = -C^{-1} g$$

a (PCG)

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

pro $1 \le i \le n$.

Poznámka 51 Abychom nemuseli vyšetřovat zvlášť případy, kdy $i = 1$ a $i > 1$, budeme formálně předpokládat, že $\beta_0 = 0$ a $p_0 = 0$.

Poznámka 52 Podle (PCG) platí $g_i = B s_i + g$.

Poznámka 53 Hodnota $\alpha_i = g_i^T C^{-1} g_i / p_i^T q_i$ realizuje přesný výběr délky kroku, neboť podle (PCG) platí

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

Poznámka 54 Některé vlastnosti nepředpodmíněné metody sdružených gradientů (algoritmus (PCG) s $C = I$) jsou ukázány ve větě 20. Připomeňme zde nejdůležitější vztahy

$$p_j^T g_i = 0, \tag{α}$$

$$g_j^T g_i = 0, \tag{β}$$

$$p_j^T B p_i = 0, \tag{γ}$$

které platí pro $1 \leq j < i \leq m+1$, kde m je index takový, že $\|g_m\| > 0$. Vztahy pro předpodmíněnou metodu sdružených gradientů dostaneme formálně tak, že místo s, g, p, q a B dosazujeme $\tilde{s} = C^{1/2}s$, $\tilde{g} = C^{-1/2}g$, $\tilde{p} = C^{1/2}p$, $\tilde{q} = C^{-1/2}q$ a $\tilde{B} = C^{-1/2}BC^{-1/2}$. Vztahy (α) a (γ) zůstanou beze změny, ale vztah (β) je třeba nahradit vztahem

$$g_j^T C^{-1} g_i = 0, \tag{β'}$$

Tento postup budeme používat i nadále. Nejprve zformulujeme a dokážeme tvrzení pro $C = I$ a pak jako důsledek uvedeme tvrzení pro $C \neq I$.

Věta 28 *Aplikujeme-li algoritmus (PCG) s $C = I$ na kvadratickou funkci $Q(s)$ a platí-li $g_j^T g_j > 0$ a $p_j^T B p_j > 0$ pro $1 \leq j \leq i$, pak*

$$Q(s_{i+1}) < Q(s_i), \tag{ρ}$$

$$\|s_{i+1}\| > \|s_i\|, \tag{σ}$$

$$\frac{g^T s_{i+1}}{\|g\|\|s_{i+1}\|} > \frac{g^T s_i}{\|g\|\|s_i\|}. \tag{τ}$$

Důkaz (a) Použijeme-li vztahy (α)–(γ) a (PCG) s $C = I$, dostaneme

$$\begin{aligned}
Q(s_{i+1}) &= g^T(s_i + \alpha_i p_i) + \frac{1}{2}(s_i + \alpha_i p_i)^T B(s_i + \alpha_i p_i) \\
&= Q(s_i) + \alpha_i (g + Bs_i)^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i \\
&= Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i \\
&= Q(s_i) - \frac{(g_i^T g_i)^2}{p_i^T B p_i} + \frac{1}{2}\frac{(g_i^T g_i)^2}{p_i^T B p_i} \\
&= Q(s_i) - \frac{1}{2}\frac{(g_i^T g_i)^2}{p_i^T B p_i} < Q(s_i).
\end{aligned}$$

(b) Jelikož $s_i = \sum_{j=1}^{i-1} \alpha_j p_j$, platí

$$\begin{aligned}
s_{i+1}^T s_{i+1} &= (s_i + \alpha_i p_i)^T (s_i + \alpha_i p_i) = s_i^T s_i + \alpha_i^2 p_i^T p_i + 2\alpha_i s_i^T p_i \\
&= s_i^T s_i + \alpha_i^2 p_i^T p_i + 2\alpha_i \sum_{j=1}^{i-1} \alpha_j p_j^T p_i \\
&= s_i^T s_i + \alpha_i^2 p_i^T p_i + 2\alpha_i g_i^T g_i \sum_{j=1}^{i-1} \alpha_j \frac{p_j^T p_j}{g_j^T g_j} > s_i^T s_i,
\end{aligned}$$

neboť

$$\begin{aligned}
p_j^T p_i &= p_j^T(-g_i + \beta_{i-1} p_{i-1}) = \beta_{i-1} p_j^T p_{i-1} = \\
&= \left(\prod_{k=j}^{i-1} \beta_k\right) p_j^T p_j = \frac{g_i^T g_i}{g_j^T g_j} p_j^T p_j.
\end{aligned}$$

(c) Použijeme-li (PCG) s $C = I$, můžeme pro $1 \le j \le i$ 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

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

Použijeme-li (β), dostaneme

$$-g^T s_{i+1} = \sum_{k=1}^{i} \alpha_k \|g_k\|^2 > 0$$

a

$$s_{i+1}^T s_{i+1} = \sum_{j=1}^{i} \left(\sum_{k=j}^{i} \alpha_k \|g_k\|^2 \right)^2 \frac{1}{\|g_j\|^2},$$

takže

$$\frac{s_{i+1}^T s_{i+1}}{(g^T s_{i+1})^2} = \sum_{j=1}^{i} \left(\frac{\sum_{k=j}^{i} \alpha_k \|g_k\|^2}{\sum_{k=1}^{i} \alpha_k \|g_k\|^2} \right)^2 \frac{1}{\|g_j\|^2}.$$

Nyní použijeme toho, že racionální funkce $\varphi(t) = (a+t)/(b+t)$ je pro $a < b$ rostoucí (můžeme se o tom přesvědčit derivováním). Platí tedy

$$\frac{\sum_{k=j}^{i} \alpha_k \|g_k\|^2}{\sum_{k=1}^{i} \alpha_k \|g_k\|^2} > \frac{\sum_{k=j}^{i-1} \alpha_k \|g_k\|^2}{\sum_{k=1}^{i-1} \alpha_k \|g_k\|^2},$$

což po dosazení dává

$$\begin{aligned} \frac{s_{i+1}^T s_{i+1}}{(g^T s_{i+1})^2} &= \sum_{j=1}^{i-1} \left(\frac{\sum_{k=j}^{i} \alpha_k \|g_k\|^2}{\sum_{k=1}^{i} \alpha_k \|g_k\|^2} \right)^2 \frac{1}{\|g_j\|^2} + \left(\frac{\alpha_i \|g_i\|^2}{\sum_{k=1}^{i} \alpha_k \|g_k\|^2} \right)^2 \frac{1}{\|g_i\|^2} \\ &> \sum_{j=1}^{i-1} \left(\frac{\sum_{k=j}^{i-1} \alpha_k \|g_k\|^2}{\sum_{k=1}^{i-1} \alpha_k \|g_k\|^2} \right)^2 \frac{1}{\|g_j\|^2} = \frac{s_i^T s_i}{(g^T s_i)^2}. \end{aligned}$$

Bezprostředním použitím této nerovnosti dostaneme (τ).

Důsledek 2 *Aplikujeme-li algoritmus (PCG) s positivně definitní maticí C na kvadratickou funkci $Q(s)$ a platí-li $g_j^T C^{-1} g_j > 0$ a $p_j^T B p_j > 0$ pro $1 \le j \le i$, pak*

$$Q(s_{i+1}) < Q(s_i),$$

$$\|s_{i+1}\|_C > \|s_i\|_C, \qquad (\sigma')$$

$$\frac{g^T s_{i+1}}{\|g\|_D \|s_{i+1}\|_C} > \frac{g^T s_i}{\|g\|_D \|s_i\|_C}, \qquad (\tau')$$

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 54.

Věta 29 *Jsou-li splněny předpoklady věty 28, platí*

$$-Q(s_{i+1}) \geq \frac{\|g\|^2}{2\|B\|}.$$

Je-li navíc matice B positivně definitní, platí

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

Důkaz (a) Protože

$$s_2 = s_1 + \alpha_1 p_1 = \frac{g_1^T g_1}{p_1^T B p_1} p_1 = -\frac{g^T g}{g^T B g} g,$$

platí

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

takže podle (ρ) dostaneme $-Q(s_{i+1}) \geq -Q(s_2) \geq (1/2)\|g\|^2/\|B\|$.
(b) Jelikož B je positivně definitní, platí $p_j^T B p_j > 0$, kdykoliv $\|g_j\| > 0$, neboť (α) implikuje nerovnost $\|p_j\| > \|g_j\|$. Podle věty 20 existuje index $m \leq n$ takový, že $\|g_j\| > 0$ pro $1 \leq j \leq m$ a $g_{m+1} = 0$. Podle (τ) můžeme pro $i \leq m$ psát

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

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

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

Po dosazení do předchozí nerovnosti dostaneme dokazované tvrzení.

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

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

Je-li navíc matice B positivně definitní, platí

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

Důkaz (a) Podobně jako v důkazu věty 29 dostaneme

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

což spolu s (ρ) dává dokazovanou nerovnost.
(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 $1 \le i \le m$, můžeme podle (τ') psát

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

a jelikož vektor s_{m+1} je řešením soustavy $g + Bs = 0$, můžeme použít stejnou nerovnost jako v důkazu věty 29.

Poznámka 55 Předpokládáme-li, že matice C je vybrána tak, že $\kappa(\tilde{B}) \le \kappa(B)$, můžeme druhou nerovnost v důsledku 3 nahradit nerovností

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

neboť podle věty 29 platí

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

Není-li matice B positivně definitní, mohou nastat problémy. Jednak může být $p_i^T B p_i \approx 0$, což vede k selhání algoritmu ($\alpha_i \approx \infty$ a $\|s_{i+1}\| \approx \infty$), nebo platí $p_i^T B p_i < 0$, což může vést k porušení podmínky spádovosti $g^T s_{i+1} < 0$. Proto je třeba výpočet ukončit pokud neplatí $p_i^T B p_i \ge \underline{c}\, p_i^T p_i$, kde $\underline{c}$ je zvolená dolní mez. Jelikož tato podmínka nemusí být splněna pro všechny indexy $1 \le i \le m$, nemůžeme použít druhou nerovnost v důsledku 3. Platí však tato věta

Věta 30 *Aplikujeme-li algoritmus (PCG) s positivně definitní maticí C na kvadratickou funkci $Q(s)$ a platí-li $p_j^T B p_j \ge \underline{c}\, p_j^T p_j$ pro $1 \le j \le i$, pak*

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

Důkaz Použijeme-li vztahy (α)–(γ) a (PCG), 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 $1 \le j \le i$, takže

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

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

$$\|s_{i+1}\| \le \sum_{j=1}^{i} \frac{\|p_j p_j^T\|}{p_j^T B p_j} \|g\| = \sum_{j=1}^{i} \frac{p_j^T p_j}{p_j^T B p_j} \|g\| \le \frac{n}{\underline{c}} \|g\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$-\frac{g^T s_{i+1}}{\|s_{i+1}\|\|g\|} \ge \frac{\|g\|^2}{\kappa(C)\|B\|} \frac{\underline{c}}{n\|g\|^2} = \frac{\underline{c}}{n\kappa(C)\|B\|}.$$

Poznámka 56 Je-li matice B positivně 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 uvedený v důsledku 3.

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

Algoritmus 3 (PCG) 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 $\tau < \underline{c}$, 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.

4 Metody s proměnnou metrikou

4.1 Základní vlastnosti metod s proměnnou metrikou

Definice 27 *Řekneme, že základní optimalizační metoda je metodou s proměnnou metrikou, jestliže*

$$s_i = -H_i g_i \qquad \forall i \in N, \tag{VM1}$$

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

$$H_{i+1} = \gamma_i (H_i + U_i M_i U_i^T), \tag{VM2}$$

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

$$H_{i+1} y_i = \rho_i d_i, \tag{VM3}$$

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

Poznámka 57 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 (VM3) se nazývá (zobecněnou) kvazinewtonovskou podmínkou. Předpokládáme, že $d_i \neq 0$ a $y_i \neq 0$, neboť v opačném případě nemá podmínka (VM3) smysl.

Věta 31 *(Kvadratické ukončení) Nechť x_i, $i \in N$, je posloupnost generovaná metodou s proměnnou metrikou z Broydenovy třídy s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0$ $\forall i \in N$) aplikovaná na ryze konvexní kvadratickou funkci (Q). Pak existuje index $k \leq n$ tak, že $g_{k+1} = 0$ a $x_{k+1} = x^*$.*

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

$$H_i y_j = \lambda_i^j d_j \qquad \forall 1 \leq j < i \leq n+1, \tag{α}$$

$$s_j^T g_i = 0 \qquad \forall 1 \leq j < i \leq n+1, \tag{β}$$

$$s_j^T G s_i = 0 \qquad \forall 1 \leq j < i \leq n. \tag{γ}$$

Z (VM1) a (γ) plyne, že s_i, $1 \leq i \leq n$, jsou nenulové a vzájemně sdružené (G-ortogonální), tudíž lineárně nezávislé, takže podle (β) nutně $g_{n+1} = 0$. Pro $i = 1$ platí $s_1^T g_1 = -s_1^T H_1 s_1 < 0$ (H_1 je SPD) takže $s_1 \neq 0$ a $\alpha_1 \neq 0$ a dále není co dokazovat. Indukční krok:

(a) Nechť $i \leq n$. Z (γ) a (Q) plyne $d_i^T y_j = d_i^T G d_j = \alpha_i \alpha_j s_i^T G s_j = 0$ a (α) navíc dává $y_i^T H_i y_j = \lambda_i^j y_i^T d_j = \lambda_i^j d_i^T G d_j = 0$, takže $U_i^T y_j = 0$ $\forall 1 \leq j < i$. Podle (VM2) a (α) tedy platí

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

$\forall 1 \leq j < i$. Použijeme-li (VM3) dostaneme $H_{i+1} y_i = \rho_i d_i \triangleq \lambda_{i+1}^i d_i$, takže $H_{i+1} y_j = \lambda_{i+1}^j d_i$ $\forall 1 \leq j \leq i$.

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

(c) Podle (VM1) 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 (VM1), (Q), (a), (b) dostaneme

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

$\forall 1 \leq j \leq i$.

Věta 32 *(Aproximace Hessovy matice). Nechť jsou splněny předpoklady věty 31 s $\gamma_i = 1$ a $\rho_i = 1$ $\forall i \in N$. Pak platí $H_{n+1} = G^{-1}$.*

Důkaz Z důkazu věty 31 plyne, že

$$H_{n+1} y_j = d_j \qquad \forall 1 \leq j \leq n$$

a že vektory d_j a $y_j = G d_j$, $1 \leq j \leq n$, jsou lineárně nezávislé. Z tohoto důvodu musí platit $H_{n+1} = G^{-1}$.

Nyní se budeme zabývat vyšetřováním aktualizace (VM2). Pro zjednodušení budeme index i vynechávat a index $i+1$ nahradíme symbolem $+$. Nejprve uvedeme několik pomocných tvrzení týkajících se aktualizací.

Lemma 8 *Nechť $U \in R^{n \times m}$, $V \in R^{n \times m}$ jsou matice s lineárně nezávislými sloupci (takže $m \leq n$). Pak:*

(a) Matice UV^T má stejná nenulová vlastní čísla jako matice $V^T U$.

(b) Matice $I + UV^T$ má stejná nejednotková vlastní čísla jako matice $I + V^T U$.

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

(d) Je-li matice $I + UV^T$ regulární, platí $(I + UV^T)^{-1} = I - U(I + V^T U)^{-1} V^T$.

Důkaz (a) Nechť $UV^T x = \lambda x$, $x \neq 0$ a $\lambda \neq 0$. Pak nutně $V^T x \neq 0$ a můžeme psát $V^T U V^T x = \lambda V^T x$, neboli $V^T U y = \lambda y$, kde $y = V^T x \neq 0$. Nechť naopak $V^T U y = \lambda y$, $y \neq 0$ a $\lambda \neq 0$. Pak nutně $Uy \neq 0$ a můžeme psát $UV^T U y = \lambda U y$, neboli $UV^T x = \lambda x$, kde $x = Uy \neq 0$.
(b) Zřejmě $(I + UV^T)x = \lambda x$ právě tehdy, jestliže $UV^T x = (\lambda - 1)x$, a $(I + V^T U)y = \lambda y$ právě tehdy, jestliže $V^T U y = (\lambda - 1)y$. Tvrzení (b) tedy plyne z (a).
(c) Determinant matice je roven součinu jejích vlastních čísel. Tvrzení (c) tedy plyne z (b).
(d) Platí

$$(I + UV^T)(I - U(I + V^T U)^{-1} V^T) = I + UV^T - U(I + V^T U)(I + V^T U)^{-1} V^T = I.$$

Důsledek 4 *(Woodbury) Nechť jsou splněny předpoklady lemmatu 8 a nechť $H \in R^{n \times n}$ je regulární matice. Pak*

$$\det(H + UV^T) = \det H \ \det(I + V^T H^{-1} U)$$

a

$$(H + UV^T)^{-1} = H^{-1} - H^{-1} U (I + V^T H^{-1} U)^{-1} V^T H^{-1}.$$

Důkaz Platí $H + UV^T = H(I + H^{-1} UV^T)$, takže můžeme použít (c) a (d) z lemmatu 8 (matice U se nahradí maticí $H^{-1}U$).

Důsledek 5 *Nechť $U \in R^{n \times m}$ je matice s lineárně nezávislými sloupci (takže $m \leq n$) a $M \in R^{m \times m}$ je symetrická regulární matice. Pak:*

(a) Matice UMU^T má stejná nenulová vlastní čísla jako matice $MU^T U$ (nebo jako matice $U^T UM$).

(b) Matice $I + UMU^T$ má stejná nejednotková vlastní čísla jako matice $I + MU^T U$ (nebo jako matice $I + U^T UM$).

(c) Platí $\det(I + UMU^T) = \det(I + MU^T U) = \det(I + U^T UM) = \det M \ \det(M^{-1} + U^T U)$.

(d) Je-li matice $I + UMU^T$ regulární, platí $(I + UMU^T)^{-1} = I - U(M^{-1} + U^T U)^{-1} U^T$.

Důkaz Stačí v lemmatu 8 použít UM místo V (nebo UM místo U a U místo V).

Důsledek 6 *Nechť jsou splněny předpoklady důsledku 5 a nechť $H \in R^{n\times n}$ je symetrická positivně definitní matice. Pak*

$$\det(H + UMU^T) = \det H \, \det M \, \det(M^{-1} + U^T H^{-1} U)$$

a

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

Důkaz Platí $H + UMU^T = H^{1/2}(I + H^{-1/2}UMU^T H^{-1/2})H^{1/2}$, takže můžeme použít (c) a (d) z důsledku 5 (matice U se nahradí maticí $H^{-1/2}U$).

Nyní se vrátíme k vyšetřování aktualizace (VM2). Budeme předpokládat, že vektory d a Hy jsou lineárně nezávislé. Jsou-li tyto vektory lineárně závislé, má matice UMU^T hodnost 1 a všechny aktualizace z Broydenovy třídy jsou ekvivalentní (dávají stejnou matici H_+). V tomto případě je výhodné používat metodu hodnosti 1. Kvazinewtonovskou podmínku $H_+ y = \rho d$ můžeme v tomto případě splnit prostým vynásobením matice H číslem $\gamma = \rho\, y^T d / y^T H y$.

Věta 33 *Nechť $H_+ = \gamma(H + UMU^T)$, kde H je SPD matice a $U = [d, Hy]$. Pak $H_+ y = \rho d$ platí právě tehdy, jestliže*

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

kde η je volný parametr a kde

$$a = y^T H y, \qquad b = y^T d, \qquad c = d^T H^{-1} d.$$

Důkaz Podle (VM2) a (VM3) musí platit

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

takže nutně

$$m_1 b + m_2 a = \rho/\gamma,$$

$$m_2 b + m_3 a = -1.$$

Jeden parametr je nadbytečný. Zvolíme $m_2 = -\eta/b$ a zbylé prvky m_1, m_3 určíme řešením soustavy. Tím dostaneme matici M uvedenou ve větě 33.

Poznámka 58 Z pozitivní definitnosti matice H a z nenulovosti vektorů d a y plyne, že $a > 0$ a $c > 0$. Vybíráme-li délku kroku podle (S3b), platí

$$b = y^T d = \alpha(g_+ - g)^T s \geq \alpha(\varepsilon_2 - 1) g^T s > 0.$$

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 59 Při vyšetřování metod s proměnnou metrikou budeme často používat označení

$$\delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac-b^2)+b^2),$$

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

Přímým výpočtem se snadno přesvědčíme, že $\mu = -\det M$, kde M je matice vystupující ve větě 33. Podle poznámky 62 platí $\det((1/\gamma)H_+) = \delta \det H$.

Poznámka 60 Vztah $H_+ = \gamma(H + UMU^T)$ můžeme roznásobit. Pak platí

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

(Broydenova třída). Nejznámější členy Broydenovy třídy dostaneme, položíme-li $\eta = 0$ (metoda DFP):

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

nebo $\eta = 1$ (metoda BFGS):

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

nebo $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$ (metoda hodnosti 1):

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

nebo $\eta = (\rho/\gamma)/(\rho/\gamma + a/b)$ (Hoshinova metoda):

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

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 63), takže může dojít ke ztrátě globální konvergence vlivem porušení podmínky spádovosti (S1a).

Poznámka 61 Uvedeme několik dalších tvarů aktualizace (H), které se používají při implementaci nebo při teoretickém vyšetřování metod s proměnnou metrikou. Předně jsou to dvoučlenné symetrické vztahy

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

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

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

První je zobecněním vztahu pro metodu DFP, druhý se používá při implementaci metody BFGS, pro kterou dosazením $\eta = 1$ dostaneme

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

a třetí je zobecněním metody hodnosti 1. Pro konstrukci metod s omezenou pamětí je užitečný pseudosoučinový tvar

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

který lze použít pouze tehdy, pokud $\eta \geq 0$. O správnosti všech těchto vztahů se můžeme přesvědčit jejich roznásobením a porovnáním odpovídajících si členů.

Lemma 9 *Nechť H je SPD matice, $a > 0$, $b > 0$, $c > 0$ a nechť H_+ je matice získaná pomocí aktualizace (H), kde $\gamma > 0$ a $\rho > 0$. Pak matice $(1/\gamma)H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}}$ má $n-2$ jednotkových vlastních čísel a zbylá dvě vlastní čísla jsou řešením kvadratické rovnice.*

$$\lambda^2 - \sigma\lambda + \delta = 0,$$

kde

$$\sigma = \frac{1}{b^2}(\eta(ac-b^2)+b^2) + \frac{\rho}{\gamma}\frac{c}{b} = \frac{\rho}{\gamma}\frac{c}{b} + \left(1 - \frac{\eta}{\eta^*}\right),$$

$$\delta = \frac{\rho}{\gamma}\frac{1}{ab}(\eta(ac-b^2)+b^2) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right)$$

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 (VM2) platí

$$\frac{1}{\gamma}H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}} = I + H^{-\frac{1}{2}}UMU^TH^{-\frac{1}{2}}.$$

Tato matice má $n-2$ jednotkových vlastních čísel odpovídajících $n-2$ vlastním vektorům kolmým k $H^{-\frac{1}{2}}U$. Zbylá dvě vlastní čísla jsou podle důsledku 5 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 33 a pro $U^TH^{-1}U$ vztah

$$U^TH^{-1}U = \begin{bmatrix} c & b \\ b & a \end{bmatrix},$$

můžeme psát

$$\begin{aligned} \det((1-\lambda)I + MU^TH^{-1}U) &= \det\left(\begin{bmatrix} 1-\lambda, & 0 \\ 0, & 1-\lambda \end{bmatrix} + M \begin{bmatrix} c, & b \\ b, & a \end{bmatrix}\right) \\ &= \det\begin{bmatrix} \eta\frac{ac-b^2}{b^2} + \frac{\rho c}{\gamma b} + 1 - \lambda, & \frac{\rho}{\gamma} \\ -\eta\frac{ac-b^2}{ab} - \frac{b}{a}, & -\lambda \end{bmatrix} = 0, \end{aligned}$$

což po úpravě dává $\lambda^2 - \sigma\lambda + \delta = 0$ s koeficienty uvedenými v lemmatu 9.

Poznámka 62 Poznamenejme, že δ se jako součin vlastních čísel matice $I + MU^T H^{-1}U$ rovná (podle důsledku 5) determinantu matice $\frac{1}{\gamma}H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}} = I + H^{-\frac{1}{2}}UMU^TH^{-\frac{1}{2}}$. Platí tedy $\det((1/\gamma)H_+) = \delta \det H$, což lze zapsat ve tvaru.

$$\det\left(\frac{1}{\gamma}H_+\right) = \frac{\rho}{\gamma}\frac{b}{a}\left(1 - \frac{\eta}{\eta^*}\right)\det H$$

(pokud $ac - b^2 = 0$, můžeme položit $1/\eta^* = 0$).

Věta 34 *Nechť jsou splněny předpoklady lemmatu 9. Pak H_+ je SPD 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 9 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 63 Z věty 34 plyne, že matice H_+ je SPD, 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, jestiže 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 (H).

Věta 35 *(Aktualizace matice $B = H^{-1}$). Nechť jsou splněny předpoklady lemmatu 9. Nechť $B = H^{-1}$ a $B_+ = H_+^{-1}$. Pak platí*

$$\gamma B_+ = B + \frac{\gamma}{\rho}\frac{1}{b}yy^T - \frac{1}{c}Bd(Bd)^T + \frac{\beta}{c}\left(\frac{c}{b}y - Bd\right)\left(\frac{c}{b}y - Bd\right)^T, \tag{B}$$

kde

$$\beta\eta(ac - b^2) + (\beta + \eta)b^2 = b^2.$$

Důkaz Inverzí vztahu $\frac{1}{\gamma}H_+ = H + UMU^T$ podle důsledku 6 dostaneme

$$\gamma B_+ = B - BU(M^{-1} + U^TBU)^{-1}U^TB \triangleq B + BUKU^TB,$$

kde $K \in R^{2\times 2}$. Jelikož podle (VM3) platí $H_+y = \rho d$, musí platit $B_+d = (1/\rho)y$ neboli

$$\gamma B_+d = Bd + [Bd, y]\begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix}\begin{bmatrix} c \\ b \end{bmatrix} = Bd + (k_1c + k_2b)Bd + (k_2c + k_3b)y = \frac{\gamma}{\rho}y,$$

takže nutně

$$k_1c + k_2b = -1,$$
$$k_2c + k_3b = \gamma/\rho.$$

Zvolíme $k_2 = -\beta/b$ a zbylé prvky k_1, k_3 určíme řešením soustavy. 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 (B). Vztah svazující β s η lze získat například z rovnosti

$$K = -(M^{-1} + U^T BU)^{-1}.$$

Jednodušší způsob je uveden v poznámce 66.

Poznámka 64 (Dualita) Vztah (B) dostaneme ze vztahu (H) 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 $\beta = 1$:

$$\gamma B_+ = B + \left(\frac{c}{b} + \frac{\gamma}{\rho}\right)\frac{1}{b}yy^T - \frac{1}{b}\left(Bdy^T + y(Bd)^T\right). \tag{BD}$$

Metodu BFGS dostaneme pro $\beta = 0$:

$$\gamma B_+ = B + \frac{\gamma}{\rho}\frac{1}{b}yy^T - \frac{1}{c}Bd(Bd)^T. \tag{BB}$$

Metoda hodnosti 1 je samoduální, dostaneme ji pro $\beta = (\gamma/\rho)/(\gamma/\rho - c/b)$:

$$\gamma B_+ = B + \frac{1}{(\gamma/\rho)b - c}\left(\frac{\gamma}{\rho}y - Bd\right)\left(\frac{\gamma}{\rho}y - Bd\right)^T. \tag{BR}$$

Hoshinova metoda je také samoduální, dostaneme ji pro $\beta = (\gamma/\rho)/(\gamma/\rho + c/b)$:

$$\gamma B_+ = B + \frac{2c}{b^2}yy^T - \frac{1}{(\gamma/\rho)b + c}\left(\frac{\gamma}{\rho}y + Bd\right)\left(\frac{\gamma}{\rho}y + Bd\right)^T. \tag{BH}$$

Poznámka 65 Z duality plyne, že matice B_+ je pozitivně definitní právě tehdy, jestliže $\beta > \beta^*$, kde $\beta^* = \eta^* = -b^2/(ac - b^2) < 0$.

Poznámka 66 Z duality lze snadno určit vztah mezi β a η. Platí totiž

$$\det(\gamma B_+) = \frac{\gamma}{\rho}\frac{b}{c}\left(1 - \frac{\beta}{\beta^*}\right)\det B,$$

což spolu s výrazem pro $\det H_+$ (poznámka 62) a identitou $\det B_+ \det H_+ = 1$ dává

$$\frac{b^2}{ac}\left(1 - \frac{\beta}{\beta^*}\right)\left(1 - \frac{\eta}{\eta^*}\right) = 1.$$

Po roznásobení dostaneme rovnost uvedenou ve větě 35.

Poznámka 67 Z úvahy použité v poznámce 66 plyne, že při přechodu od vztahu (H) ke vztahu (B) provádíme záměnu $\delta \to 1/\delta$. Z důkazu věty 35 víme, že $-K^{-1} = M^{-1} + U^T BU$, což podle důsledku 6 dává $\det(K^{-1}) = \delta \det M$ (neboť $K \in R^{2\times 2}$, takže $\det(-K^{-1}) = \det K^{-1}$). Odtud plyne, že při přechodu od vztahu (H) ke vztahu (B) provádíme záměnu $\mu \to \mu/\delta$. Z duality plyne, že

$$\frac{1}{\delta} = \frac{\gamma}{\rho}\frac{1}{bc}(\eta(ac - b^2) + b^2),$$

$$\frac{\mu}{\delta} = \frac{1}{bc}\left(\eta\frac{c}{b} + (1-\eta)\frac{\rho}{\gamma}\right).$$

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

$$\frac{1}{\sqrt{\gamma}}S_+ = S + p\tilde{q}^T, \tag{S}$$

kde $p \in R^n$ a $\tilde{q} \in R^m$. Součinový tvar metod s proměnnou metrikou se používá zejména v případě, že $m \neq n$.

Poznámka 68 Existují dvě možnosti, buď

$$\frac{1}{\sqrt{\gamma}}S_+ = S(I + \tilde{p}\tilde{q}^T), \tag{Sa}$$

kde $\tilde{p} \in R^m$, nebo

$$\frac{1}{\sqrt{\gamma}}S_+ = (I + pq^T)S, \tag{Sb}$$

kde $q \in R^n$. První možnost je výhodná tam, kde je třeba zachovat $\mathcal{L}(S)$, například při minimalizaci s lineárními omezeními, kde podprostor $\mathcal{L}(S)$ je dán aktivními omezeními. V tomto případě musí existovat vektory $\tilde{d} \in R^m$ a $\tilde{p} \in R^m$ takové, že

$$d = S\tilde{d}, \qquad p = S\tilde{p}.$$

Druhá možnost se používá tam, kde je třeba měnit $\mathcal{L}(S)$, například u metod s proměnnou metrikou s omezenou pamětí. V tomto případě musí existovat vektory $y \in R^n$ a $q \in R^n$ takové, že

$$\tilde{y} = S^T y, \qquad \tilde{q} = S^T q.$$

Čísla a, b, c lze určit podle vzorců

$$a = \tilde{y}^T\tilde{y}, \qquad b = \tilde{y}^T\tilde{d}, \qquad c = \tilde{d}^T\tilde{d}.$$

Nejprve ukážeme, jak musí vypadat aktualizace (S), aby byla splněna kvazinewtonovská podmínka.

Lemma 10 *Uvažujme aktualizaci (S), s vektorem $\tilde{q}$ zvoleným tak, že*

$$D^2 \triangleq (\tilde{q}^T\tilde{y})^2 + (\frac{\rho}{\gamma}b - a)\tilde{q}^T\tilde{q} > 0,$$

kde $\tilde{y} = S^T y$ (pokud $\rho/\gamma \geq a/b$, může být vektor $\tilde{q}$ libovolný). Pak kvazinewtonovská podmínka $S_+S_+^T y = \rho d$ 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 $\tau = p^T y$ se vypočte z rovnosti $\tilde{q}^T\tilde{y} + \tau\tilde{q}^T\tilde{q} = D$. Pokud $\rho/\gamma = a/b$, lze volit $\tau = 0$.

Důkaz Použitím vztahu (S) 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

$$S\tilde{y} + p\tilde{q}^T\tilde{y} + S\tilde{q}p^T y + p\tilde{q}^T\tilde{q}p^T y = \frac{\rho}{\gamma}d,$$

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

$$D^2 = (\tilde{q}^T\tilde{y})^2 + (\frac{\rho}{\gamma}b - a)\tilde{q}^T\tilde{q}.$$

Toto číslo musí být kladné, což poněkud omezuje volbu vektoru $\tilde{q}$.

Nyní se budeme zabývat součinovým tvarem metod z Broydenovy třídy. Poznamenejme, že Broydenovu třídu lze vyjádřit v součinovém tvaru pouze tehdy, když $d \in \mathcal{L}(S)$ (takže $d = S\tilde{d}$), což platí například tehdy, je-li matice SS^T regulární (pak $\tilde{d} = S^T(SS^T)^{-1}d$). V opačném příadě můžeme sice všechny vztahy formálně použít, ale výsledná metoda nebude patřit do Broydenovy třídy.

Lemma 11 *Nechť $\tilde{U} = [\tilde{d}, \tilde{y}]$, kde $d = S\tilde{d}$ a $\tilde{y} = S^T y$ (takže $S\tilde{U} = U$). Uvažujme aktualizaci (S) kde $\tilde{q} = \tilde{U}\hat{q}$ a kde vektor $\hat{q} \in R^{2\times 2}$ je zvolen tak, že $D^2 = (\tilde{q}^T\tilde{y})^2 + ((\rho/\gamma)b - a)\tilde{q}^T\tilde{q} > 0$. Pak pro matici $H_+ = S_+S_+^T$ platí $H_+ = \gamma(H + UMU^T)$, kde $\delta = \det((1/\gamma)H_+)/\det H \geq 0$ a $\mu = -\det M \geq 0$.*

Důkaz Jestliže $\tilde{q} = \tilde{U}\hat{q}$ a $D^2 > 0$, pak podle lemmatu 10 platí

$$p = \frac{(\rho/\gamma)d - S(\tilde{y} + \tau\tilde{q})}{D} = \frac{(\rho/\gamma)d - SS^T y + \tau U\hat{q}}{D} \triangleq U\hat{p},$$

kde $\hat{p} \in R^{2\times 2}$ Dosadíme-li obě tato vyjádření do vztahu pro $(1/\gamma)S_+S_+^T$ uvedeného v důkazu lemmatu 10, můžeme psát

$$\frac{1}{\gamma}S_+S_+^T = SS^T + U\hat{p}\hat{q}^TU^T + U\hat{q}\hat{p}^TU^T + U\hat{p}\tilde{q}^T\tilde{q}\hat{p}^TU^T = SS^T + UMU^T,$$

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

$$\det M = -(\det[\hat{p}, \hat{q}])^2 \leq 0.$$

Dále platí $SS^T + UMU^T = S(I + \tilde{U}M\tilde{U}^T)S^T$ a použijeme-li (Sb), můžeme psát $SS^T + UMU^T = S(I + \tilde{p}\tilde{q}^T)(I + \tilde{q}\tilde{p}^T)S$, takže podle lemmatu 8 platí

$$\delta = \det(I + \tilde{U}M\tilde{U}^T) = \det(I + \tilde{p}\tilde{q}^T)\det(I + \tilde{q}\tilde{p}^T) = (1 + \tilde{q}^T\tilde{p})^2 \geq 0.$$

Poznámka 69 Podle lemmatu 11 existuje součinový tvar pouze pro ty metody z Broydenovy třídy, pro které $\delta \geq 0$ a $\mu \geq 0$. Ve větě 37 ukážeme, že tyto nutné podmínky jsou v jistém smyslu i podmínkami postačujícími.

Nyní ukážeme, jak lze volit vektor $\tilde{q} = \tilde{U}\hat{q}$, abychom dostali jednotlivé metody z Broydenovy třídy. Jak vyplývá z důkazu lemmatu 10, je vektor p určen vektorem $\tilde{q}$ (existují obvykle dvě řešení). Navíc výsledná aktualizace nezávisí na normě vektoru $\tilde{q}$, neboť z (S) 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 36 *Nechť jsou splněny předpoklady lemmatu 11 a nechť $\tilde{q} = \tilde{y} + \vartheta\tilde{d}$. Pak aktualizace (S) je ekvivalentní aktualizaci (H), pokud*

$$\frac{\rho}{\gamma}\frac{(b+\vartheta c)^2}{D^2} = \frac{1}{ab}(\eta(ac-b^2)+b^2), \qquad (*)$$

kde

$$D^2 = \frac{\rho}{\gamma}b(a+2\vartheta b+\vartheta^2 c) - \vartheta^2(ac-b^2).$$

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

$$\frac{1}{\vartheta} = \chi \pm \sqrt{\left(\chi+\frac{\gamma}{\rho}\right)\left(\chi+\frac{c}{b}\right)}, \quad \chi = \frac{(1-\eta)b}{\eta a},$$

což lze zapsat ve tvaru

$$\frac{1}{\vartheta} = \frac{b}{\eta}\left(\frac{1-\eta}{a} \pm \frac{\gamma}{\rho}\sqrt{\delta\mu}\right).$$

Důkaz V důkazu lemmatu 11 jsme ukázali, že $\delta = (1+\tilde{q}^T\tilde{p})^2$. Výraz vystupující na pravé straně této rovnosti je podle lemmatu 10 roven číslu

$$1+\tilde{q}^T\tilde{p} = 1+q^Tp = \frac{\rho}{\gamma}\frac{q^Td}{D} = \frac{\rho}{\gamma}\frac{b+\vartheta c}{D}.$$

Dosadíme-li za δ výraz z lemmatu 9, dostaneme $(*)$. Jelikož $\tilde{q}^T\tilde{y} = a+\vartheta b$ a $\tilde{q}^T\tilde{q} = a+2\vartheta b+\vartheta^2 c$, dostaneme po úpravě výraz pro D^2 uvedený v tvrzení věty 36. Dosadíme-li tento vztah do $(*)$, dostaneme po úpravě

$$\vartheta^2\left(\frac{\rho}{\gamma}c+b\right) + 2\vartheta\frac{\rho}{\gamma}b = \frac{\eta}{b}D^2.$$

Jestliže $\eta = 0$, pak buď $\vartheta = 0$ nebo $1/\vartheta = -(\gamma/\rho + c/b)/2$. V opačném případě dosazením za D^2 a dalšími úpravami dostaneme

$$\frac{(1-\eta)b}{\eta a}\left[\vartheta^2\left(\frac{\rho c}{\gamma b}+1\right)+2\vartheta\frac{\rho}{\gamma}\right] + \vartheta^2\frac{c}{b} - \frac{\rho}{\gamma} = 0,$$

neboli

$$\frac{1}{\vartheta^2} - \frac{2}{\vartheta}\chi - \left(\frac{\gamma c}{\rho b} + \chi\left(\frac{\gamma}{\rho}+\frac{c}{b}\right)\right) = 0.$$

Tato kvadratická rovnice má řesení

$$\frac{1}{\vartheta} = \chi \pm \sqrt{\chi^2 + \chi\left(\frac{\gamma}{\rho} + \frac{c}{b}\right) + \frac{\gamma c}{\rho b}} = \chi \pm \sqrt{\left(\chi + \frac{\gamma}{\rho}\right)\left(\chi + \frac{c}{b}\right)}.$$

Poslední dokazovaný vztah plyne z toho, že

$$\chi + \frac{c}{b} = \frac{(1-\eta)b^2 + \eta ac}{\eta ab} = \frac{\eta(ac - b^2) + b^2}{\eta ab} = \frac{\gamma\delta}{\rho\eta}$$

a

$$\chi + \frac{\gamma}{\rho} = \frac{(1-\eta)b/a + \eta\gamma/\rho}{\eta} = \frac{\gamma b}{\rho a}\frac{(1-\eta)\rho/\gamma + \eta a/b}{\eta} = \frac{\gamma b^2}{\rho\eta}\mu.$$

Poznámka 70 Věta 36 udává způsob, jak lze k dané metodě s proměnnou metrikou (charakterizované parametrem η) nalézt součinový tvar (S). K dané hodnotě η najdeme podle věty 36 hodnotu ϑ určující vektor $\hat{q}$ a číslo D^2 (existují obvykle dvě řešení). Pak podle lemmatu 10 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 $\chi = 0$, 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 $\chi = \gamma/\rho$, 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 63).

Použití věty 36 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 37 *Uvažujme aktualizaci (H), kde $H = SS^T$, $a > 0$, $b > 0$, $c > 0$, $ac - b^2 > 0$, $\delta \ge 0$, $\mu \ge 0$ a $\delta + \mu > 0$. Nechť $d = S\tilde{d}$ a $\tilde{y} = S^T y$. Pak platí $H_+ = S_+^T S_+$, přičemž*

$$\frac{1}{\sqrt{\gamma}} S_+ = S + U\hat{p}\hat{q}^T\tilde{U}^T, \tag{S1}$$

kde

$$\hat{p}\hat{q}^T = \frac{1}{\lambda}\begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{\mu} \\ -1 - b\sqrt{\mu} \end{bmatrix}\begin{bmatrix} \sqrt{\delta} - b\sqrt{\mu} \\ -\frac{\gamma}{\rho}\sqrt{\delta} + c\sqrt{\mu} \end{bmatrix}^T$$

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 11 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 (S1), 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 11 víme, že

$$\det(I + \tilde{U}\hat{p}\hat{q}^T\tilde{U}^T) = \sqrt{\delta},$$

takže podle lemmatu 8 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

$$\hat{q}_1\left(1 + b\sqrt{\mu}\right) + \hat{q}_2\left(\frac{\rho}{\gamma} + a\sqrt{\mu}\right) = \sqrt{\mu}.$$

Řešením těchto dvou rovnic je vektor

$$\hat{q} = \frac{1}{\lambda} \begin{bmatrix} \sqrt{\delta} - b\sqrt{\mu} \\ -\frac{\gamma}{\rho}\sqrt{\delta} + c\sqrt{\mu} \end{bmatrix},$$

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 znaménka $\sqrt{\delta}$ docílit toho, že $\lambda \neq 0$.
(c) Nechť nyní $\delta = 0$. Jelikož podle předpokladu platí $ac - b^2 > 0$ a $\mu > 0$, můžeme vhodnou volbou znaménka $\sqrt{\mu}$ docílit toho, že $\lambda \neq 0$. Jelikož podle poznámky 59 je δ 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 (S1) závisí spojitě na ε, lze použít limitní přechod a tvrzení platí i pro $\delta = \delta(\eta) = 0$.

Poznámka 71 Ve větě 37 jsme použili předpoklad $\delta + \mu > 0$, neboť v opačném případě je matice $\hat{p}\hat{q}^T$ vystupující v (S1) nulová. I v tomto případě je však možné vyjádřit aktualizaci (H) v součinovém tvaru. Hodnota $\mu = 0$ odpovídá metodě hodnosti 1, pro kterou podle důsledku 7 platí vyjádření (SR) a pokud $ac - b^2 > 0$ nemůže být jmenovatel v (SR) nulový.

Obě předchozí věty obsahují poměrně komplikované výrazy. Tyto výrazy se velmi zjednoduší pro základní metody (HD), (HB), (HR).

Důsledek 7 Pro metodu DFP platí $\eta = 0$, čili $\delta = \rho b/(\gamma a)$ a $\mu = \rho/(\gamma ab)$, takže

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{1}{a}\left(SS^T y \pm \sqrt{\frac{\rho a}{\gamma b}}d\right)\tilde{y}^T. \tag{SD}$$

Pro metodu BFGS platí $\eta = 1$, čili $\delta = \rho c/(\gamma b)$ a $\mu = 1/b^2$, takže

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{1}{b}d\left(\tilde{y} \pm \sqrt{\frac{\rho b}{\gamma c}}\tilde{d}\right)^T. \tag{SB}$$

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

$$\frac{1}{\sqrt{\gamma}}S_+ = S + \frac{\sqrt{\delta} - 1}{\left(\frac{\rho}{\gamma}\right)^2 c - 2\frac{\rho}{\gamma}b + a}\left(\frac{\rho}{\gamma}d - SS^T y\right)\left(\frac{\rho}{\gamma}\tilde{d} - \tilde{y}\right)^T. \tag{SR}$$

V těchto vzorcích je $d = S\tilde{d}$ (pokud $d \in \mathcal{L}(S)$) a $SS^T y = S\tilde{y}$ (pokud $\tilde{y} = S^T y$).

Důkaz K odvození těchto vztahů můžeme použít buď větu 36 nebo větu 37. Použití věty 36 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 37. Přímé dosazení do (S1) 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

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{1}{a}\begin{bmatrix} \frac{a}{b}\sqrt{\delta} \\ -1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = -\frac{1}{a}\begin{bmatrix} \pm\sqrt{\frac{\rho a}{\gamma b}} \\ 1 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

(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

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}} \\ -1 \end{bmatrix} = -\frac{1}{b}\begin{bmatrix} 1 \\ 0 \end{bmatrix}\begin{bmatrix} \pm\sqrt{\frac{\rho b}{\gamma c}} \\ 1 \end{bmatrix}.$$

(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

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta} = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c - b}{\frac{\rho}{\gamma}b - a} + \sqrt{\delta}\right) = \frac{\gamma}{\rho}\left(\frac{\rho}{\gamma}b - a\right)\sqrt{\delta}(\sqrt{\delta}+1),$$

$$\hat{p} = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}, \quad \hat{q} = \sqrt{\delta}\begin{bmatrix} 1 \\ -\frac{\gamma}{\rho} \end{bmatrix} = \frac{\rho}{\gamma}\sqrt{\delta}\begin{bmatrix} \frac{\gamma}{\rho} \\ -1 \end{bmatrix},$$

takže po vykrácení

$$\frac{1}{\lambda}\hat{p}\hat{q}^T = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T}{\left(\frac{\rho}{\gamma}b - a\right)(\sqrt{\delta}+1)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T(\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}b - a\right)\left(\frac{\rho}{\gamma}\frac{\frac{\rho}{\gamma}c-b}{\frac{\rho}{\gamma}b-a} - 1\right)} = \frac{\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}\begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}^T(\sqrt{\delta}-1)}{\left(\frac{\rho}{\gamma}\right)^2 c - 2\frac{\rho}{\gamma}b + a}.$$

Poznámka 72 Vzorec (SR) lze upravit tak, aby se v něm neodečítala blízká čísla. Dosadíme-li do (SR) výraz pro δ a rozšíříme-li zlomek číslem $\sqrt{\delta}+1$, vykrátí se nový čitatel s původním jmenovatelem a po úpravách dostaneme

$$\frac{1}{\sqrt{\gamma}}S_+ = S + \frac{1}{\frac{\rho}{\gamma}b - a \pm \sqrt{\frac{\rho}{\gamma}(\frac{\rho}{\gamma}b-a)(\frac{\rho}{\gamma}c-b)}}\left(\frac{\rho}{\gamma}d - SS^Ty\right)\left(\frac{\rho}{\gamma}\tilde{d}-\tilde{y}\right)^T.$$

Poznámka 73 V součinovém tvaru lze vyjádřit také vztah (B). Z praktických důvodů se inverzní součinový vztah používá pouze v případě, že matice B je regulární (potřebujeme řešit soustavu $Bs+g=0$). Předpokládejme, že $B=A^TA$ a $B_+=A_+^TA_+$, kde matice $A\in R^{m\times n}$ a $A_+\in R^{m\times n}$ mají plnou hodnost, což nastává například v případě, že $F(x)=(1/2)f^T(x)f(x)$ (minimalizace součtu čtverců) a matice A aproximuje Jacobiovu matici zobrazení $f:R^n\to R^m$.

Věta 38 *Uvažujme aktualizaci (B), kde $B=A^TA$, $a>0$, $b>0$, $c>0$, $\delta>0$ a $\mu>0$. Nechť $\tilde{d}=Ad$ a $\tilde{y}=A(A^TA)^{-1}y$ (takže $y=A^T\tilde{y}$). Pak platí $B_+=A_+^TA_+$, přičemž*

$$\sqrt{\gamma}A_+ = A - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T(BU)^T, \tag{A1}$$

kde $\hat{p}$ a $\hat{q}$ jsou vektory vystupující ve větě 37.

Důkaz Podle (Sa) platí

$$\frac{1}{\gamma}S_+S_+^T = (I+pq^T)SS^T(I+qp^T),$$

což s použitím lemmatu 8 dává

$$\gamma A_+^TA_+ = (I+qp^T)^{-1}A^TA(I+pq^T)^{-1} = \left(I-\frac{1}{\sqrt{\delta}}qp^T\right)A^TA\left(I-\frac{1}{\sqrt{\delta}}pq^T\right).$$

Platí tedy

$$\sqrt{\gamma}A_+ = A\left(I-\frac{1}{\sqrt{\delta}}pq^T\right) = A - \frac{1}{\sqrt{\delta}}\tilde{U}\hat{p}\hat{q}^T(BU)^T,$$

neboť

$$Ap = AU\hat{p} = A\left[d,\,(A^TA)^{-1}y\right]\hat{p} = \left[\tilde{d},\,\tilde{y}\right]\hat{p} = \tilde{U}\hat{p}$$

a

$$q = A^T\tilde{q} = A^T\tilde{U}\hat{q} = A^T\left[\tilde{d},\,\tilde{y}\right]\hat{q} = \left[A^TAd,\,y\right]\hat{q} = BU\hat{q}.$$

Poznámka 74 Pro metodu DFP platí

$$\sqrt{\gamma}A_+ = A - \frac{1}{b}\left(\tilde{d}\pm\sqrt{\frac{\gamma b}{\rho a}}\tilde{y}\right)y^T. \tag{AD}$$

Pro metodu BFGS platí

$$\sqrt{\gamma}A_+ = A - \frac{1}{c}\tilde{d}\left(A^TAd\pm\sqrt{\frac{\gamma c}{\rho b}}y\right)^T. \tag{AB}$$

Pro metodu hodnosti 1 platí

$$\sqrt{\gamma}A_+ = A + \frac{1/\sqrt{\delta} - 1}{\left(\frac{\gamma}{\rho}\right)^2 a - 2\frac{\gamma}{\rho}b + c}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho}y - A^TAd\right)^T. \qquad \text{(AR)}$$

kde $1/\delta = (\gamma/\rho)((\gamma/\rho)a - b)/((\gamma/\rho)b - c)$. Vzorec (AR) lze upravit na tvar

$$\sqrt{\gamma}A_+ = A + \frac{1}{\frac{\gamma}{\rho}b - c \pm \sqrt{\frac{\gamma}{\rho}(\frac{\gamma}{\rho}b - c)(\frac{\gamma}{\rho}a - b)}}\left(\frac{\gamma}{\rho}\tilde{y} - \tilde{d}\right)\left(\frac{\gamma}{\rho}y - A^TAd\right)^T.$$

Ve všech těchto vzorcích je $y = A^T\tilde{y}$ (pokud $y \in \mathcal{L}(A^T)$) a $A^TAd = A^T\tilde{d}$ (pokud $\tilde{d} = Ad$). Poznamenejme, že pro minimalizaci součtu čtverců má praktický význam pouze metoda BFGS, která používá vektor $\tilde{d} = Ad$. Ostatní metody potřebují navíc vektor $\tilde{y} = A(A^TA)^{-1}y$, takže je nutné invertovat matici A^TA.

Poznámka 75 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) Definitní 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 definitní 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 definitní pouze tehdy, jestliže 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.

Věta 39 *Nechť W je SPD matice. Pak Frobeniova norma $\|W^{-1/2}((1/\gamma)H_+ - H)W^{-1/2}\|_F$ je minimální na množině všech matic splňujících kvazinewtonovskou podmínku*

$$\left(\frac{1}{\gamma}H_+ - H\right)y = \frac{\rho}{\gamma}d - Hy \triangleq w$$

právě tehdy, platí-li

$$\frac{1}{\gamma}H_+ = \left(H + \frac{Wyw^T + w(Wy)^T}{y^TWy} - \frac{w^Ty}{y^TWy}\frac{Wy(Wy)^T}{y^TWy}\right). \qquad (\overline{\text{H}})$$

Důkaz Jelikož matice $(1/\gamma)H_+ - H$ je symetrická, můžeme položit $(1/\gamma)H_+ - H = X + X^T$, kde X je zatím neznámá čtvercová matice . Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{4}\|W^{-1/2}(X + X^T)W^{-1/2}\|_F^2 + u^T(w - (X + X^T)y) = \\ &= \frac{1}{4}\sum_{i=1}^n\sum_{j=1}^n(e_i^TW^{-1/2}(X + X^T)W^{-1/2}e_j)^2 + \sum_{i=1}^n u_i\left(w_i - \sum_{j=1}^n e_i^T(X + X^T)e_jy_j\right), \end{aligned}$$

kde $u \in R^n$ je vektor Lagrangeových multiplikátorů (e_i, e_j jsou sloupce jednotkové matice řádu n). Derivováním Langrangeovy funkce dostaneme

$$\begin{aligned}
\frac{\partial L}{\partial X_{kl}} &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j)(e_i^T W^{-1/2}(e_k e_l^T + e_l e_k^T)W^{-1/2}e_j) - \\
&\quad -\sum_{i=1}^{n} u_i \sum_{j=1}^{n} e_i^T(e_k e_l^T + e_l e_k^T)e_j y_j = \\
&= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} e_k^T W^{-1/2} e_i e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j e_j^T W^{-1/2} e_l + \\
&\quad +\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} e_l^T W^{-1/2} e_i e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j e_j^T W^{-1/2} e_k + \\
&\quad +\sum_{i=1}^{n}\sum_{j=1}^{n} u_i(e_i^T e_k e_l^T e_j + e_i^T e_l e_k^T e_j)y_j = \\
&= e_k^T W^{-1}(X+X^T)W^{-1}e_l - (u_k y_l + u_l y_k).
\end{aligned}$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar

$$\frac{1}{\gamma}H_+ - H = W(uy^T + yu^T)W.$$

Dosadíme-li tento vektor do vztahu $((1/\gamma)H_+ - H)y = w$, dostaneme po úpravě

$$(y^T W y \cdot W + Wy(Wy)^T)u = w.$$

Podle důsledku 6, můžeme vypočítat inverzi

$$(y^T W y \cdot W + Wy(Wy)^T)^{-1} = \frac{1}{y^T W y}\left(W^{-1} - \frac{yy^T}{2y^T W y}\right)$$

(z pozitivní definitnosti matice W plyne $y^T W y > 0$ pro $y \neq 0$). Platí tedy

$$u = \frac{1}{y^T W y}\left(W^{-1} - \frac{yyT}{2y^T W y}\right)w,$$

což po dosazení do vztahu pro $(1/\gamma)H_+ - H$ dává tvrzení věty. Postačitelnost plyne z konvexity Frobeniovy normy.

Je zřejmé, že metoda získaná aktualizací ($\overline{\mathrm{H}}$) patří do Broydenovy třídy právě tehdy, je-li vektor Wy lineární kombinací vektorů d a Hy. Protože aktualizace ($\overline{\mathrm{H}}$) nezávisí na normě vektoru Wy, budeme předpokládat, že $Wy = d + \vartheta Hy$.

Věta 40 *Nechť jsou splněny předpoklady věty 39 a nechť $Wy = d + \vartheta Hy$. Pak aktualizace ($\overline{\mathrm{H}}$) je ekvivalentní aktualizaci (H), pokud*

$$\eta = \frac{b(b - \vartheta^2(\rho/\gamma)a)}{(b+\vartheta a)^2}. \qquad (*)$$

Pokud $\eta = 1$, pak buď $\vartheta = 0$, nebo $1/\vartheta = -(\rho/\gamma + a/b)/2$. Pokud $\eta \neq 1$ a $\mu \geq 0$, platí

$$\frac{1}{\vartheta} = \frac{a}{1-\eta}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

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 ($\overline{\text{H}}$) a porovnáme-li záporně vzaté koeficienty u smíšených členů (v aktualizaci (H) je tento koeficient roven η/b), dostaneme

$$\frac{b - \vartheta^2(\rho/\gamma)a}{(b + \vartheta a)^2} = \frac{\eta}{b},$$

odkud plyne $(*)$. Jestliže $\eta = 1$, dostaneme z $(*)$ $b^2 - \vartheta^2(\rho/\gamma)ab = b^2 + 2\vartheta ab + \vartheta^2 a^2$,
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 úpravě dává

$$\frac{\eta - 1}{\vartheta^2} + \frac{2}{\vartheta}\eta\frac{a}{b} + \eta\left(\frac{a}{b}\right)^2 + \frac{\rho}{\gamma}\frac{a}{b} = 0.$$

Tato kvadratická rovnice má řesení

$$\frac{1}{\vartheta} = \frac{\eta a}{(1-\eta)b} \pm \frac{\sqrt{\left[\eta\left(\frac{a}{b} - \frac{\rho}{\gamma}\right) + \frac{\rho}{\gamma}\right]\frac{a}{b}}}{1-\eta} = \frac{a}{1-\eta}\left(\frac{\eta}{b} \pm \sqrt{\mu}\right).$$

Poznámka 76 Analogický postup lze použít pro aktualizaci matice B. Nechť V je SPD matice. Pak Frobeniova norma $\|V^{-1/2}(\gamma B_+ - B)V^{-1/2}\|_F$ je minimální na množině všech matic splňujících kvazinewtonovskou podmínku

$$(\gamma B_+ - B)\, d = \frac{\gamma}{\rho} y - By \triangleq v$$

právě tehdy, platí-li

$$\gamma B_+ = \left(B + \frac{Vdv^T + v(Vd)^T}{d^T Vd} - \frac{v^T d}{d^T Vd}\frac{Vd(Vd)^T}{d^T Vd}\right). \qquad (\overline{\text{B}})$$

Zvolíme-li matici V tak že $Vd = y + \vartheta Bd$, je aktualizace ($\overline{\text{B}}$) ekvivalentní aktualizaci (B), pokud

$$\beta = \frac{b(b - \vartheta^2(\gamma/\rho)c)}{(b + \vartheta c)^2}.$$

Pokud $\beta = 1$, pak buď $\vartheta = 0$, nebo $1/\vartheta = -(\gamma/\rho + c/b)/2$. Pokud $\beta \neq 1$ a $\mu \geq 0$, platí

$$\frac{1}{\vartheta} = \frac{c}{1-\beta}\left(\frac{\beta}{b} \pm \sqrt{\mu}\right).$$

Poznámka 77 Zvolíme-li $V = I$, dostaneme metodu, která nepatří do Broydenovy třídy a která se nazývá Powellovou symetrizací Broydenovy metody (PSB).

$$B_+ = \frac{1}{\gamma}\left(B + \frac{dv^T + vd^T}{d^T d} - \frac{v^T d}{d^T d}\frac{dd^T}{d^T d}\right).$$

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 82).

Minimalizační postup lze použít i k odvození součinového tvaru metod s proměnnou metrikou. V tomto případě dostaneme vyjádření, které je obecnější než (S1) a které obsahuje i aktualizace hodnosti 2. Abychom mohli použít variační princip, zapíšeme kvazinewtonovskou podmínku ve tvaru

$$\frac{1}{\sqrt{\gamma}}S_+^T y = z, \qquad \frac{1}{\sqrt{\gamma}}S_+ z = \frac{\rho}{\gamma}d, \qquad z^T z = \frac{\rho}{\gamma}b, \tag{$*$}$$

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 41 *Nechť T je SPD matice. Pak Frobeniova norma $\|T^{-1/2}(S_+/\sqrt{\gamma} - S)\|_F$ je minimální na množině všech matic splňujících kvazinewtonovskou podmínku (*) právě tehdy, platí-li*

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \frac{Ty}{y^T Ty}y^T S + \left(\frac{\rho}{\gamma}d - Sz + \frac{y^T Sz}{y^T Ty}Ty\right)\frac{z^T}{z^T z}. \tag{$\overline{\text{S2}}$}$$

Důkaz Označme $X = S_+/\sqrt{\gamma}$. Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2}\left\|T^{-1/2}(X - S)\right\|_F^2 + u^T\left(X^T y - z\right) + v^T\left(Xz - \frac{\rho}{\gamma}d\right) \\ &= \sum_{i=1}^{m}\left[\frac{1}{2}(x_i - s_i)^T T^{-1}(x_i - s_i) + u_i y^T x_i + z_i v^T x_i\right] - u^T z - \frac{\rho}{\gamma}v^T d, \end{aligned}$$

kde $S = [s_1, \dots s_m]$ a $X = [x_1, \dots x_m]$. Derivováním Langrangeovy funkce dostaneme

$$\frac{\partial L}{\partial x_i} = T^{-1}(x_i - s_i) + u_i y + z_i v.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $T^{-1}(x_i - s_i) + u_i y + z_i v = 0$, $1 \le i \le m$, neboli

$$X - S = -Tyu^T - Tvz^T.$$

Použitím první podmínky z (*) dostaneme

$$X^T y = S^T y - y^T Tyu - v^T Tyz = z \quad \Rightarrow \quad u = \frac{1}{y^T Ty}\left(S^T y - (1 + v^T Ty)z\right),$$

což po dosazení do předchozí rovnosti dává

$$X - S = -\frac{Ty}{y^T Ty}y^T S + wz^T,$$

kde $w \in R^n$ je zatím neznámý vektor (jednoznačně určený vektorem v). Užitím druhé podmínky z (*) dostaneme

$$Xz = Sz - \frac{y^T Sz}{y^T Ty}Ty + z^T z w = \frac{\rho}{\gamma}d \quad \Rightarrow \quad w = \frac{1}{z^T z}\left(\frac{\rho}{\gamma}d - Sz + \frac{y^T Sz}{y^T Ty}Ty\right),$$

což po dosazení do předchozí rovnosti (s využitím vztahu $X = S_+/\sqrt{\gamma}$) dává ($\overline{\text{S2}}$). Postačitelnost plyne z konvexity Frobeniovy normy.

Nyní ukážeme, jak lze volit vektory Ty a z, abychom dostali jednotlivé metody z Broydenovy třídy.

Věta 42 *Uvažujme aktualizaci (H), kde $H = SS^T$, $a > 0$, $b > 0$, $c > 0$ a $\eta \geq 0$ (takže $\delta > 0$). Nechť S_+ je matice určená podle ($\overline{S2}$), kde*

$$Ty = \frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy, \qquad z = \frac{\rho}{\gamma}\frac{b}{\sqrt{\delta}}S^T BTy$$

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 v poznámce 59). Pak platí $H_+ = S_+S_+^T$.

Důkaz (a) Položme $z = \vartheta bS^T BTy$, kde hodnota ϑ se vybírá tak, aby platilo $z^T z = (\rho/\gamma)b$ (vztah (*)). Jelikož $z^T z = \vartheta^2 b^2 y^T TBTy$, je tato hodnota dána výrazem

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{1}{b\, y^T TBTy}.$$

Speciální volbu $z = \vartheta bS^T BTy$ používáme proto, že se tím značně zjednoduší aktualizace ($\overline{\text{S2}}$), neboť v tomto případě platí

$$\frac{y^T Sz}{y^T Ty}Ty - Sz = \frac{\vartheta b\, y^T Ty}{y^T Ty}Ty - \vartheta b\, Ty = 0.$$

(b) Jelikoz vynásobení matice T kladným číslem nezmění tvar aktualizace ($\overline{\text{S2}}$), budeme předpokládat, že $y^T Ty = 1$. Položíme-li

$$Ty = \frac{\alpha_1}{b}d + \frac{\alpha_2}{a}Hy,$$

pak z $y^T Ty = 1$ plyne $\alpha_1 + \alpha_2 = 1$. Dále platí

$$y^T TBTy = \frac{\alpha_1^2}{b^2}c + 2\frac{\alpha_1\alpha_2}{ab}b + \frac{\alpha_2^2}{a^2}a = \frac{\alpha_1^2(ac-b^2)+b^2}{ab^2}$$

(používáme vztah $\alpha_1 + \alpha_2 = 1$). Dosadíme-li tento výsledek do výrazu odvozeného v (a), dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{1}{b\, y^T TBTy} = \frac{\rho}{\gamma}\frac{ab}{\alpha_1^2(ac-b^2)+b^2}$$

(c) Nyní využijeme toho, že vektory z a Ty uvedené v (a) a (b) umožňují zapsat aktualizaci ($\overline{\text{S2}}$) ve velmi jednoduchém tvaru

$$\frac{1}{\sqrt{\gamma}}S_+ = S - Tyy^T S + \vartheta dy^T TBS.$$

Položíme-li $H = SS^T$ a $H_+ = S_+S_+^T$, dostaneme z předchozího vztahu (po vynásobení)

$$\begin{aligned}\frac{1}{\gamma}H_+ &= H - (Tyy^T H + Hyy^T T) + \vartheta(dy^T T + Tyd^T) + aTyy^T T - \vartheta(dy^T T + Tyd^T) + \vartheta^2 y^T TBTydd^T \\ &= H - (Tyy^T H + Hyy^T T) + aTyy^T T + \vartheta^2 y^T TBTydd^T.\end{aligned}$$

Jelikož vektor Ty 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 33. K určení parametru η stačí porovnat koeficienty u $Hyy^T H$ v obou vyjádřeních. Podle věty 33 se tento koeficient rovná $(\eta-1)/a$ a dosazením vektoru $Ty = (\alpha_1/b)d + (\alpha_2/a)Hy$ 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 59, dostaneme

$$\vartheta^2 = \frac{\rho}{\gamma}\frac{ab}{\eta(ac-b^2)+b^2} = \left(\frac{\rho}{\gamma}\right)^2\frac{1}{\delta}.$$

Důsledek 8 *Nechť jsou splněny předpoklady věty 42 a nechť*

$$\frac{1}{\sqrt{\gamma}}S_+ = S - \left(\frac{\sqrt{\eta}}{b}d + \frac{1-\sqrt{\eta}}{a}Hy\right)y^TS + \frac{\rho}{\gamma}\frac{1}{\sqrt{\delta}}d\left(\frac{\sqrt{\eta}}{b}d^TB + \frac{1-\sqrt{\eta}}{a}y^T\right)S$$

$$= S - \left(\left(\frac{\sqrt{\eta}}{b} - \vartheta\frac{1-\sqrt{\eta}}{a}\right)d - \frac{1-\sqrt{\eta}}{a}Hy\right)y^TS + \vartheta\frac{\sqrt{\eta}}{b}dd^TBS, \tag{S2}$$

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 Ty a čísla ϑ uvedených ve větě 42 do vzorce $(1/\sqrt{\gamma})S_+ = S - Tyy^TS + \vartheta dy^TTBS$ použitého v důkazu této věty.

Poznámka 78 Věta 42 používá jiné předpoklady než věta 37, nerovnost $\mu \geq 0$ je nahražena nerovností $\eta \geq 0$. Vztah (S2) lze tedy použít pro každou perfektní metodu z Broydenovy třídy. Na druhé straně matice $(1/\sqrt{\gamma})S_+ - S$ v (S2) má obecně hodnost 2, takže (S2) vyžaduje více numerických operací než (S1). Je zajímavé, že pro $\eta = 0$ (DFP) nebo $\eta = 1$ (BFGS) dávají oba vztahy stejný výsledek, dosazení do (S2) je však nesrovnatelně jednodušší. Je také zajímavé porovnat (S2) s pseudosoučinovým tvarem uvedeným v poznámce 61.

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 26.

Věta 43 *Nechť $\tilde{G}$ je matice taková, že $\tilde{G}d = y$, tedy například*

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

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 (H), 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

$$\frac{1}{\gamma}v^TR'_+v = v^TRv + \frac{\eta - 1}{a}v^TRz \leq v^TRv$$

(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 64) 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 79 Podle věty 43 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 75). Metoda hodnosti 1 však vyžaduje, aby $0 < \rho/\gamma < b/c$ nebo $a/b < \rho/\gamma$ (poznámka 63), neboť jinak není matice H_+ pozitivně definitní. Interval $0 < \rho/\gamma < b/c$ je nevhodný, neboť v tomto případě $\eta^{HR} < 0$. Zbývá tedy interval $a/b < \rho/\gamma$. Pak $\eta^{HR} > 1$ a metoda hodnosti 1 patří mezi perfektní metody s proměnnou metrikou. Bližší podrobnosti týkající se volby podílu ρ/γ jsou uvedeny v poznámce 80.

K volbě parametru η lze použít různé minimalizační principy. Nejvíce se ujal princip spočívající v minimalizaci čísla podmíněnosti matice $(1/\gamma)H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}}$.

Lemma 12 *Nechť jsou splněny předpoklady lemmatu 9 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

$$\eta = \eta^{HM} = \frac{bc(\rho/\gamma - b/c)}{ac - b^2} > \eta^*. \tag{HM}$$

Důkaz (a) Podle lemmatu 9 jsou čísla $\underline{\lambda}(\eta)$ a $\overline{\lambda}(\eta)$ vlastními čísly matice $(1/\gamma)H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}} \triangleq T + \eta uu^T$ (použili jsme vztah (H)). Nechť $\eta_1 > \eta_2$. 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 (VM2) platí $(1/\gamma)H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}} = I + H^{-\frac{1}{2}}UMU^TH^{-\frac{1}{2}}$, 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 $H^{-\frac{1}{2}}UMU^TH^{-\frac{1}{2}}$, 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 34 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 9, 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 44 *Nechť jsou splněny předpoklady lemmatu 12. Označme $\kappa(\eta)$ číslo podmíněnosti matice $(1/\gamma)H^{-\frac{1}{2}}H_+H^{-\frac{1}{2}}$. Pak:*

(a) Pokud $0 < \rho/\gamma < b/c$, je $\kappa(\eta)$ minimální právě tehdy, jestliže $\eta = \max(\eta^{HR}, \eta^{HM})$.

(b) Pokud $b/c \le \rho/\gamma \le a/b$, je $\kappa(\eta)$ minimální právě tehdy, jestliže $\eta = \eta^{HM}$.

(a) Pokud $a/b < \rho/\gamma$, je $\kappa(\eta)$ minimální právě tehdy, jestliže $\eta = \min(\eta^{HR}, \eta^{HM})$.

Důkaz Zřejmě

$$\kappa(\eta) = \frac{\max(1, \overline{\lambda}(\eta))}{\min(1, \underline{\lambda}(\eta))}. \qquad (*)$$

Jestliže $\mu \ge 0$, podle (b) lemmatu 12 platí $0 < \underline{\lambda}(\eta) \le 1 \le \overline{\lambda}(\eta)$, takže podle (c) lemmatu 12 je $\kappa(\eta)$ minimální, pokud $\eta = \eta^{HM}$. Jestliže $\mu < 0$, lze podle (a) lemmatu 12 oba kořeny rovnice $\lambda^2 + \sigma\lambda + \delta = 0$ současně zvětšit nebo zmenšit změnou parametru η. Podíl $(*)$ 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 63 a poznámky 75.

Poznámka 80 Větu 44 lze použít k volbě parametru η. Jestliže $b/c \le \rho/\gamma \le a/b$, je vhodné použít hodnotu $\eta = \eta^{HM}$. Mnohem praktičtější aplikací věty 44 je však určení vhodného podílu ρ/γ pro danou hodnotu parametru η. V tomto případě z $\eta = \eta^{HM}$ 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 12 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

$$\rho/\gamma = \frac{\eta(ac-b^2)+b^2}{bc} = \frac{a}{b}\frac{\rho/\gamma + b/c}{\rho/\gamma + a/b},$$

neboli $\rho/\gamma = \sqrt{a/c}$

(d) Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)/(\rho/\gamma - a/b)$. Této hodnotě odpovídá podle věty 44 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 44 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 $ac - b^2 > 0$).

Pokud $\eta > 1$, platí podle věty 44 $\rho/\gamma > a/b$. Použití této hodnoty však není vhodné, neboť v tomto případě nejsou splněny předpoklady věty 43. 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 81 Větu 44 můžeme 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 $\eta > 1$).

Zatím jsme popsali, jak lze volit parametr η a podíl ρ/γ. Nyní ukážeme jak se určují parametry ρ a γ. Parametr ρ slouží jako korekce 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.

Poznámka 82 Použijeme-li větu o střední hodnotě (tvrzení 1) v přímém nebo zpětném směru (s aproximací B_+ místo $G(\tilde{x})$), dostaneme

$$F_+ = F + d^T g + \frac{1}{2} d^T B_+ d$$

nebo

$$F = F_+ - d^T g + \frac{1}{2} d^T B_+ d.$$

Použijeme-li modifikovanou kvazinewtonovskou podmínku $d^T B_+ d = (1/\rho) d^T y$, můžeme z předchozích rovnic určit hodnotu parametru ρ. Platí

$$\rho = \frac{d^T y}{2(F_+ - F - d^T g)}$$

nebo

$$\rho = \frac{d^T y}{2(F - F_+ + d^T g_+)}.$$

Tyto hodnoty používáme pouze tehdy, jestliže $\underline{\rho} \leq \rho \leq \overline{\rho}$ (obvykle $\underline{\rho} = 0.01$ a $\overline{\rho} = 100$). V opačném případě pokládáme $\rho = 1$. Hodnota založená na zpětném použití věty o střední hodnotě (druhý vzorec) dává lepší praktické výsledky.

Poznámka 83 Existují další způsoby, jak lze určit vhodnou 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. Velmi se osvědčilo použití homogenního modelu

$$\varphi(\alpha) = a\alpha^r + b\alpha + c,$$

kde a, b, c jsou neznáme koeficienty a r je neznámý exponent.

Věta 45 *Uvažujme homogenní model $\varphi(\alpha) = a\alpha^r + b\alpha + c$. Pak platí*

$$\rho = \frac{A - 1}{B - A},$$

kde

$$\begin{aligned} A &= \frac{\varphi(\alpha) - \varphi(0)}{\alpha \varphi'(0)} = \frac{F_+ - F}{d^T g}, \\ B &= \frac{\varphi'(\alpha)}{\varphi(0)} = \frac{d^T g_+}{d^T g}. \end{aligned}$$

Důkaz Platí

$$\begin{aligned} \varphi'(\alpha) &= ar\alpha^{r-1} + b, \\ \varphi''(\alpha) &= ar(r-1)\alpha^{r-2}, \end{aligned}$$

takže

$$\begin{aligned} B - 1 &= \frac{\varphi'(\alpha)}{\varphi(0)} - 1 = \frac{ar\alpha^{r-1} + b}{b} - 1 = \frac{ar\alpha^{r-1}}{b}, \\ A - 1 &= \frac{\varphi(\alpha) - \varphi(0)}{\alpha \varphi'(0)} - 1 = \frac{a\alpha^r - b\alpha}{\alpha b} - 1 = \frac{a\alpha^{r-1}}{b}, \end{aligned}$$

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 83) dává

$$\rho = \frac{\varphi'(\alpha)-\varphi'(0)}{\varphi'(0)}\frac{\varphi'(0)}{\alpha\varphi''(\alpha)} = (B-1)\frac{A-1}{(B-1)(B-A)} = \frac{A-1}{B-A}.$$

Je zajímavé, že výsledný vztah neobsahuje neznámý exponent r.

Parametr γ slouží ke škálování matice H, neboť aktualizace (H) s $\gamma \neq 1$ je ekvivalentní aktualizaci (H) s $\gamma = 1$ aplikované na matici γH. Důvod pro škálování poskytuje věta 43 a následující věta.

Věta 46 *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 84 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 85 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 91). 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 $\underline{\gamma} = 0.7$ 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 $F_1 = F(x+\alpha_1 s)$, $g_1 = g(x + \alpha_1 s)$, $\lambda_1 = s^T g_1 / s^T g$, a nechť $\lambda > 0$ je vhodná konstanta (například $\lambda = 0.2$). Pak hodnotu $\gamma = 1$ použijeme navíc v následujících případech (γ je původní hodnota určená podle (IS)):
   (a) Jestliže $|\lambda_1| \leq \lambda$ a $F_1 \leq F$.
   (b) Jestliže $\gamma > 1$ a buď $F_1 > F$ nebo $\lambda_1 < 0$.
   (c) Jestliže $\gamma < 1$ a $F_1 \leq F$ a $\lambda_1 > 0$.

(AS) Permanentní škálování. V každém iteračním kroku postupujeme tak jako v prvním iteračním kroku strategie (PS).

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 13 *Uvažujme metodou 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 $F : R^n \to R$ splňuje podmínky (F1), (F3) a (F4). Pak:*

(a) Existuje konstanta $\overline{C}$ taková, že $Tr\, B_{i+1} \leq \overline{C}^i$.

(b) Existuje konstanta $\underline{K}$ taková, že

$$\prod_{j=1}^{i} \frac{c_j}{b_j} \geq \underline{K}^i, \qquad \sum_{j=1}^{i} \frac{c_j}{b_j} \geq i\underline{K}.$$

Důkaz (a) Vztah (B) 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 43), kde matice $\tilde{G}_i$ vyhovuje nerovnostem v podmínkách (F3)–(F4), můžeme psát

$$\begin{aligned}\frac{y_i^T y_i}{y_i^T d_i} &= \frac{d_i^T \tilde{G}_i^2 d_i}{d_i^T \tilde{G}_i d_i} \le \overline{G},\\ \frac{d_i^T d_i}{y_i^T d_i} &= \frac{d_i^T d_i}{d_i^T \tilde{G}_i d_i} \le \frac{1}{\underline{G}},\\ \frac{\|y_i\|\|d_i\|}{y_i^T d_i} &= \sqrt{\frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T d_i}{y_i^T d_i}} \le \sqrt{\frac{\overline{G}}{\underline{G}}}.\end{aligned}$$

Dosadíme-li tyto nerovnosti spolu s nerovností $\|B_i\| \le Tr\, B_i$ do vztahu pro $Tr\, B_{i+1}$, dostaneme

$$\begin{aligned}Tr\, B_{i+1} &\le Tr\, B_{i+1} + 1 \le \frac{1}{\underline{\gamma}}\left(1 + \frac{\overline{G}}{\underline{G}} + 2\sqrt{\frac{\overline{G}}{\underline{G}}}\right) Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + 1\\ &\le \underline{C}(Tr\, B_i + 1) \le \underline{C}^i (Tr\, B_1 + 1) \le \overline{C}^i,\end{aligned}$$

kde $\underline{C} = \max\left(\left(1 + \overline{G}/\underline{G} + 2\sqrt{\overline{G}/\underline{G}}\right)/\underline{\gamma},\ \overline{G}/\underline{\rho} + 1\right)$ a $\overline{C} = \underline{C}(Tr\, B_1 + 1)$.

(b) Použijeme-li vztah pro $\det B_{i+1}$ (poznámka 66), můžeme psát

$$\frac{\det B_{i+1}}{\det B_i} = \left(\frac{1}{\gamma_i}\right)^n \frac{\gamma_i}{\rho_i}\frac{b_i}{c_i}\left(1 - \frac{\beta_i}{\beta_i^*}\right) \ge \left(\frac{1}{\overline{\gamma}}\right)^n \frac{\underline{\gamma}}{\overline{\rho}}\frac{b_i}{c_i}\lambda = \overline{K}\frac{b_i}{c_i},$$

kde $\overline{K} = (1/\overline{\gamma})^n(\underline{\gamma}/\overline{\rho})\lambda$. Platí tedy

$$\frac{\det B_{i+1}}{\det B_1} \ge \overline{K}^i \prod_{j=1}^{i}\frac{b_j}{c_j}$$

a protože $\det H_{i+1} = 1/\det B_{i+1}$, můžeme psát

$$\det H_{i+1} \le \frac{\det H_1}{\overline{K}^i}\prod_{j=1}^{i}\frac{c_j}{b_j} \le \frac{1}{K^i}\prod_{j=1}^{i}\frac{c_j}{b_j},$$

kde $K = \overline{K}/(\det H_1 + 1)$. Podle (a) platí

$$\det B_{i+1} \le \left(\frac{Tr\, B_{i+1}}{n}\right)^n \le (Tr\, B_{i+1})^n \le \overline{C}^{in} \triangleq C^i$$

(používáme nerovnost mezi geometrickým a aritmetickým průměrem), takže

$$\frac{1}{C^i} \le \det H_{i+1} \le \frac{1}{K^i}\prod_{j=1}^{i}\frac{c_j}{b_j},$$

neboli

$$\prod_{j=1}^{i}\frac{c_j}{b_j} \ge \left(\frac{K}{C}\right)^i \triangleq \underline{K}^i.$$

Použijeme-li nerovnost mezi geometrickým a aritmetickým průměrem, dostaneme

$$\sum_{j=1}^{i}\frac{c_j}{b_j} \ge i\left(\prod_{j=1}^{i}\frac{c_j}{b_j}\right)^{1/i} \ge i\underline{K}.$$

Věta 47 *(Globální konvergence) Nechť jsou splněny předpoklady lemmatu 13, 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 13. Protože $y_i = \tilde{G}_i d_i$ a $B_i d_i = -\alpha_i g_i$, můžeme psát

$$\begin{aligned}
\frac{|y_i^T B_i d_i|}{y_i^T d_i} &= \frac{|y_i^T B_i d_i|}{d_i^T B_i d_i}\frac{c_i}{b_i} \leq \frac{\|\tilde{G} d_i\|\|\alpha_i g_i\|}{-\alpha_i d_i^T g_i}\frac{c_i}{b_i} \leq \frac{\overline{G}}{\cos\theta_i}\frac{c_i}{b_i},\\
\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i} &= \frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i}\frac{y_i^T d_i}{d_i^T B_i d_i}\frac{c_i}{b_i} \geq \frac{\alpha_i^2\|g_i\|^2\underline{G}\|d_i\|^2}{\alpha_i^2(d_i^T g_i)^2}\frac{c_i}{b_i} = \frac{\underline{G}}{\cos^2\theta_i}\frac{c_i}{b_i},
\end{aligned}$$

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 $\sum_{i=1}^{\infty}\cos\theta_i < \infty$, takže $\cos\theta_i \to 0$ a tedy $\xi_i \to -\infty$. Existuje tedy index $k \in N$ takový, že $\xi_i < -2\overline{G}/(\underline{\rho}\underline{K})$ $\forall i \geq k$. 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 13 platí

$$Tr\, B_{i+1} \leq Tr\, B_i + \frac{\overline{G}}{\underline{\rho}} + \xi_i\frac{c_i}{b_i} \leq Tr\, B_1 + i\frac{\overline{G}}{\underline{\rho}} - 2\frac{\overline{G}}{\underline{\rho}\underline{K}}\sum_{j=1}^{i}\frac{c_j}{b_j} \leq Tr\, B_1 - i\frac{\overline{G}}{\underline{\rho}}.$$

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 SPD matice je kladná.

Poznámka 86 Věta 47 je nejobecnějším doposud známým tvrzením o globální konvergenci metod s proměnnou metrikou. Tato věta vyžaduje aby byla splněna podmínka (F4) (exitence konstanty $\underline{G} > 0$), takže ji lze použít pouze pro konvexní funkce. Podmínku $\gamma_i \geq 1$, $i \in N$, která je pro důkaz věty 47 důležitá, můžeme poněkud oslabit. Stačí, když existuje konstanta $\underline{\gamma}$ taková, že $\prod_{j=k}^{l}\gamma_j \geq \underline{\gamma}\ \forall k < l$. To bývá v praxi většinou splněno, neboť hodnoty $\gamma_i < 1$ a $\gamma_i > 1$ mají tendenci se vzájemně kompenzovat.

Poznámka 87 Věta 47 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.

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 43 nebo transformaci proměnných studovanou ve větě 46. 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 $G^* = I$.

Věta 48 *(Lineární konvergence) Nechť jsou splněny předpoklady lemmatu 13, přičemž $\gamma_i \geq 1\ \forall i \in N$. Nechť $x_i \to x^*$ a $G^* = I$. Pak platí*

$$\sum_{i=1}^{\infty} \|e_i\| = 0.$$

Důkaz Jelikož $x_i \to x^*$, 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 47 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, 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 13 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 15 a poznámky 24.

V dalších úvahách budeme používat princip omezeného znehodnocení zformulovaný v následujícím lemmatu.

Lemma 14 *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). 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 24), 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 $\exp(\overline{C})$ (limita se snadno určí pomocí l'Hospitalova pravidla). Platí tedy $\kappa_i \leq \kappa_1 \exp(\overline{C})$ $\forall i \in N$.

Poznámka 88 Podle lemmatu 14 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 14 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 17), budeme požadovat splnění předpokladů věty 32 ($\rho_i = 1$, $\gamma_i = 1$, $\forall i \in N$). 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 43 (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 89 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

$$\|uv^T\|_F^2 = Tr(vu^T uv^T) = u^T u Tr(vv^T) = u^T u v^T v.$$

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 15 *Uvažujme aktualizaci*

$$R'_+ = R + \frac{zz^T}{z^T z} - \frac{Rz(Rz)^T}{z^T Rz} + \frac{\eta}{z^T Rz}\left(\frac{z^T Rz}{z^T z}z - Rz\right)\left(\frac{z^T Rz}{z^T z}z - Rz\right)^T.$$

Pak platí

$$\begin{aligned}\|R'_+ - I\|_F^2 \;=\; \|R - I\|_F^2 \;&-\; (1-\eta)\left(\left(1 - \frac{z^T R^2 z}{z^T R z}\right)^2 + 2\left(\frac{z^T R^3 z}{z^T R z} - \left(\frac{z^T R^2 z}{z^T R z}\right)^2\right)\right)\\ &-\; \eta\left(\left(1 - \frac{z^T R z}{z^T z}\right)^2 + 2\eta\left(\frac{z^T R^2 z}{z^T z} - \left(\frac{z^T R z}{z^T z}\right)^2\right)\right)\\ &-\; \eta(1-\eta)\left(\left(\frac{z^T R^2 z}{z^T R z}\right)^2 - \left(\frac{z^T R z}{z^T z}\right)^2\right).\end{aligned}$$

Důkaz Aplikujeme-li pravidla uvedená v poznámce 89 na vztah

$$R'_+ - I = R - I + \frac{z z^T}{z^T z} + \eta\frac{z^T R z}{z^T z}\frac{z z^T}{z^T z} - \eta\frac{z z^T R}{z^T z} - \eta\frac{R z z^T}{z^T z} + (\eta - 1)\frac{R z z^T R}{z^T R z},$$

získaný úpravou aktualizace z lemmatu 15, dostaneme

$$\begin{aligned}\|R'_+ - I\|_F^2 = \|R - I\|_F^2 \;&+\; 1 + \eta^2\left(\frac{z^T R z}{z^T z}\right)^2 + 2\eta^2\frac{z^T R^2 z}{z^T z} + (\eta-1)^2\left(\frac{z^T R^2 z}{z^T R z}\right)^2\\ &+\; 2\frac{z^T R z}{z^T z} + 2\eta\left(\frac{z^T R z}{z^T z}\right)^2 - 4\eta\frac{z^T R^2 z}{z^T z} + 2(\eta-1)\frac{z^T R^3 z}{z^T R z}\\ &-\; 2 - 2\eta\frac{z^T R z}{z^T z} + 4\eta\frac{z^T R z}{z^T z} - 2(\eta-1)\frac{z^T R^2 z}{z^T R z} + 2\eta\frac{z^T R z}{z^T z}\\ &-\; 4\eta\frac{z^T R z}{z^T z} + 2(\eta-1)\frac{z^T R z}{z^T z} - 4\eta^2\left(\frac{z^T R z}{z^T z}\right)^2\\ &+\; 2\eta(\eta-1)\left(\frac{z^T R z}{z^T z}\right)^2 + 2\eta^2\left(\frac{z^T R z}{z^T z}\right)^2 - 4\eta(\eta-1)\frac{z^T R^2 z}{z^T z}\\ = \|R - I\|_F^2 \;&-\; 1 + 2\eta\frac{z^T R z}{z^T z} - 2\eta^2\frac{z^T R^2 z}{z^T z} - 2(\eta-1)\frac{z^T R^2 z}{z^T R z}\\ &+\; 2(\eta-1)\frac{z^T R^3 z}{z^T R z} + \eta^2\left(\frac{z^T R z}{z^T z}\right)^2 + (\eta-1)^2\left(\frac{z^T R^2 z}{z^T R z}\right)^2.\end{aligned}$$

Stejný výsledek dostaneme roznásobením vztahu uvedeného v lemmatu 15.

Důsledek 9 *Jsou-li splněny předpoklady lemmatu 15 s* $0 \le \eta \le 1$, *platí* $\|R'_+ - I\|_F \le \|R - I\|_F$.

Důkaz Použijeme-li Schwarzovu nerovnost, dostaneme

$$\begin{aligned}\frac{z^T R^3 z}{z^T R z} - \left(\frac{z^T R^2 z}{z^T R z}\right)^2 &= \frac{z^T R^3 z z^T R z - (z^T R^2 z)^2}{(z^T R z)^2} \ge 0,\\ \frac{z^T R^2 z}{z^T z} - \left(\frac{z^T R z}{z^T z}\right)^2 &= \frac{z^T R z z^T z - (z^T R z)^2}{(z^T z)^2} \ge 0,\\ \left(\frac{z^T R^2 z}{z^T R z}\right)^2 - \left(\frac{z^T R z}{z^T z}\right)^2 &= \frac{(z^T R^2 z z^T z)^2 - (z^T R z)^4}{(z^T R z z^T z)^2}\\ &= \frac{z^T R^2 z z^T z + (z^T R z)^2}{z^T R z z^T z}\,\frac{z^T R^2 z z^T z - (z^T R z)^2}{z^T R z z^T z} \ge 0.\end{aligned}$$

Všechny závorky ve vztahu pro $\|R'_+ - I\|_F^2$ v lemmatu 15 jsou tedy nezáporné a jelikož $0 \le \eta \le 1$, platí $\|R'_+ - I\|_F^2 \le \|R - I\|_F^2$.

Lemma 16 *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 $0 \le \eta_i \le 1$ taková, že $x_i \to x^*$, kde x^* je stacionárním bodem funkce $F : R^n \to R$ 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 15 (viz důkaz věty 43). Použijeme-li důsledek 9 dostaneme

$$\begin{aligned}\|R_{i+1} - I\|_F &= \|\tilde{R}_{i+1} - I\|_F \le \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|\tilde{R}'_{i+1} - I\|_F \\ &= \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R'_{i+1} - I\|_F \le \|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F + \|R_i - I\|_F.\end{aligned}$$

Stačí tedy dokázat, že $\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F = (\|R_i - I\|_F + 1)O(\|e_i\|)$. Použijeme-li definiční vztah pro matici $\tilde{G}_i$ uvedený ve větě 43 a nerovnost (F5), můžeme psát

$$\begin{aligned}\|\tilde{G}_{i+1} - \tilde{G}_i\| &= \|\int_0^1 G(x_{i+1} + \lambda d_{i+1})\mathrm{d}\lambda - \int_0^1 G(x_i + \lambda d_i)\mathrm{d}\lambda\| \\ &\le \int_0^1 \|G(x_{i+1} + \lambda d_{i+1}) - G(x_i + \lambda d_i)\|\mathrm{d}\lambda \\ &\le \overline{L}\int_0^1 \|e_{i+1} + \lambda d_{i+1} - e_i - \lambda d_i\|\mathrm{d}\lambda \\ &\le \overline{L}\left(\|e_{i+1}\| + \|e_i\| + \frac{1}{2}\|d_{i+1}\| + \frac{1}{2}\|d_i\|\right) = O(\|e_i\|),\end{aligned}$$

neboť podle poznámky 23 platí $\|e_{i+1}\| = O(\|e_i\|)$ a $\|d_i\| = O(\|e_i\|)$. Platí tedy

$$\begin{aligned}\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| &\le \|H_{i+1}^{1/2}(\tilde{G}_{i+1} - \tilde{G}_i)H_{i+1}^{1/2}\| \le \|H_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \\ &= \|\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2}H_{i+1}\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \le \|\tilde{G}_i^{-1}\|\|R'_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| \\ &\le \frac{1}{\underline{G}}\|R'_{i+1}\|\|\tilde{G}_{i+1} - \tilde{G}_i\| = \|R'_{i+1}\|O(\|e_i\|).\end{aligned}$$

Ale

$$\|R'_{i+1}\| = \|I + R'_{i+1} - I\| \le 1 + \|R'_{i+1} - I\| \le 1 + \|R'_{i+1} - I\|_F \le 1 + \|R_i - I\|_F,$$

takže

$$\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\|_F \le \sqrt{n}\|\tilde{R}_{i+1} - \tilde{R}'_{i+1}\| = \|R'_{i+1}\|O(\|e_i\|) = (\|R_i - I\|_F + 1)O(\|e_i\|).$$

Důsledek 10 *Jsou-li splněny předpoklady lemmatu 16, existují konstanty $\overline{R}$ a $\overline{H}$ takové, že $\|R_i\| \le \overline{R}$ a $\|H_i\| \le \overline{H}$ $\forall i \in N$.*

Důkaz Jelikož $\|R_i\|_F \le \sqrt{n} + \|R_i - I\|_F$ a posloupnost $\|R_i - I\|_F$, $i \in N$, je podle lemmatu 16 a poznámky 88 omezená, je i posloupnost $\|R_i\| \le \|R_i\|_F$, $i \in N$, omezená. Jelikož

$$\|H_i\| = \|\tilde{G}_i^{-1/2}\tilde{G}_i^{1/2}H_i\tilde{G}_i^{1/2}\tilde{G}_i^{-1/2}\| \le \|\tilde{G}_i^{-1}\|\|R_i\| \le \frac{1}{\underline{G}}\|R_i\|,$$

je i posloupnost $\|H_i\|$, $i \in N$, omezená.

Poznámka 90 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 75), můžeme použít stejné úvahy jako v důkazu lemmatu 15 a lemmatu 16. 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 49 *(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 $F : R^n \to R$ splňující podmínky (F3)–(F5). Pak $x_i \to x^*$ superlineárně.*

Důkaz Z důkazu lemmatu 16 (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 10 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 48, 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 15. 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 10 a poznámky 90 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 $H_i \le \overline{H}$ a $H_i^{-1} \le 1/\underline{H}$ $\forall i \in N$ (důsledek 10 a a poznámka 90), ekvivalentní vztahu

$$\lim_{i\to\infty} \frac{\|(B_i - G^*)d_i\|}{\|d_i\|} = 0.$$

Závěr důkazu plyne z věty 17.

Poznámka 91 Věta 49 předpokládá, že platí $0 \le \eta_i \le 1$ (neboli $0 \le \beta_i \le 1$) $\forall i \in N$. Tento předpoklad nelze příliš zeslabit. Dá se pouze dokázat, že věta zůstane v platnosti, pokud $\beta_i \le 1$ $\forall i \in N$ 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 $\gamma_i \ne 1$ věta 49 neplatí a to zejména proto, že volba $\alpha_i = 1$ nemá při použití škálování žádné výsadní postavení.

4.7 Implementace metod s proměnnou metrikou

Poznámka 92 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 82) nebo použitím homogenního modelu (poznámka 83) a upravený tak, aby platilo $\underline{\rho} \le \rho \le \underline{\rho}$.
- Škálování (parametr γ): Vhodné škálování značně zvyšuje účinnost omezených metod s proměnnou metrikou (kdy $0 \le \eta \le 1$), pokud volíme parametr γ tak, aby platilo $b/c \le \rho/\gamma \le a/b$. Škálování v každé iteraci však není účelné, je třeba používat nějakou strategii, která omezuje použití hodnoty $\gamma \ne 1$ v těch iteracích, kde je to nevhodné. Nejvíce se osvědčilo řízené škálování popsané v poznámce 85.
- 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 (VM) 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 $x_1 \in R^n$ vypočteme $F_1 = F(x_1)$, $g_1 = g(x_1)$, zvolíme počáteční SPD 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 82) nebo použitím homogenního modelu (poznámka 83). Jestliže $\rho_i < \underline{\rho}$ nebo $\rho_i > \overline{\rho}$ položíme $\rho_i = 1$. Použijeme řízené škálování (poznámka 85) 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 (H) (pro metodu BFGS volíme $\eta_i = 1$). 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 (HD), metody BFGS (HB) a Hoshinovy metody (HH) uvedena metoda, která používá hodnotu

$$\eta_i = \frac{\max(0, \sqrt{c/a} - b^2/(ac))}{\max(\varepsilon, 1 - b^2/(ac)}, \tag{HP}$$

kde číslo $\varepsilon = 10^{-60}$ slouží k zaručení nenulovosti jmenovatele, který by mohl být nulový vlivem zaokrouhlovacích chyb. Hodnotu (HP) uvádíme, abychom demonstrovali, že efektivita metody BFGS může být překonána vhodnou volbou parametru η_i. Označení typu škálování v prvním sloupci tabulky má stejný význam jako v poznámce 85. Pro NS a PS se používá hodnota $\rho_i = 1$. Pro CS se používá parametr ρ_i určený zpětným použitím věty o střední hodnotě (poznámka 82). První sada sloupců odpovídá dimenzi $n = 50$ a druhá dimenzi $n = 200$.

Metoda	NIT	NFV	F	Čas	NIT	NFV	F	Čas
HD + NS	78846	83583	35	12.31	119772	131865	33	3:13.60
HD + PS	91398	93393	41	14.00	144317	148434	42	4:28.87
HD + CS	11682	15135	1	2.49	33103	44893	4	55.87
HB + NS	13199	21222	1	3.71	31532	56521	1	1:17.95
HB + PS	15009	16536	1	3.03	34376	38172	3	53.86
HB + CS	7982	9487	-	2.21	19091	22751	-	32.89
HH + NS	15509	21355	1	3.77	33869	49213	1	1:13.47
HH + PS	18623	19739	1	3.05	40325	42884	3	1:01.07
HH + CS	8449	9853	-	2.17	20864	24543	-	34.64
HP + NS	11244	17171	1	3.06	28521	44458	1	1:07.85
HP + PS	11903	13725	1	2.19	27039	29645	2	42.12
HP + CS	8183	9260	-	1.73	18430	20422	-	30.98
CG	57787	121776	6	12.84	136781	288655	10	3:18.42

Poznámka 93 Z výsledků uvedených v této tabulce lze učinit několik závěrů.

- Metoda DFP je velmi neefektivní.
- Řízené škálování velmi zvyšuje efektivitu metod s proměnnou metrikou.
- Efektivita metody BFGS může být překonána vhodnou volbou parametru η_i.
- Metody s proměnou metrikou jsou pro standardní (husté) úlohy menších rozměrů (řekněme 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 94 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 $F : R^n \to R$.

Definice 28 *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 $0 < \underline{\delta} < 1 < \overline{\delta}$, $0 < \underline{\sigma} < 1$ a $0 \leq \overline{\omega} < 1$, kde 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 kde čísla $0 < \Delta_i \leq \overline{\Delta}$, $i \in N$, se volí tak, že

$$\rho_i(s_i) < \overline{\rho} \Rightarrow \underline{\beta} \|s_i\| \leq \Delta_{i+1} \leq \overline{\beta} \|s_i\|, \tag{T3a}$$

$$\rho_i(s_i) \geq \overline{\rho} \Rightarrow \underline{\Delta}_i \leq \Delta_{i+1} \leq \overline{\Delta}_i, \tag{T3b}$$

kde $\underline{\Delta}_i = \min(\Delta_i, \overline{\gamma}\|s_i\|)$, $\overline{\Delta}_i = \min(\underline{\gamma}\underline{\Delta}_i, \overline{\Delta})$, $0 < \underline{\beta} \leq \overline{\beta} < 1 < \underline{\gamma} < \infty$, $1 < \overline{\gamma} \leq \infty$ a $0 \leq \underline{\rho} < \overline{\rho} < 1$, přičemž $\overline{\beta\delta} < 1$ a $\overline{\gamma}\underline{\delta} > 1$.

Poznámka 95 Označíme $N_1 \subset N$ množinu indexů takových, že $\|s_i\| < \underline{\delta}\Delta_i$, $N_2 \subset N$ množinu indexů takových, že $\rho_i(s_i) > \underline{\rho} \geq 0$, a $N_3 \subset N$ množinu indexů takových, že $\rho_i(s_i) \geq \overline{\rho} > \underline{\rho}$ (takže $N_3 \subset N_2$).

Poznámka 96 Jestliže $\overline{\omega} = 0$ nebo $\overline{\omega} > 0$, dostaneme přesné nebo nepřesné metody s lokálně omezeným krokem. Obvykle volíme $\underline{\rho} = 0$. Pokud $\underline{\rho} > 0$, dostaneme silnější tvrzení o globální konvergenci (věta 51). Číslo $\overline{\Delta} > 0$ slouží k omezení délky kroku, abychom se nedostali mimo definiční obor funkce $F : R^n \to R$. Číslo $\overline{\gamma}$ má pouze teoretický význam (věta 52), většinou pokládáme $\overline{\gamma} = \infty$. Vždy platí $\underline{\Delta}_i \leq \Delta_i$. Pokud $\overline{\gamma} < \infty$, může být $\underline{\Delta}_i < \Delta_i$. V tomto případě $\|s_i\| \to 0$ implikuje $\Delta_i \to 0$, což je někdy nevýhodné (věta 70). Poznamenejme, že z nerovnosti $\overline{\gamma}\underline{\delta} > 1$ plyne $\underline{\Delta}_i = \Delta_i$, pokud $i \notin N_1$.

Poznámka 97 Číslo $\underline{\sigma}$ není vnějším parametrem. Jeho existence musí být zaručena, ale jeho velikost závisí na zvolené metodě (obvykle $\underline{\sigma} = 1/2$). Čísla $\underline{\delta} < 1 < \overline{\delta}$ mají význam při přibližném výpočtu optimálního lokálně omezeného kroku. V ostatních případech lze položit $\underline{\delta} = 1 = \overline{\delta}$.

Poznámka 98 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 pouze tehdy, když $\|s_i\| \geq \underline{\delta}\Delta_i$ (viz důkaz věty **??**).

Poznámka 99 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.

Lemma 17 *Aplikujeme-li metodu s lokálně omezeným krokem (T1)-(T3) na funkci $F : R^n \to R$, která splňuje podmínku (F3), existuje konstanta $\underline{c} > 0$ taková, že*

$$\|s_i\| \geq \underline{c} m_i/M_i, \qquad (*)$$

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

$$F(x_i + s_i) - F(x_i) \geq \overline{\rho} Q_i(s_i) = \overline{\rho}\left(g_i^T s_i + \frac{1}{2} s_i^T B_i s_i\right) \geq \overline{\rho}\left(g_i^T s_i - \frac{1}{2}\|B_i\|\|s_i\|^2\right).$$

Z druhé strany podle (UB1) dostaneme

$$F(x_i + s_i) - F(x_i) \leq g_i^T s_i + \frac{1}{2}\overline{G}\|s_i\|^2,$$

což dohromady dává

$$\frac{1}{2}(\overline{G} + \overline{\rho}\|B_i\|)\|s_i\|^2 \geq (\overline{\rho} - 1) g_i^T s_i.$$

Z (T1c) dostaneme

$$-\underline{\sigma}\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|) \geq Q_i(s_i) \geq g_i^T s_i - \frac{1}{2}\|B_i\|\|s_i\|^2,$$

což spolu s předchozí nerovností dává

$$\begin{aligned} \frac{1}{2}(\overline{G} + \overline{\rho}\|B_i\|)\|s_i\|^2 &\geq (\overline{\rho} - 1) g_i^T s_i \geq \frac{1}{2}(\overline{\rho} - 1)\|B_i\|\|s_i\|^2 - \\ &- \underline{\sigma}(\overline{\rho} - 1)\|g_i\| \min(\Delta_i, \|g_i\|/\|B_i\|), \end{aligned}$$

neboli

$$\frac{1}{2}(\overline{G}+\|B_i\|)\|s_i\|^2 \geq \underline{\sigma}(1-\overline{\rho})\|g_i\|\min(\Delta_i, \|g_i\|/\|B_i\|),$$

takže buď $\|s_i\| \geq \underline{\delta}\Delta_i \geq \underline{\delta}\|g_i\|/\|B_i\| \geq \underline{\delta}m_i/M_i$, nebo

$$\frac{1}{2}(\overline{G}+\|B_1\|)\frac{M_i}{\|B_1\|}\|s_i\|^2 \geq \frac{1}{2}(\overline{G}+\|B_i\|)\|s_i\|^2 \geq \underline{\sigma}(1-\overline{\rho})\|g_i\|\Delta_i \geq \frac{\underline{\sigma}(1-\overline{\rho})}{\overline{\delta}}\|g_i\|\|s_i\|,$$

což dává $\|s_i\| \geq \left(2\underline{\delta}\underline{\sigma}(1-\overline{\rho})\|B_1\|/\left(\overline{\delta}(\overline{G}+\|B_1\|)\right)\right)m_i/M_i$.
(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

$$\underline{c} = \underline{\beta}\underline{\delta}\min\left((1-\overline{\omega}), \underline{\delta}, \frac{2\underline{\sigma}(1-\overline{\rho})\|B_1\|}{\overline{\delta}(\overline{G}+\|B_1\|)}, \frac{\|s_1\|\|B_1\|}{\|g_1\|}\right).$$

Poznámka 100 V důkazu lemmatu 17 jsme používali odhad $\|B_i s_i\| \leq \|B_i\|\|s_i\|$ obsahující normu matice B_i. Místo toho jsme mohli předpokládat existenci čísla $\overline{B_i}$ takového, že $\|B_i s_i\| \leq \overline{B_i}\|s_i\|$ (není třeba aby platilo $\|B_i s\| \leq \overline{B_i}\|s\|$ pro $s \neq s_i$). To znamená, že podmínku (T1c) můžeme zapsat ve tvaru

$$-Q_i(s_i) \geq \underline{\sigma}\|g_i\|\min(\Delta_i, \|g_i\|/\overline{B_i}),$$

kde $\|B_i s_i\| \leq \overline{B_i}\|s_i\|$. Tuto variantu použijeme v důkazu superlineární konvergence metod s lokálně omezeným krokem.

Věta 50 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)-(T3) taková, že*

$$\sum_{i=1}^{\infty}\frac{1}{M_i} = \infty,$$

kde M_i, $i \in N$, jsou čísla definovaná v lemmatu 17. Nechť funkce $F: R^n \to R$ splňuje podmínky (F1) a (F3). Pak platí

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

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 17 platí

$$\Delta_i \geq \frac{1}{\overline{\delta}}\|s_i\| \geq \frac{\underline{c}}{\overline{\delta}}\frac{\underline{\varepsilon}}{M_i} \qquad (*)$$

$\forall i \in N$. Protože $N_3 \subset N_2$, můžeme psát

$$F_i - F_{i+1} = F(x_i) - F(x_i + s_i) \geq -\overline{\rho} Q_i(s_i) \geq \overline{\rho}\underline{\sigma\varepsilon} \min\left(\Delta_i, \frac{\underline{\varepsilon}}{M_i}\right) \geq \frac{\overline{\rho}\underline{\sigma}\underline{\varepsilon}^2\underline{c}}{\overline{\delta}} \frac{1}{M_i}$$

$\forall i \in N_3$, 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) Nechť $i \in N$, nechť r je přirozené číslo takové, že $(\overline{\beta\delta})^{r-1}\underline{\gamma} < 1$ (takové číslo existuje neboť $(\overline{\beta\delta} < 1$ a $\underline{\gamma} < \infty$) 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 (T1a) a (T3) platí

$$\Delta_i \leq \underline{\gamma}^{p(i-1)}(\overline{\beta\delta})^{i-1-p(i-1)}\Delta_1 \leq \underline{\gamma}^{(i-1)/r}(\overline{\beta\delta})^{(r-1)(i-1)/r}\Delta_1 \leq \left(\underline{\gamma}(\overline{\beta\delta})^{(r-1)}\right)^{(i-1)/r}\Delta_1.$$

Protože podle předpokladu je $\underline{\gamma}(\overline{\beta\delta})^{r-1} < 1$, můžeme psát

$$\sum_{i\in N_4}\Delta_i \leq \sum_{i\in N_4}\left(\underline{\gamma}(\overline{\beta\delta})^{(r-1)}\right)^{(i-1)/r}\Delta_1 \leq \sum_{i=1}^{\infty}\left(\underline{\gamma}(\overline{\beta\delta})^{(r-1)}\right)^{(i-1)/r}\Delta_1 = \frac{\Delta_1}{1-\left(\underline{\gamma}(\overline{\beta\delta})^{(r-1)}\right)^{1/r}} < \infty.$$

Použijeme-li nyní (T1a) a $(*)$, dostaneme

$$\sum_{i\in N_4}\frac{1}{M_i} \leq \frac{1}{\underline{c\varepsilon}}\sum_{i\in N_4}\|s_i\| \leq \frac{\overline{\delta}}{\underline{c\varepsilon}}\sum_{i\in N_4}\Delta_i < \infty.$$

(c) Nyní stačí dokázat, že

$$\sum_{i\in N_5}\frac{1}{M_i} < \infty,$$

kde $N_5 = N \setminus N_4$, takže $N_5 = \{i \in N : rp(i) \geq i\}$. Označme

$$N_3 = \{i_1, i_2, i_3 \ldots\}, \qquad N_5 = \{k_1, k_2, k_3 \ldots\}$$

(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

$$rp(k_j) \geq k_j \geq j \quad \forall j \in N,$$

takže podle definice množiny N_6 dostaneme

$$l_j \leq l_{rp(k_j)} \leq i_{p(k_j)} \leq k_j \quad \forall j \in N,$$

neboť $i_{p(k_j)}$ je poslední prvek množiny $[1,k_j] \cap N_3$. Platí tedy $M_{l_j} \leq M_{k_j}\ \forall j \in N$, takže podle (a) dostaneme

$$\sum_{i\in N_5} \frac{1}{M_i} = \sum_{j=1}^{\infty} \frac{1}{M_{k_j}} \leq \sum_{j=1}^{\infty} \frac{1}{M_{l_j}} = \sum_{i\in N_6} \frac{1}{M_i} = r \sum_{i\in N_3} \frac{1}{M_i} < \infty.$$

Poznámka 101 Předpoklady věty 50 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 50 pozměnit tak, že

$$F_1 - \underline{F} \geq \lim_{i\to\infty} (F_1 - F_{i+1}) = \sum_{i=1}^{\infty} (F_i - F_{i+1}) \geq \sum_{i\in N_3} (F_i - F_{i+1}) \geq \frac{\overline{\rho}\sigma\underline{\varepsilon}^2 c}{\overline{\delta}} \sum_{i\in N_3} \frac{1}{\overline{B}}.$$

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 spolu s (T1a) dává $\|s_i\| \to 0$, což je ve sporu s $(*)$, neboť $m_i \geq \underline{\varepsilon}$ a $M_i \leq \overline{B}$.

Poznámka 102 Předpoklady věty 50 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

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

kde $C_1 > 0$ a $\overline{C} \geq 0$ jsou vhodné konstanty. V tomto případě platí

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \geq \frac{1}{C_1} + \sum_{i=1}^{\infty} \frac{1}{C_1 + \overline{C\delta\Delta} i} \geq \frac{1}{C_1} + \frac{1}{C_1 + \overline{C\delta\Delta}} \sum_{i=1}^{\infty} \frac{1}{i} = \infty,$$

neboť harmonická řada je divergentní.

V poznámce 15 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 51 *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 $F : R^n \to R$ splňuje podmínky (F1) a (F3). Pak platí*

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

Důkaz V poznámce 101 jsme ukázali, že platí

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

Předpokládejme, že

$$\limsup_{i\to\infty} \|g_i\| > \underline{\varepsilon} > 0.$$

Jelikož ke každému indexu $k \notin N_2$ existuje index $i \in N_2$ takový, že $x_k = x_i$, musí být podle tohoto předpokladu množina N_2 nekonečná a musí obsahovat nekonečnou podmnožinu $\overline{N}_2 \subset N_2$ takovou, že

$\|g_i\| \geq \underline{\varepsilon}\ \forall i \in \overline{N}_2$. 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 nerostoucí (pravidlo (T2)) a zdola omezená (pomínka (F1)), 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 (e) z důkazu věty 13, 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 $\|g_{k_j}\| \geq \underline{\varepsilon}\ \forall k_j \in \overline{N}_2$.

Věta 52 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem (T1)-(T3) s $\overline{\gamma} < \infty$, kde matice B_i, $i \in N$, jsou dostatečně omezené. 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 (F3) a (F4). Pak matice B_i, $i \in N$, 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) a (T3) platí

$$\|s_j\| \leq \overline{\delta}\Delta_j \leq \overline{\beta\delta}\|s_{j-1}\| \leq \ldots \leq (\overline{\beta\delta})^{j-k-1}\|s_{k+1}\| \leq \frac{1}{\overline{\beta}}(\overline{\beta\delta})^{j-k}\Delta_{k+1} \leq \frac{\overline{\gamma\gamma}}{\overline{\beta}}(\overline{\beta\delta})^{j-k}\|s_k\|$$

$\forall k < j < l$, neboli

$$\sum_{j=k}^{l-1} \|s_j\| \leq \frac{\underline{\gamma}\overline{\gamma}}{\overline{\beta}} \|s_k\| \sum_{j=k}^{l-1} (\overline{\beta\delta})^{j-k} \leq \frac{\underline{\gamma}\overline{\gamma}}{\overline{\beta}(1-\overline{\beta\delta})} \|s_k\| = \overline{D}\|s_k\|,$$

kde $\overline{D} = \underline{\gamma}\overline{\gamma}/(\overline{\beta}(1-\overline{\beta\delta})) > 1$, 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 14. 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

$$\frac{F_i - F_{i+1}}{\|g_i\|} \geq \frac{\overline{\rho}\underline{\sigma}}{\overline{\delta}} \frac{\|g_i\|\|s_i\|}{\|g_i\| + \frac{2}{\underline{G}}C_i\|g_i\|} \geq \frac{\overline{\rho}\underline{\sigma}\underline{G}}{3\overline{\delta}} \frac{\|s_i\|}{C_i} \geq \frac{\overline{\rho}\underline{\sigma}\underline{G}}{3\overline{\delta}\overline{C}} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j=1}^{i} \overline{C}\|s_j\|} \geq \frac{\overline{\rho}\underline{\sigma}\underline{G}}{3\overline{\delta}\overline{C}\overline{D}} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j \in N_3}^{i} \overline{C}\|s_j\|}$$

$\forall i \in N_3$, takže jako v důkazu věty 14 platí

$$\frac{\overline{\rho}\underline{\sigma}\underline{G}}{3\overline{\delta}\overline{C}\overline{D}} \sum_{i \in N_3} \frac{\overline{C}\|s_i\|}{C_1 + \sum_{j \in N_3}^{i} \overline{C}\|s_j\|} \leq \sum_{i \in N_3} \frac{F_i - F_{i+1}}{\|g_i\|} \leq \sum_{i=1}^{\infty} \frac{F_i - F_{i+1}}{\|g_i\|} \leq \frac{\sqrt{2\overline{G}}}{\underline{G}} \sqrt{F_1 - F^*}.$$

Existuje tedy číslo $\underline{\omega}$, takové, že

$$C_k \leq C_1 + \sum_{j=1}^{k-1} \overline{C}\|s_j\| \leq \overline{D}(C_1 + \sum_{j \in N_3}^{k-1} \overline{C}\|s_j\|) \leq \overline{D}C_1/\underline{\omega}$$

$\forall k \in N$ (viz důkaz věty 10), takže $\|B_k\| \leq C_k \leq \overline{B} \ \forall k \in N$, kde $\overline{B} = \overline{D}C_1/\underline{\omega}$. Z toho že $C_1 + \sum_{j=1}^{k-1} \overline{C}\|s_j\| \leq \overline{B} \ \forall k \in N$ plyne nerovnost

$$\sum_{i=1}^{\infty} \|s_i\| < \infty.$$

Věta 53 *(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 $F \in \mathcal{C}^2 : R^n \to R$, která vyhovuje podmínkám (F3) a (F4). 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 15. Nechť $i \in N_3$. Podle lemmatu 17 existuje index $k \le i$ takový, že $\|s_i\| \ge (\underline{c}/\overline{B})\|g_k\|$. Jelikož posloupnost $F(x_i)$, $i \in N$, je nerostoucí, platí

$$1 \ge \frac{F(x_i) - F(x^*)}{F(x_k) - F(x^*)} \ge \frac{\underline{G}^2}{2} \frac{\|x_i - x^*\|^2}{\|g_k\|^2} \ge \frac{\underline{G}^2}{2\overline{G}^2} \frac{\|g_i\|^2}{\|g_k\|^2}$$

(používáme vztahy (c)–(g) z důkazů vět 13 a 15), takže

$$\|s_i\| \ge \frac{1}{\sqrt{2}} \frac{\underline{cG}}{\overline{BG}} \|g_i\|.$$

Jelikož $i \in N_3$, můžeme použít (T1c) a (T2b), takže platí

$$F_i - F_{i+1} \ge \overline{\rho}\underline{\sigma}\|g_i\|^2 \min\left(\frac{1}{\sqrt{2}} \frac{\underline{cG}}{\overline{\delta BG}}, \frac{1}{\overline{B}}\right) = \frac{\overline{\rho}}{\sqrt{2}} \frac{\underline{\sigma cG}}{\overline{\delta BG}} \|g_i\|^2 \ge \frac{c}{\overline{G}} \|g_i\|^2.$$

kde

$$c = \frac{\overline{\rho}\underline{\sigma cG}}{2\overline{\delta B}} < \frac{\overline{\rho}\underline{\sigma G}}{\overline{\delta G}} < 1,$$

neboť podle lemmatu 17 platí

$$\underline{c} \le \frac{2\underline{\sigma}(1-\overline{\rho})\|B_1\|}{\overline{\delta}(\overline{G} + \|B_1\|)} < 2\frac{\overline{B}}{\overline{G}}.$$

Použijeme-li ještě jednou vztah (g) z důkazu věty 15, dostaneme

$$F_{i+1} - F^* \le \left(1 - c\frac{\underline{G}}{\overline{G}}\right)(F_i - F^*) \quad \forall i \in N_3,$$

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 $j \notin N_3$ $\forall i < j \le k$. Pak platí

$$\|s_j\| \le \overline{\delta}\Delta_j \le \overline{\beta\delta}\|s_{j-1}\| \le \ldots \le (\overline{\beta\delta})^{j-i-1}\|s_{i+1}\| \qquad (*)$$

$\forall i < j \le k$. V (a) jsme ukázali, že $\|s_i\| \ge c_0\|g_i\| \ge c_0\underline{G}\|e_i\|$, kde $c_0 = (1/\sqrt{2})(\underline{cG})/(\overline{BG})$. Jelikož posloupnost $F(x_i)$, $i \in N$, je nerostoucí, pak pro libovolný index $j > i$ platí

$$1 \ge \frac{F(x_j) - F(x^*)}{F(x_i) - F(x^*)} \ge \frac{\underline{G}}{\overline{G}} \frac{\|e_j\|}{\|e_i\|},$$

neboli $\|e_j\| \le \sqrt{\overline{G}/\underline{G}}\,\|e_i\|$. Můžeme tedy psát

$$\|s_{i+1}\| = \|x_{i+2} - x_{i+1}\| \le \|e_{i+2}\| + \|e_{i+1}\| \le 2\sqrt{\frac{\overline{G}}{\underline{G}}}\|e_i\| \le \frac{2}{c_0\underline{G}}\sqrt{\frac{\overline{G}}{\underline{G}}}\|s_i\| \triangleq \overline{c}\|s_i\|,$$

což po dosazení do $(*)$ dává

$$\sum_{j=i}^{k} \|s_j\| \le \|s_i\| + \sum_{j=i+1}^{k} \overline{c}(\overline{\beta\delta})^{j-i-1}\|s_i\| \le \left(1 + \frac{\overline{c}}{1 - \overline{\beta\delta}}\right)\|s_i\|,$$

takže podle (a) platí

$$\sum_{i=1}^{\infty}\|x_i - x^*\| = \sum_{i\in N_2}\|s_i\| \le \left(1 + \frac{\overline{c}}{1-\overline{\beta\delta}}\right)\sum_{i\in N_3}\|s_i\| < \infty.$$

Poznámka 103 Ve větě 53 jsme nepředpokládali, že $\overline{\gamma} < \infty$. Pokud $\overline{\gamma} < \infty$, platí silnější tvrzení věty 52.

Věta 54 *(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 $F : R^n \to R$ splňuje podmínky (F3) a (F4). Nechť*

$$\lim_{i\to\infty}\overline{\omega}_i = 0, \tag{α}$$

$$\lim_{i\to\infty}\frac{\|(B_i - G_i)s_i\|}{\|s_i\|} = 0. \tag{β}$$

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz Nechť $0 < \underline{G} < \underline{\lambda}(G^*)$ a $\overline{G} > \overline{\lambda}(G^*)$. Důkaz provedeme v několika krocích.
(a) Ukážeme, že existuje index $k_2 \in N$ takový, že

$$\|g_i\| \ge \frac{1}{2}\underline{G}\|s_i\|$$

a

$$-Q_i(s_i) \ge \frac{\sigma\underline{G}^2}{4\overline{G}}\|s_i\|^2$$

$\forall i \ge k_2$. 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

$$\|B_i s_i\| \le \overline{\lambda}(G_i)\|s_i\| + \|\vartheta_i\|\|s_i\|,$$

$$s_i^T B_i s_i \ge \underline{\lambda}(G_i)\|s_i\|^2 - \|\vartheta_i\|\|s_i\|^2$$

a jelikož $\|\vartheta_i\| \to 0$ (podle (β)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) < \underline{G}$, $\overline{\lambda}(G_i) \to \overline{\lambda}(G^*) < \overline{G}$, existuje index $k_2 \in N$ takový, že $\|B_i s_i\| \le \overline{G}\|s_i\|$ a $s_i^T B_i s_i \ge \underline{G}\|s_i\|^2$ $\forall i \ge k_2$. Z definice $Q_i(s_i)$ pak plyne plyne

$$0 \ge Q_i(s_i) = g_i^T s_i + \frac{1}{2}s_i^T B_i s_i \ge \frac{1}{2}\underline{G}\|s_i\|^2 - \|g_i\|\|s_i\|,$$

což dává $\|g_i\| \ge (\underline{G}/2)\|s_i\|$ $\forall i \ge k_2$. Použijeme-li (T1c) a přihlédneme-li k poznámce 100, můžeme psát

$$-Q_i(s_i) \ge \underline{\sigma}\|g_i\|\min(\|s_i\|, \|g_i\|/\overline{G}) \ge \frac{\sigma\underline{G}}{2}\min(1, \frac{\underline{G}}{2\overline{G}})\|s_i\|^2 = \frac{\sigma\underline{G}^2}{4\overline{G}}\|s_i\|^2.$$

(b) Ukážeme, že existuje index $k_3 \ge k_2$ tak, že $i \in N_3$ $\forall i \ge k_3$. 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

$$\left|\frac{s_i^T(G_i - B_i)s_i + o(\|s_i\|^2)}{2Q_i(s_i)}\right| \leq \frac{2\overline{G}}{\underline{\sigma G}^2}\frac{\|\vartheta_i\|\|s_i\|^2 + o(\|s_i\|^2)}{\|s_i\|^2} \to 0,$$

neboť $\|\vartheta_i\| \to 0$. Platí tedy $\rho_i(s_i) \to 1$ a jelikož $\overline{\rho} < 1$, existuje index $k_3 \geq k_2$ takový, že $\rho_i(s_i) \geq \overline{\rho}\ \forall i \geq k_3$.
(c) Ukážeme, že existuje index $k \geq k_3$ 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 $k \geq k_3$ 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 $i \in N_3\ \forall i \geq k \geq k_3$. 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 $i \geq k_3$, $i \in N_1$, a označme $\omega_i = \omega_i(s_i)$. Podle (α), (β) 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 17 se dá ukázat, že existuje index $k_4 \geq k_3$, $k_4 \in N_1$, takový, že

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

$\forall i \geq k_4$, $i \in N_1$. Použijeme-li větu 3, můžeme pro $i \geq k_4$ 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 $i \geq k_4 \geq k_3$. 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 $i \geq k_4$). Pak z nerovnosti $\underline{G}\|s_i\| \leq \|g_i\|$, platící pro $i \geq k_4$, $i \in N_1$, 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 $k \geq k_4$, $k \in N_1$ 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 $\Delta_{i+1} \geq \min(\overline{\gamma}\|s_i\|, \Delta_i)$, což dává

$$\|s_{i+1}\| < \|s_i\|/\overline{\delta} \leq \min(\overline{\gamma}\|s_i\|, \Delta_i) \leq \Delta_{i+1},$$

neboť $\overline{\gamma}\overline{\delta} > 1$, 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 29 *Metody s optimálním lokálně omezeným krokem používají místo podmínky (T1c) silnější podmínku*

$$Q_i(s_i) \leq \underline{\delta}^2 Q_i(s_i^*), \tag{$\overline{\text{T1}}$c}$$

kde

$$s_i^* = \arg \min_{\|s\| \leq \Delta_i} Q_i(s), \tag{$\overline{\text{T1}}$d}$$

přičemž $\|s_i^*\| = \Delta_i$, pokud toto minimum není jediné. Navíc používají hodnotu $\overline{\omega} = 0$ v podmínce (T1a).

Věta 55 *Směrový vektor $s_i^* \in R^n$ určený podle ($\overline{\text{T1}}$d) vyhovuje podmínkám (T1) s $\underline{\delta} = 1 = \overline{\delta}$, $\overline{\omega} = 0$ a $\underline{\sigma} = 1/2$.*

Důkaz (a) Podmínka (T1a) je přímo součástí podmínky ($\overline{\text{T1}}$d). Předpokládejme, že $s_i^* \in R^n$ je řešením úlohy ($\overline{\text{T}}$1d), 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 $\|s_i^*\| = \Delta_i$. Položme

$$s = -\frac{g_i^T g_i}{g_i^T B_i g_i} g_i$$

a předpokládejme, že $\|s\| \leq \Delta_i$ a $g_i^T B_i g_i > 0$. Pak platí

$$-Q_i(s) = \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} - \frac{1}{2} \frac{(g_i^T g_i)^2 g_i^T B_i g_i}{(g_i^T B_i g_i)^2} = \frac{1}{2} \frac{(g_i^T g_i)^2}{g_i^T B_i g_i} \geq \frac{1}{2} \|g_i\|^2 / \|B_i\|.$$

Podle ($\overline{\text{T1}}$d) musí být $Q_i(s_i^*) \leq Q_i(s)$, takže nutně

$$-Q_i(s_i^*) \geq -Q_i(s) \geq \frac{1}{2} \|g_i\|^2 / \|B_i\|.$$

(c) Nechť $\|s_i^*\| = \Delta_i$ a buď $\|s\| > \Delta_i$ nebo $g_i^T B_i g_i \leq 0$, kde $s \in R^n$ je vektor definovaný v (b). Jestliže $\|s\| > \Delta_i$ a $g_i^T B_i g_i > 0$, pak $\|g_i\|^3 / g_i^T B_i g_i > \Delta_i$ neboli

$$g_i^T B_i g_i < \|g_i\|^3 / \Delta_i.$$

Stejná nerovnost platí pro $g_i^T B_i g_i < 0$. Položme $\tilde{s} = -(\Delta_i / \|g_i\|) g_i$ takže $\|\tilde{s}\| \leq \Delta_i$. Pak platí

$$-Q_i(\tilde{s}) = \Delta_i \|g_i\| - \frac{1}{2} \frac{\Delta_i^2}{\|g_i\|^2} g_i^T B_i g_i > \Delta_i \|g_i\| - \frac{1}{2} \Delta_i \|g_i\| = \frac{1}{2} \Delta_i \|g_i\| = \frac{1}{2} \|s_i^*\| \|g_i\|,$$

neboť $\|s_i^*\| = \Delta_i$. Podle ($\overline{\text{T}}$1d) musí být $Q_i(s_i^*) \leq Q_i(\tilde{s})$ takže nutně

$$-Q_i(s_i^*) \geq -Q_i(\tilde{s}) \geq \frac{1}{2} \|g_i\| \|s_i\|.$$

Poznámka 104 Podle definice 29 a věty 55 splňuje metoda s optimálním lokálně omezeným krokem podmínku (T1c) s $\underline{\sigma} = \underline{\delta}^2/2$.

5.3 Výpočet optimálního lokálně omezeného kroku

Věta 56 *Vektor $s_i^* \in R^n$ je řešením úlohy ($\overline{T1}d$) 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 $\lambda_i^* \geq 0$. 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 $v = \pm(s - s_i^*)/\|s - s_i^*\|$, $s \neq s_i^*$, 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 $\|s\| \leq \|s_i\| = \Delta_i$.

Tvrzení věty 56 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 57 *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 ($\overline{T1}c$).*

Důkaz Zřejmě

$$\begin{aligned} Q(s) &= g^T s + \frac{1}{2}s^T B s = -s^T(B + \lambda I)s + \frac{1}{2}s^T B s \\ &= -\frac{1}{2}\left(s^T(B + \lambda I)s + \lambda s^T s\right) \leq -\frac{1}{2}\underline{\delta}^2\left(s^T(B + \lambda I)s + \lambda\Delta^2\right) \end{aligned}$$

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 57 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 18 *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 11 *Nechť jsou splněny předpoklady lemmatu 18. Nechť λ_i, $1 \leq i \leq n$, jsou vlastní čísla matice B (seřazená vzestupně) a v_i, $1 \leq i \leq n$, jim odpovídající ortonormální vlastní vektory. Nechť $g = \sum_{1=1}^n \gamma_i v_i$ (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 18 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 ortogonalitu matice V, dostaneme dokazované tvrzení.

Poznámka 105 Jelikož požadujeme, aby matice $B + \lambda^* I$ byla positivně semidefinitní, musí platit $\lambda^* \geq -\lambda_1$, kde λ_1 je nejmenší vlastní číslo matice B. Abychom zjednodušili některé úvahy, budeme bez újmy na obecnosti předpokládat, že nejmenší vlastní číslo λ_1 je jednoduché. Budeme rozlišovat dva případy: regulární případ, kdy $\lambda^* > -\lambda_1$, a singulární případ, kdy $\lambda^* = -\lambda_1$. 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 108). 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 106 Jestliže $\gamma_1 = v_1^T g \neq 0$, lze se snadno přesvědčit (ze vzorců uvedených v důsledku 11), že platí

$$\lim_{\lambda \to \lambda_1} \phi(\lambda) = \frac{1}{\Delta}, \quad \lim_{\lambda \to \lambda_1} \phi'(\lambda) = -\frac{1}{|\gamma_1|}$$

a

$$\lim_{\lambda \to \infty} \phi(\lambda) = -\infty, \quad \lim_{\lambda \to \infty} \phi'(\lambda) = -\frac{1}{\|g\|}.$$

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 107 V singulárním případě nelze použít Newtonovu metodu, neboť $\phi(\lambda) < 0 \ \forall \lambda > \lambda_1$. V tomto případě lze vektor $s(\lambda^*)$ vyjádřit ve tvaru $s(\lambda^*) = s + \alpha v_1$, kde s je libovolné řešení rovnice $(B - \lambda_1 I)s = -g$ 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 ($\overline{\text{T}}$1c). Kvantitativní vztahy udává následující věta.

Věta 58 *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 ($\overline{T}$1c).

Důkaz Tak jako v důkazu věty 57 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 57, dostaneme dokazované tvrzení.

Poznámka 108 Abychom zabránili selhaní 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 109 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 67 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 Určíme $\underline{\mu}$ jako maximální diagonální prvek matice $-B$. Položíme $\overline{\lambda} = \|g\|/\Delta + \|B\|$, $\underline{\lambda} = \max(0, \underline{\mu}, \|g\|/\Delta - \|B\|)$ 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 67), 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 $\underline{\delta}\Delta \leq \|s\| \leq \overline{\delta}\Delta$ ukončíme výpočet. Jestliže $\|s\| < \underline{\delta}\Delta$ a $\lambda = 0$ ukončíme výpočet. Jestliže $\|s\| < \underline{\delta}\Delta$ a $\lambda \neq 0$ položíme $\overline{\lambda} = \lambda$ a přejdeme na krok 5.

Krok 5 Určíme vektor $v \in R^n$ tak, aby tento vektor byl dobrou aproximací vlastního vektoru matice B příslušného vlastnímu číslu $\underline{\lambda}(B)$ a aby platilo $\|v\| = 1$ a $v^T s \geq 0$ (tento vektor lze určit z rozkladu $R^T R$ způsobem, který používají programy knihovny LAPACK). Určíme číslo $\alpha \geq 0$ tak, aby platilo $\|s + \alpha v\| = \Delta$ (poznámka 111). Jestliže $\alpha^2 \|Rv\|^2 \leq (1 - \underline{\delta}^2)(\|Rs\|^2 + \lambda\Delta^2)$, položíme $s := s + \alpha v$ a ukončíme výpočet. V opačném případě položíme $\underline{\mu} = \lambda - \|Rv\|^2$, $\underline{\lambda} = \max(\underline{\mu}, \underline{\lambda})$ 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

5.4 Využití směru největšího spádu (metody psí nohy)

Nevýhodou metod s optimálním lokálně omezeným krokem je nutnost řešení úlohy ($\overline{\text{T1}}$d), 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 ($\overline{\text{T1}}$d) často nahražuje úlohou

$$s_i = \arg \min_{\|s(\alpha,\beta)\| \le \Delta_i} Q_i(s(\alpha,\beta)), \qquad (\tilde{\text{T1}}\text{d})$$

kde

$$s(\alpha,\beta) = \alpha g_i + \beta B_i^{-1} g_i.$$

Úloha ($\tilde{\text{T1}}$d) 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 ($\tilde{\text{T1}}$d) vyhovuje opět podmínkám (T1) s $\overline{\omega} = 0$ 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 ($\tilde{\text{T1}}$d) 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 kroku s_C a Newtonova kroku s_N, kde

$$s_C = -\frac{g^T g}{g^T B g} g, \qquad s_N = -B^{-1} g.$$

Cauchyův krok je spádovým směrem právě tehdy, platí-li $g^T B g > 0$. Proto budeme rozlišovat dva případy, buď $g^T B g > 0$ nebo $g^T B g \le 0$. Jestliže $g^T B g \le 0$, můžeme položit $s = -(\Delta/\|g\|)g$, neboť v tomto případě pro $\alpha \ge 0$ platí

$$s^T(g + \alpha B s) = -\frac{\Delta}{\|g\|}\left(g^T g - \frac{\alpha\Delta}{\|g\|} g^T B g\right) \le -\frac{\Delta}{\|g\|} g^T g,$$

takže kvadratická funkce $Q(x + \alpha s)$ (funkce proměnné α, jejíž derivace je $s^T(g + \alpha B s)$) klesá pro $\alpha \ge 0$. Jestliže $g^T B g > 0$ a $\|s_C\| \ge \Delta$, můžeme položit $s = (\Delta/\|s_C\|)s_C$. Platí totiž $g = -(s_C^T B s_C / s_C^T s_C)s_C$, což dává

$$s_C^T(g + \alpha B s_C) = -\frac{s_C^T B s_C}{s_C^T s_C} s_C^T s_C + \alpha s_C^T B s_C = -(1-\alpha) s_C^T B s_C = -(1-\alpha)\frac{(g^T g)^2}{g^T B g},$$

takže kvadratická funkce $Q(x + \alpha s_C)$ klesá pro $0 \le \alpha < 1$ a nabývá minima pro $\alpha = 1$.

Pokud $g^T B g > 0$ a $\|s_C\| < \Delta$, mohou opět nastat dva případy, buď $(s_N - s_C)^T s_C \ge 0$ nebo $(s_N - s_C)^T s_C < 0$. Vyšetříme nejprve případ, kdy $(s_N - s_C)^T s_C \ge 0$. V tomto případě, který nastane například tehdy, je-li matice B positivně definitní, lze volit směrový vektor d jako průsečík hranice oblasti určené omezením $\|d\| \le \Delta$ a úsečky spojující body s_C a τs_N, kde $0 < \tau \le 1$ je vhodně zvolený parametr.

Věta 59 *Nechť $g^T B g > 0$ a $(s_N - s_C)^T s_C \ge 0$. Pak $0 < s_C^T s_C / s_C^T s_N \le 1$ a pokud*

$$s_C^T s_C / s_C^T s_N \le \tau \le 1, \qquad (*)$$

je kvadratická funkce $Q(s_C + \alpha(\tau s_N - s_C))$ nerostoucí pro $\alpha \le 1$ (pokud $(s_N - s_C)^T s_C > 0$, je tato funkce klesající pro $\alpha < 1$). Dále platí $\|\tau s_N\| \ge \|s_C\|$ (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 kolineární).

Důkaz Prostým dosazením dostaneme

$$(s_N - s_C)^T s_C = \frac{g^T g}{(g^T B g)^2}\left(g^T B g g^T B^{-1} g - (g^T g)^2\right),$$

takže nerovnost $(s_N - s_C)^T s_C \geq 0$ je splněna právě tehdy, jestliže platí $g^T B g g^T B^{-1} g - (g^T g)^2 \geq 0$ (je-li matice B positivně definitní plyne tato nerovnost ze Schwarzovy nerovnosti). Musí tedy platit $g^T B^{-1} g > 0$, neboli

$$\frac{s_C^T s_C}{s_N^T s_C} = \frac{(g^T g)^2}{(g^T B^{-1} g^T B g)} > 0,$$

což spolu s nerovností $(s_N - s_C)^T s_C \geq 0$ dává $0 < s_C^T s_C / s_C^T s_N \leq 1$. Je-li splněna nerovnost (*) můžeme psát

$$\begin{aligned}
(\tau s_N - s_C)^T s_C &= \frac{g^T g}{(g^T B g)^2} \left(\tau g^T B g g^T B^{-1} g - (g^T g)^2\right) \geq 0, \\
(\tau s_N - s_C)^T B (\tau s_N - s_C) &= \tau g^T B^{-1} g - 2\tau \frac{(g^T g)^2}{g^T B g} + \frac{(g^T g)^2}{g^T B g} \\
&= \frac{\tau}{g^T B g} \left(\tau g^T B g g^T B^{-1} g - (g^T g)^2\right) + (1-\tau)\frac{(g^T g)^2}{g^T B g} \geq 0,
\end{aligned}$$

přičemž poslední nerovnost je rovností právě tehdy, pokud $(s_N - s_C)^T s_C = 0$. Dále platí

$$\begin{aligned}
(\tau s_N - s_C)^T (g + B s_C) &= (\tau s_N - s_C)^T (\tau g + B s_C) + (1-\tau)(\tau s_N - s_C)^T g \\
&= (\tau s_N - s_C)^T B (B^{-1} g + s_C) + (1-\tau)(\tau s_N - s_C)^T g \\
&= -(\tau s_N - s_C)^T B (\tau s_N - s_C) - (1-\tau)\frac{g^T B g}{g^T g} s_C,
\end{aligned}$$

takže pro derivaci kvadratické funkce $Q(s_C + \alpha(\tau s_N - s_C))$ dostaneme

$$(\tau s_N - s_C)^T (g + B(s_C + \alpha(\tau s_N - s_C)) = -(\tau s_N - s_C)^T B (\tau s_N - s_C)(1-\alpha) - (1-\tau)\frac{g^T B g}{g^T g} s_C.$$

Pokud $\alpha \leq 1$, je tato derivace nekladná, takže funkce $Q(s_C + \alpha(\tau s_N - s_C))$ je nerostoucí (pokud $(s_N - s_C)^T s_C > 0$ je tato funkce klesající 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, jestliže $\tau s_N = s_C$. Zvolíme-li $\tau = s_C^T s_C / s_C^T s_N$ (nejmenší možná hodnota) platí

$$\tau \|s_N\| - \|s_C\| = \frac{\|s_C\|^2}{s_N^T s_C} \|s_N\| - \|s_C\| \geq \frac{\|s_C\|^2}{\|s_N\| \|s_C\|} \|s_N\| - \|s_C\| = 0,$$

přičemž rovnost nastane právě tehdy, jsou-li vektory s_C a s_N kolineární. Jelikož $s_N^T s_C > 0$, platí v tomto případě $\tau s_N = s_C$.

Poznámka 110 Případ, kdy $(s_N - s_C)^T s_C \geq 0$ je nejdůležitější, nastane například tehdy, je-li matice B positivně definitní. V tomto případě lze postupovat tak, že položíme $s = (\Delta / \|s_C\|) d_c$, pokud $s_C \geq \Delta$, $s = s_N$, pokud $\|s_N\| \leq \Delta$ a $s = s_C + \alpha(\tau s_N - s_C)$, pokud $\|s_C\| < \Delta < \|s_N\|$, kde α se volí tak aby platilo $\|d\| = \Delta$. Přitom buď $\tau = 1$ (klasická metoda psí nohy) nebo $\tau = \max(s_C^T s_C / s_C^T s_N, \Delta / \|s_N\|)$ (modifikovaná metoda psí nohy pocházející od Dennise a Meie). Ve všech uvedených případech je splněna podmínka (T1c) se $\sigma = 1/2$.

Poznámka 111 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 vzorce

$$\alpha = \frac{\sqrt{(v^T s)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} - v^T s}{\|v\|^2} = \frac{\Delta^2 - \|s\|^2}{\sqrt{(v^T s)^2 + (\Delta^2 - \|s\|^2)\|v\|^2} + v^T s}.$$

První vztah volíme pokud $v^T s \leq 0$ a druhý v opačném případě. Oba vztahy se zjednoduší, pokud $\|v\| = 1$.

Nyní vyšetříme případ, kdy $(s_N - s_C)^T s_C < 0$. V tomto případě není matice B positivně semidefinitní a nemá význam pokládat $s = s_N$, neboť tento bod není minimem kvadratické funkce $Q(s)$.

Věta 60 *Nechť $g^T Bg > 0$ a $(s_N - s_C)^T s_C < 0$. Pak kvadratická funkce $Q(s_C + \alpha(s_C - s_N))$ klesá pro $\alpha \geq 0$.*

Důkaz Z nerovnosti $(s_N - s_C)^T s_C < 0$ plyne $g^T Bg g^T B^{-1} g - (g^T g)^2 < 0$. 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

$$(s_C - s_N)^T (g + B(s_C + \alpha(s_C - s_N)) = (1 + \alpha)(s_C - s_N)^T B(s_C - s_N) < 0.$$

Dosavadní úvahy můžeme shrnout ve formě algoritmu.

Algoritmus 6 Data $\Delta > 0$.

Krok 1 Jestliže $g^T Bg \leq 0$ položíme $s = -(\Delta/\|g\|)g$ a ukončíme výpočet.

Krok 2 Vypočteme Cauchyův krok $s_C = -(g^T g/g^T Bg)g$. Jestliže $\|s_C\| \geq \Delta$, položíme $s = (\Delta/\|s_C\|)s_C$ a ukončíme výpočet.

Krok 3 Vypočteme Newtonův krok $s_N = -B^{-1}g$. 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 $s_C^T s_C / s_C^T s_N \leq \tau \leq 1$, zvolíme $\alpha > 0$ tak aby platilo $\|s_C + \alpha(\tau s_N - s_C)\| = \Delta$ (poznámka 111), 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 $\|s_C + \alpha(s_C - s_N)\| = \Delta$ (poznámka 111), položíme $s = s_C + \alpha(s_C - s_N)$ a ukončíme výpočet.

5.5 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 (PCG)

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

pro $1 \leq i \leq n$. Můžeme používat větu 28 a důsledek 2.

Poznámka 112 Určujeme-li lokálně omezený krok pomocí metody sdružených gradientů, zastavujeme iterační proces nejen tehdy, když $\|g_i\| \leq \omega\|g\|$, ale také tehdy, když $\|s_i\| < \Delta$ a buď $p_i^T B p_i \leq 0$ nebo $\|s_{i+1}\| \geq \Delta$. Pokud $p_i^T B p_i \leq 0$, platí

$$Q(s_i + \alpha_i p_i) = Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i \leq Q(s_i) - \alpha_i \|g_i\|^2 < Q(s_i)$$

pro libovolnou hodnotu $\alpha_i \geq 0$ (neboť $g_i^T p_i = -g_i^T g_i$ podle poznámky 54). Pokud $\|s_{i+1}\| \geq \Delta$, platí $Q(s_i + \alpha_i p_i) < Q(s_i)$ pro libovolnou hodnotu $\alpha_i > 0$ takovou, že $\|s_i + \alpha_i p_i\| \leq \Delta$ (neboť $Q(s_{i+1}) < Q(s_i)$ podle důsledku 2). V obou případech určíme číslo $\alpha_i \geq 0$ tak, aby platilo $\|s_i + \alpha_i p_i\| = \Delta$ (poznámka 111) a položíme $s = s_i + \alpha_i p_i$. Podle důsledku 3 platí $Q(s) \leq -(1/2)\|g\|^2/(\kappa(C)\|B\|)$ pro $i \geq 2$ a podle části (c) důkazu věty 55 platí $Q(s) \leq -(1/2)\|g\|\|s\|$ pro $i = 1$.

Poznámka 113 Předpodmíněná metoda sdružených gradientů generuje cestu v oblasti určené omezením $\|s\|_C \leq \Delta$, to znamená křivku $s(t) \in R^n$ takovou, že

$$\frac{d\|s(t)\|_C}{dt} > 0,$$

$$\frac{dQ(s(t))}{dt} < 0.$$

Proto je předpodmíněná metoda sdružených gradientů vhodná zejména pro metody s lokálně omezeným krokem používající omezení $\|s\|_C \leq \Delta$. Pokud $C \neq I$, mohou pro metody s lokálně omezeným krokem používající omezení $\|s\| \leq \Delta$ nastat jisté obtíže, neboť posloupnost $\|s_i\|$, $i \in N$, není v tomto případě monotonně rostoucí. Tyto obtíže jsou však vyváženy velkou efektivitou předpodmíněné metody sdružených gradientů.

Uvedené úvahy tvoří základ jednoduchého algoritmu:

Algoritmus 7 Data $C \succ 0$, $0 < \omega < 1$, $\Delta > 0$, $m \geq n$ (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.

(Obvykle volíme $m = n + 3$).

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 61 *Nechť jsou splněny předpoklady věty 28 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*

$$\frac{dQ(s_i + \alpha(s_N - s_i))}{d\alpha} = (1 - \alpha)(s_N - s_i)^T g_i < 0.$$

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 61 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 $0 < \Delta$, $m < n$

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 $m \leq 3$. Pro $m = 1$ dostaneme klasickou metodu psí nohy).

5.6 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*

$$q_0 = 0, \quad \gamma_1 q_1 = g$$

a

$$\delta_i = q_i^T B q_i, \quad \gamma_{i+1} q_{i+1} = B q_i - \delta_i q_i - \gamma_i q_{i-1} \tag{SL}$$

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 určeným maticí B a vektorem g.

Poznámka 114 Nechť $\gamma_i \neq 0$, $1 \leq i \leq k$, pro nějaký index $1 \leq k \leq n$. Pak podle (SL) platí $g = Q_k(\gamma_1 e_1)$ a

$$BQ_k = Q_k T_k + \gamma_{k+1} q_{k+1} e_k^T \qquad (\overline{\text{SL}}),$$

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 (SL).

Věta 62 *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 (SL) 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 (SL) 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 115 Jelikož $Q_k^T Q_k = I$ a $Q_k^T q_{k+1} = 0$ (důkaz věty 62), můžeme psát

$$Q_k^T B Q_k = T_k,$$

takže symetrický Lanczosův proces lze použít k tridiagonalizaci matice B.

Poznámka 116 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 $(\overline{\text{SL}})$ 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 63 *Vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 116 jsou shodné s vektory s_{i+1}, $1 \leq i \leq k$, generovanými metodou sdružených gradientů (algoritmus (PCG) s $C = I$). Navíc platí $\delta_1 = 1/\alpha_1$, $\varepsilon_1 = 1$ a*

$$\delta_{i+1} = \frac{\beta_i}{\alpha_i} + \frac{1}{\alpha_{i+1}}, \qquad \gamma_{i+1} = \frac{\sqrt{\beta_i}}{|\alpha_i|}, \qquad \varepsilon_{i+1} = -\varepsilon_i \text{sgn}(\alpha_i)$$

a

$$q_i = \varepsilon_i \frac{g_i}{\|g_i\|}$$

pro $1 \leq i \leq k$.

Důkaz Z důkazu Věty 20 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 116. 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 $D_k = \text{diag}\ (\varepsilon_1 \parallel g_1 \parallel, \ldots, \varepsilon_k \parallel g_k \parallel)$ (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 20 plyne, že matice $P_k^T BP_k$ je diagonální. Použijeme-li rekurentní vztahy metody CG, dostaneme $P_k^T BP_k = \text{diag}(\parallel g_1 \parallel^2 / \alpha_1, \ldots, \parallel g_k \parallel^2 / \alpha_k) = D_k \text{diag}(1/\alpha_1, \ldots, 1/\alpha_k) D_k$, takže

$$\begin{aligned} T_k &= Q_k^T BQ_k = D_k^{-1} G_k^T BG_k D_k^{-1} = D_k^{-1} B_k^T P_k^T BP_k B_k D_k^{-1} = \\ &= D_k^{-1} B_k^T D_k \text{diag}\ (1/\alpha_1, \ldots, 1/\alpha_k) D_k B_k D_k^{-1}. \end{aligned}$$

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 $(\overline{\text{SL}})$ 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 $\varepsilon_i \varepsilon_{i+1} = -\text{sgn}(\alpha_i)$ pro $1 \leq i \leq k$. Protože podle (SL) platí $\gamma_1 q_1 = g = g_1$ a $\gamma_1 \geq 0$, dostaneme $\varepsilon_1 = 1$.

Poznámka 117 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

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i, \|s\| \leq \Delta} \left(\frac{1}{2} s^T B s + g^T s \right). \qquad (\overline{\text{T1}}\text{e})$$

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

$$z_i = \arg \min_{z \in R^i, \|s\| \leq \Delta} \left(\frac{1}{2} z^T T_i z + \gamma_1 e_1^T z \right) \qquad (\overline{\text{T1}}\text{f})$$

(plyne to z úvah použitých v poznámce 116 a z toho, že ortogonalita matice Q_i implikuje $\|s\| = \|Q_i z\| = \|z\|$).

Je-li vektor z_i řešením úlohy ($\overline{\text{T1}}$f), zajímá nás, jak dobře aproximuje vektor s_{i+1} řešení úlohy ($\overline{\text{T1}}$d). Podle věty 56 je vektor z_i řešením úlohy ($\overline{\text{T1}}$f) 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 56. Kriteriem aproximace tedy může být hodnota $\|(B + \lambda_i) s_{i+1} + g\|$ (norma rezidua).

Věta 64 *Nechť $s_{i+1} = Q_i z_i$, kde vektor z_i je řešením úlohy ($\overline{\text{T1}}$f). 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 ($\overline{\text{SL}}$) 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 ($\overline{\text{T1}}$f).

Definice 31 *řekneme, že matice T_i (jejíž tvar je uveden v poznámce 114) je ireducibilní, jestliže $\gamma_j \neq 0$ $\forall 1 < j \leq i$.*

Věta 65 *Je-li matice T_i ireducibilní, nenastane v úloze ($\overline{\text{T1}}$f) singulární případ (matice $T_i + \lambda_i I$ je positivně definitní). Je-li matice T_n ireducibilní, nenastane singulární případ ani v úloze ($\overline{\text{T1}}$d). Nenastane-li singulární případ v úloze ($\overline{\text{T1}}$d) a platí-li $\gamma_{i+1} = 0$, je vektor $s_{i+1} = Q_i z_i$ řešením úlohy ($\overline{\text{T1}}$d).*

Důkaz (a) Je-li vektor z_i řešením úlohy ($\overline{\text{T1}}$f), 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 ($\overline{\text{T1}}$f) tedy nenastane singulární případ.

(b) Pro $i = n$ je úloha ($\overline{\text{T1}}$d) ekvivalentní úloze ($\overline{\text{T1}}$e) a tedy i úloze ($\overline{\text{T1}}$f). Je-li matice T_n ireducibilní, nenastane singulární případ v úloze ($\overline{\text{T1}}$f) a tedy ani v úloze ($\overline{\text{T1}}$d).

(c) Platí-li $\gamma_{i+1} = 0$, můžeme podle ($\overline{\text{SL}}$) 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 ($\overline{\text{T1}}$d) nenastane singulární případ takže podle poznámky 105 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 64, vidíme, že jsou splněny nutné a postačující podmínky pro to, aby vektor $s_{i+1} = Q_i z_i$ byl řešením úlohy ($\overline{\text{T1}}$d).

Poznámka 118 Symetrický Lanczosův proces můžeme předpodmínit. V tomto případě se používají rekurentní vztahy

$$w_0 = 0, \qquad \gamma_1 w_1 = g, \qquad q_1 = C^{-1} w_1$$

a

$$\delta_i = q_i^T B q_i, \quad \gamma_{i+1} w_{i+1} = B q_i - \delta_i w_i - \gamma_i w_{i-1}, \quad q_{i+1} = C^{-1} w_{i+1} \tag{PSL}$$

pro $1 \leq i \leq n$, kde koeficienty $\gamma_i \geq 0$, $1 \leq i \leq n$ se volí tak, aby platilo $w_i^T C^{-1} w_i = 1$, $1 \leq i \leq n$. Pak vektory q_i, $1 \leq i \leq n$ jsou C-ortogonální. Pro libovolný index $1 \leq k \leq n$ platí $Q_k^T C Q_k = I$ a

$$BQ_k = CQ_k T_k + \gamma_{k+1} w_{k+1} e_k^T, \tag{$\overline{\text{PSL}}$}$$

kde $T_k = Q_k^T B Q_k$ je symetrická tridiagonální matice. Poznamenejme, že je-li vektor z_i řešením problému ($\overline{\text{T1}}$f), kde matice T_i byla získána předpodmíněným symetrickým Lanczosovým procesem, je třeba v ($\overline{\text{T1}}$e) 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 119 Shrneme základní myšlenky, které se používají v algoritmu založeném na použití symetrického Lanczosova procesu.

1. Jelikož metoda sdružených gradientů je výpočetně ekonomičtější než symetrický Lanczosův proces, začínáme metodou CG. Přitom v každém iteračním kroku počítáme a ukládáme čísla γ_i, δ_i a vektory q_i (věta 63). Pokud nemůžeme metodu CG použít (pokud $p_i^T B p_i = 0$, nebo pokud číslo α_i je příliš velké), přejdeme na symetrický Lanczosův proces a začneme pokládat $s_{i+1} = Q_i z_i$, kde vektor z_i je řešením úlohy ($\overline{\text{T1}}$f).

2. 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 ($\overline{\text{T1}}$f).

3. 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 ($\overline{\text{T1}}$f).

Algoritmus 9 Data $0 < \omega < 1$, $\Delta > 0$, $\varepsilon > 0$, $m \leq n$ (obvykle $m = \min(n, 100)$).

Krok 1 Položíme $s_1 = 0$, $g_1 = g$, $p_1 = -g$, $q_1 = g/\|g\|$, $\beta_1 = \|g\|$, $\sigma_1 = g^T g$, $L = 1$, $M = 1$ 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 63, 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 63, 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 ($\overline{\text{T1}}$f) 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 $\varepsilon_{k+1} = -\varepsilon_k \text{sgn}(\alpha_k)$, $q_{k+1} = \varepsilon_{k+1} g_{k+1}/\|g_{k+1}\|$ 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.

5.7 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 56 a směrový vektor $s = s(\lambda)$ budeme hledat řešením úlohy

$$s(\lambda) = \arg\min_{\|s\| \leq \Delta} Q_\lambda(s), \qquad Q_\lambda(s) = \frac{1}{2} s^T (B + \lambda I) s + g^T s. \qquad (\overline{\text{T1}}\lambda)$$

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 ($\overline{\text{T1}}$d) 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 66 *Nechť pro daný index $1 \leq k \leq n$, je vektor s_k řešením úlohy*

$$s_k = \arg\min_{s \in \mathcal{K}_k, \|s\| \leq \Delta} Q(s), \qquad Q(s) = \frac{1}{2} s^T B s + g^T s \qquad (*)$$

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 $1 \le i \le j \le n$, pak podle věty 28 platí $\|s_i\| \le \|s_j\|$. Speciálně $\|s_k\| \le \|s_n\| = \|s^*\|$, 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 $1 \le k < n$. 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

$$B_1 - B_2 = B_2^{\frac{1}{2}}(B_2^{-\frac{1}{2}} B_1 B_2^{-\frac{1}{2}} - I)B_2^{\frac{1}{2}}, \quad B_2^{-1} - B_1^{-1} = B_1^{-\frac{1}{2}}(B_1^{\frac{1}{2}} B_2^{-1} B_1^{\frac{1}{2}} - I)B_1^{-\frac{1}{2}}$$

a z toho, že matice $B_2^{-\frac{1}{2}} B_1 B_2^{-\frac{1}{2}}$ a $B_1^{\frac{1}{2}} B_2^{-1} B_1^{\frac{1}{2}}$ 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

$$Q_\lambda(s) = \frac{1}{2}s^T(B+\lambda I)s + g^T s = \frac{1}{2}\tilde{s}^T Z_k^T(B+\lambda I)Z_k\tilde{s} + g^T Z_k\tilde{s} \triangleq \tilde{Q}_\lambda(\tilde{s})$$

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 $Z_k^T BZ_k + \lambda_1 I$, $Z_k^T BZ_k + \lambda_2 I$ 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

$$\lambda_2 \le \lambda_1 \quad \Longleftrightarrow \quad \|s_k(\lambda_2)\| \ge \|s_k(\lambda_1)\|.$$

Použijeme-li (d), dostaneme

$$\|s_k(\lambda)\|^2 = g^T Z_k(Z_k^T(B+\lambda I)Z_k)^{-2} Z_k^T g = g^T Z_k(Z_k^T BZ_k + \lambda I)^{-2} Z_k^T g.$$

Platí tedy

$$\|s_k(\lambda_2)\|^2 - \|s_k(\lambda_1)\|^2 = g^T Z_k\left[(Z_k^T BZ_k + \lambda_2 I)^{-2} - (Z_k^T BZ_k + \lambda_1 I)^{-2}\right] Z_k^T g.$$

Označíme-li $\tilde{B}_2 = (Z_k^T BZ_k + \lambda_2 I)$ a předpokláme-li, že $\lambda_2 \le \lambda_1$, můžeme psát

$$(Z_k^T BZ_k + \lambda_1 I)^2 - (Z_k^T BZ_k + \lambda_2 I)^2 = (\tilde{B}_2 + (\lambda_1 - \lambda_2)I)^2 - \tilde{B}_2^2 = 2(\lambda_1 - \lambda_2)\tilde{B}_2 + (\lambda_1 - \lambda_2)^2 I \succeq 0,$$

což spolu s první ekvivalencí v (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 (*) 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 56). 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 ($\overline{\text{T1}}\lambda$). Položíme-li $\lambda = \lambda_k$ pro nějaký index $k \leq n$, věta 66 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 28, 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.

Krok 1: Použijeme $k \ll n$ 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 5.3). Získáme přitom Lagrangeův multiplikátor λ.

Krok 3: Aplikujeme nepřesnou metodu s lokálně omezeným krokem na úlohu ($\overline{\text{T1}}\lambda$) a získáme tak vektor s, který je aproximací vektoru $s(\lambda)$.

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), 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 7.3 (realizované jako metody s lokálně omezeným krokem určovaným pomocí uvedených algoritmů).

N	Metoda	NIT	NFV	NFG	NCG	Čas
1000	A5	1918	1955	8797	-	4.65
	A6	2515	2716	11859	-	4.42
	A8	2292	2456	10673	12203	4.61
	A7	3329	3784	16456	53573	8.20
	A9	3107	3444	15306	55632	8.53
	PA7	2631	2823	13019	910	5.14
	PSA7	1999	2046	9201	1161	4.25
5000	A5	8391	8566	35824	-	122.44
	A6	9657	10133	42425	-	115.77
	A8	8938	9276	39032	47236	122.84
	A7	16894	19163	83933	358111	364:42
	A9	14679	16383	71483	366695	401.45
	PA7	10600	11271	50365	3767	145.42
	PSA7	8347	8454	35939	4329	108.87

5.8 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

$$\begin{bmatrix} R_{(i-1),(i-1)}, & R_{(i-1),i}, & R_{(i-1),(n-i)} \\ *, & B^{(i-1)}_{ii}, & B^{(i-1)}_{i,(n-i)} \\ *, & *, & B^{(i-1)}_{(n-i),(n-i)} \end{bmatrix},$$

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:

$$\gamma_i = \max_{i<j\leq n} (|B^{(i-1)}_{i,j}|),$$

$$\rho_i^2 = \max\left(|B^{(i-1)}_{ii}|, \frac{\gamma_i^2}{\beta^2}, \delta^2\right),$$

$$R_{ii} = \rho_i,$$

$$R_{i,(n-i)} = B^{(i-1)}_{i,(n-i)}/R_{ii},$$

$$B^{(i)}_{(n-i),(n-i)} = B^{(i-1)}_{(n-i),(n-i)} - R^T_{i,(n-i)}R_{i,(n-i)},$$

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

$$E_{ii} = \rho_i^2 - B_{ii}^{(i-1)} = \rho_i^2 + R_{i,(n-i)}R^T_{i,(n-i)} - B_{ii},$$

kde B_{ii} je prvek původní matice.

Věta 67 *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

$$B_{ii}^{(i-1)} = \rho_i^2 - E_{ii} = B_{ii} - R_{i,(n-i)}R^T_{i,(n-1)} \leq B_{ii} - \beta^2 < \|B\| - \|B\| = 0.$$

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 68 *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 $L^T P v = u$ dostaneme

$$v^T B v = v^T P^T L D L^T P v = u^T D u = \sum_{i=1}^{m} u_i^T D_{ii} u_i = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}).$$

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

$$v^T B v = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) \leq \underline{\lambda}(D_{kk}) < 0.$$

5.9 Newtonova metoda

Newtonova metoda používá matice $B_i = G(x_i), i \in N$, takže z (F3) plyne $\|B_i\| = \|G(x_i)\| \leq \overline{G}, i \in N$.

Věta 69 *Nechť jsou splněny podmínky (F1) a (F3). Pak Newtonova metoda realizovaná jako metoda s lokálně omezeným krokem je globálně konvergentní. Jsou-li navíc splněny podmínky (F4) a (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 50 (platí $\|B_i\| \leq \overline{G}, i \in N$). Jestliže $x_i \to x^*$, platí $B_i = G(x_i) \to G(x^*)$, neboli

$$\frac{\|(G^* - B_i)s_i\|}{\|s_i\|} \leq \|G^* - B_i\| \to 0,$$

což spolu s $\omega_i(s_i) \to 0$ implikuje Q-superlineární konvergenci (věta 54).

Nejpoužívanější jsou tyto realizace Newtonovy metody:

- Nepřesná Newtonova metoda ($\omega_i(s_i) > 0$). Jestliže platí (F3)-(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 $\Rightarrow$ méně než $O(n^3)$ operací na iteraci, což je výhodné pro rozsáhlé úlohy).
- Newtonova metoda s optimálním lokálně omezeným krokem. Pro tuto realizaci platí obzvláště silné tvrzení:

Věta 70 *Nechť jsou splněny předpoklady (F1)-(F3). Nechť x_i, $i \in N$, je posloupnost určená Newtonovou metodou s optimálním lokálně omezeným krokem s $\overline{\gamma} = \infty$. 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

$$\Delta_i \to 0 \quad \text{a} \quad i \notin N_3 \quad \forall i \in M \tag{u}$$

(neboť $\overline{\gamma} = \infty$). Ve druhém případě existuje podposloupnost x_i, $i \in M \subset N$, taková, že

$$\Delta_i \geq \underline{\Delta} \quad \text{a} \quad i \in N_3 \quad \forall i \in M, \tag{v}$$

kde $\underline{\Delta} > 0$. Vzhledem k tomu, že platí (F2), lze v obou případech tuto podposloupnost 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 (u) 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 (T1c) a (T1a) 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 definice $Q_i(s_i)$ a věty 3 dává

$$|\rho_i(s_i) - 1| = \left| \frac{F(x_i + s_i) - F(x_i)}{Q_i(s_i)} - 1 \right| = \frac{o(\|s_i\|^2)}{|Q_i(s_i)|} = o(\|s_i\|) \to 0,$$

což je ve sporu s předpokladem, že $i \notin N_3$.
(c) Předpokládejme, že platí (u) a $G^* \not\succeq 0$. 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 ($\overline{\text{T}}$1c) a ($\overline{\text{T}}$1d) platí

$$|Q_i(s_i)| \geq \underline{\delta}^2 |Q_i(s_i^*)| \geq \underline{\delta}^2 |Q_i(v_i)| = -\underline{\delta}^2 (v_i^T g_i + \frac{1}{2} v_i^T G_i v_i) \geq -\frac{\underline{\delta}^2}{2} \underline{\lambda}_i \Delta_i^2 \geq \frac{\underline{\delta}^2}{4\overline{\delta}^2} |\underline{\lambda}^*| \|s_i\|^2 \quad \forall i \in M,\ i \geq k,$$

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 (v). 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

$$s^* = \arg \min_{\|s\| \leq \underline{\Delta}/2} Q^*(s), \tag{w}$$

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

$$Q_i(s_i) \leq \underline{\delta}^2 Q_i(s_i^*) \leq \underline{\delta}^2 Q_i(x^* + s^* - x_i) \quad \forall i \in M,\ i \geq k.$$

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 (w), což je možné pouze tehdy, pokud $g(x^*) = 0$ a $G(x^*) \succeq 0$.
(e) Podle ($\overline{F}$4) 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). 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

$$|Q_i(s_i)| \geq \frac{\sigma \underline{G}^2}{4\overline{\delta}\overline{G}} \|s_i\|^2.$$

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 (u). Předpokládejme tedy, že platí (v). 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 $\|s_i\| < \min(\varepsilon/2, \underline{\delta\Delta})$, 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 (e) a (f) z důkazu věty 13) 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 (v). 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, $i \in N$.
(f) 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 (e).

Přestože Newtonova metoda, realizovaná jako metoda s 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.

6 Metody pro minimalizaci součtu čtverců

6.1 Gaussova-Newtonova metoda

Předpokládejme, že účelová funkce $F(x)$ má tvar

$$F(x) = \frac{1}{2}f^T(x)f(x) = \frac{1}{2}\sum_{k=1}^{m} f_k^2(x),$$

kde $f_k : R^n \to R$, $1 \leq k \leq m$, jsou dvakrát spojitě diferencovatelné funkce. Pak platí

$$g(x) = J^T(x)f(x) = \sum_{k=1}^{m} f_k(x)g_k(x),$$

$$G(x) = J^T(x)J(x) + C(x) = \sum_{k=1}^{m} g_k(x)g_k^T(x) + \sum_{k=1}^{m} f_k(x)G_k^T(x).$$

Gaussova-Newtonova metoda vznikne z Newtonovy metody tím, že ve výrazu pro $G(x_i)$ zanedbáme člen $C(x_i)$, takže

$$B_i = J_i^T J_i = \sum_{k=1}^{m} g_k(x_i) g_k^T(x_i).$$

Poznámka 120 Existují dva důvody pro použití takto definované matice B_i (vystupující v kvadratické funkci $Q_i(s)$ zavedené v poznámce 94):

1) Úlohy s nulovým reziduem ($F(x^*) = 0$). Z $x_i \to x^*$ plyne $F(x_i) \to F(x^*) = 0$ a tedy $f_k(x_i) \to 0$ $\forall 1 \le k \le m$. Jestliže $\|G_k(x)\| \le \overline{G}$, pak i

$$\|C(x_i)\| = \|\sum_{k=1}^{m} f_k(x_i) G_k(x_i)\| \le \overline{G} \sum_{k=1}^{m} |f_k(x_i)| \to 0$$

a tedy $\|G(x_i) - B_i\| = \|C(x_i)\| \to 0$ z čehož plyne Q-superlineární konvergence.

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 $B_i = J_i^T J_i$.

Pro další úvahy je třeba poněkud upravit podmínky kladené na funkci $F : R^n \to R$. Podmínka (F1) je splněna vždy, neboť $F(x) \ge 0$ $\forall x \in R^n$. Podmínku (F3) nahradíme podmínkou

$$\|G_k(x)\| \le \overline{G} \qquad (\overline{\text{F3}})$$

$\forall x \in R^n$, $\forall 1 \le k \le m$. Z (F2) a ($\overline{\text{F3}}$) plyne omezenost gradientů i funkčních hodnot

$$\|g_k(x)\| \le \overline{g},$$

$$|f_k(x)| \le \overline{f}$$

$\forall x \in \mathcal{L}(F(x_1))$, $\forall 1 \le k \le m$, a tudíž i (F3).

Věta 71 *Nechť jsou splněny podmínky (F2) a ($\overline{\text{F3}}$). Pak Gaussova-Newtonova metoda realizovaná jako metoda s lokálně omezeným krokem je globálně konvergentní. Jsou-li navíc splněny podmínky (F4) a (F5) a platí-li $x_i \to x^*$, $F(x^*) = 0$ a $\omega_i(s_i) \to 0$, je rychlost konvergence Q-superlineární.*

Důkaz Z (F2) a ($\overline{\text{F3}}$) plyne $\|G_k(x)\| \le \overline{G}$, $\|g_k(x)\| \le \overline{g}$, $|f_k(x)| \le \overline{f}$ $\forall x \in R^n$, $\forall 1 \le k \le m$. Platí tedy jednak

$$\|G(x)\| \le \sum_{k=1}^{m} \|g_k(x)\|^2 + \sum_{k=1}^{m} |f_k(x)| \|G_k(x)\| \le m\overline{g}^2 + m\overline{f}\,\overline{G}$$

(podmínka (F3)) a jednak

$$\|B_i\| = \|\sum_{k=1}^{m} g_k(x_i) g_k^T(x_i)\| \leq \sum_{k=1}^{m} \|g_k(x_i)\|^2 \leq m\overline{g}^2,$$

takže podle věty 50 je Gaussova-Newtonova metoda globálně konvergentní. Jak již bylo ukázáno 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 $\omega_i(s_i) \to 0$ implikuje Q-superlineární konvergenci (věta 54).

Poznámka 121 Směrový vektor odpovídající Gaussově-Newtonově metodě můžeme určit třemi různými způsoby:

1) Řešením normální soustavy rovnic. Rovnice $B_i s_i + g_i = 0$ má tvar

$$J_i^T J_i s_i + J_i^T f_i = 0. \tag{NE}$$

2) Řešením linearizované úlohy pro součet čtverců (přeurčené soustavy rovnic). Linearizovaná úloha má tvar

$$J_i s_i + f_i \approx 0. \tag{OE}$$

Používá se QR-rozklad $J_i = Q_i \begin{bmatrix} R_i \\ 0 \end{bmatrix}$, kde $Q_i^T Q_i = I$, takže

$$Q_i^T J_i s_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix} s_i = Q_i^T f_i$$

(R_i je horní trojúhelníková matice). QR-rozklad je stabilní a je možné určit pseudohodnost matice J_i a následně snížit dimenzi soustavy. 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 systémových 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

$$\begin{bmatrix} I & J_i \\ J_i^T & 0 \end{bmatrix} \begin{bmatrix} r_i \\ s_i \end{bmatrix} + \begin{bmatrix} f_i \\ 0 \end{bmatrix} = 0, \tag{SE}$$

což je soustava $m + n$ rovnic se symetrickou indefinitní maticí. Tento způsob je vhodný pro řídké úlohy nebo 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 váhy mohou být i nekonečné (úlohy s omezeními).

6.2 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:

1) Jestliže $F_i \to F^* = 0$, volíme Gaussovu-Newtonovu metodu.

2) Jestliže $F_i \to F^* > 0$, volíme nějakou superlineárně konvergentní metodu (buď Newtonovu metodu nebo metodu s proměnnou metrikou).

Věta 72 *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 122 Velmi efektivní hybridní metodu dostaneme, zkombinujeme-li Gaussovu-Newtonovu metodu s metodou BFGS: Nechť $B_1 = J_1^T J_1$. Jestliže $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme

$$B_{i+1} = J_{i+1}^T J_{i+1}.$$

Jestliže $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, položíme

$$B_{i+1} = B_i + \frac{y_i y_i^T}{y_i^T d_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i},$$

kde $d_i = x_{i+1} - x_i$ a $y_i = g_{i+1} - g_i = J_{i+1}^T f_{i+1} - J_i^T f_i$. Obvykle $\underline{\vartheta} = 0.01$ pro metody spádových směrů a $\underline{\vartheta} = 0.0001$ pro metody s lokálně omezeným krokem. V případě řídkých součtů čtverců je výhodné kombinovat Gaussovu-Newtonovu metodu s Newtonovou metodou (oddíl 7.7).

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

$$C_+ d = z \stackrel{\Delta}{=} y - J_+^T J_+ d = J_+^T f_+ - J^T f - J_+^T J_+ d,$$

která bezprostředně plyne z podmínky $B_+ d = y$. Dostaneme tak aktualizaci

$$C_+ = C + \frac{zz^T}{d^T z} - \frac{Cd(Cd)^T}{d^T Cd} + \frac{\beta}{d^T Cd}\left(\frac{d^T Cd}{d^T z} z - Bd\right)\left(\frac{d^T Cd}{d^T z} z - Bd\right)^T. \tag{C1}$$

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

$$C_+ = C + \frac{(z - Cd)(z - Cd)^T}{d^T (z - Cd)}. \tag{CR}$$

(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 (B), kde matice B je nahražena maticí $\bar{B}$. Protože $y - \bar{B}d = z - Cd$, je výhodné použít vzorec ($\overline{\text{B}}$). Pak

$$\begin{aligned} C_+ &= C + \frac{(y - \bar{B}d)v^T + v(y - \bar{B}d)^T}{d^T v} - \frac{(y - \bar{B}d)^T d}{d^T v}\frac{vv^T}{d^T v} \\ &= C + \frac{(z - Cd)v^T + v(z - Cd)^T}{d^T v} - \frac{(z - Cd)^T d}{d^T v}\frac{vv^T}{d^T v}, \end{aligned} \tag{C2}$$

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 (CR)).

Poznámka 123 Vektory y a z mohou být definovány různým způsobem, vždy ale musí 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

$$z = \sum_{k=1}^{m} f_k^+ B_k^+ s = \sum_{k=1}^{m} f_k^+ (g_k^+ - g_k) = (J_+ - J)^T f_+.$$

Jak již bylo zmíněno (poznámka 73), je možné metody s proměnnou metrikou pro součet čtverců realizovat v součinovém tvaru. Nyní se budeme zabývat strukturovanými metodami s proměnnou metrikou, které využijí 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 efektivně použít pouze metodu BFGS (poznámka 73), omezíme se pouze 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

$$(J_+ + L_+)^T z = y, \qquad (J_+ + L_+) d = z, \qquad z^T z = d^T y, \tag{$*$}$$

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 73 *Nechť T je SPD 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 (*) 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 $(A_+ - \bar{A})^T z = -z^T T z u = y - \bar{A}^T z$, 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 (*), 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 (*).

Poznámka 124 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 (*) plyne, že $z^T z = \lambda^2 d^T \bar{A}^T \bar{A} d = d^T y$, což po dosazeni do vztahu uvedeného ve větě 74 dává

$$L_+ = L + \frac{\bar{A}d}{d^T \bar{A}^T \bar{A} d}\left(\sqrt{\frac{d^T \bar{A}^T \bar{A} d}{d^T y}}\, y - \bar{A}^T \bar{A} d\right)^T. \qquad \text{(LB)}$$

Pokud $J_+ = 0$, takže $\bar{A} = A$, přejde tento výraz v (AB).

Jistá nevýhoda aktualizace (LB) spočívá v tom, že řešení přeurčené soustavy lineárních rovnic $(J+L)d+f \approx 0$ (lineárního problému nejmenších čtverců) 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 40). 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. Dá se ukázat, že 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 $P^2 = P$).

Věta 74 *Frobeniova norma $\|P(L_+ - L)\|_F$ je minimální na množině všech matic vyhovujících kvazi-newtonovské podmínce (*) a omezení $L_+^T f_+ = 0$ právě tehdy, platí-li*

$$L_+ = PL + \frac{\tilde{A}d}{d^T \tilde{A}^T \tilde{A} d} \left(\sqrt{\frac{d^T \tilde{A}^T \tilde{A} d}{d^T y}} \tilde{y} - \tilde{A}^T \tilde{A} d \right)^T . \qquad (\overline{\text{LB}})$$

kde

$$\tilde{A} = P(J_+ + L), \qquad \tilde{y} = y - \frac{J_+ f_+ (J_+ f_+)^T d}{f_+^T f_+}.$$

Důkaz (a) Nejprve ukážeme, že pokud $(J_+ + L_+)^T d = z$ a $L_+^T f_+ = 0$, je podmínka $(J_+ + L_+)^T z = y$ ekvivalentní podmínce $(J_+ + L_+)^T P z = \tilde{y}$. Z $(J_+ + L_+)^T d = z$ a $L_+^T f_+ = 0$ totiž plyne $f_+^T J_+ d = f_+^T z$, takže

$$\begin{aligned} (J_+ + L_+)^T P z - \tilde{y} &= J_+^T z - \frac{J_+^T f_+ f_+^T J_+ d}{f_+^T f_+} + L_+^T P z - y + \frac{J_+^T f_+ f_+^T J_+ d}{f_+^T f_+} \\ &= J_+^T z + L_+^T P z - y = (J_+ + L_+)^T z - y. \end{aligned}$$

Poznamenejme, že z rovností $(J_+ + L_+)^T d = z$ a $(J_+ + L_+)^T P z = \tilde{y}$ plyne vztah $z^T P z = d^T \tilde{y}$.
(b) Nutnost dokážeme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2} \|P(L_+ - L)\|_F^2 + u^T \left((J_+ + L_+)^T P z - \tilde{y} \right) \\ &= \sum_{i=1}^{m} \left[\frac{1}{2} \left(l_i^+ - l_i \right)^T P \left(l_i^+ - l_i \right) + u_i z^T P l_i^+ \right] + u^T (J_+^T P z - \tilde{y}), \end{aligned}$$

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

$$\frac{\partial L}{\partial l_i^+} = P \left(l_i^+ - l_i \right) + u_i P z.$$

Podmínka pro stacionaritu Langrangeovy funkce má tedy tvar $P(l_i^+ - l_i) + u_i P z = 0$, $1 \le i \le m$, neboli

$$P(L_+ - L) = -P z u^T.$$

Z rovnosti $(J_+ + L_+)^T P z = \tilde{y}$ dostaneme $(L_+ - L)^T P z = -z^T P z u = \tilde{y} - \tilde{A}^T z$, takže

$$u = -\frac{\tilde{y} - \tilde{A}^T z}{z^T P z},$$

což po dosazení do předchozí rovnosti dává

$$P(L_+ - L) = \frac{P z (\tilde{y} - \tilde{A}^T P z)^T}{z^T P z} \qquad (\#)$$

(neboť $P^2 = P$ implikuje $P\tilde{A} = \tilde{A}$). Použijeme-li druhou podmínku z (*), dostaneme $P(L_+ - L)d = Pz - \tilde{A}d$, takže lze psát

$$\frac{Pz(\tilde{y} - \tilde{A}^T Pz)^T d}{z^T Pz} = Pz - \bar{A}d.$$

Z posledního vyjádření je zřejmé, že vektor Pz je rovnoběžný s vektorem $\tilde{A}d$, neboli $Pz = \lambda\tilde{A}d$. Použijeme-li vztah $z^T Pz = d^T\tilde{y}$ dokázaný v (a), můžeme psát

$$\lambda^2 d^T\tilde{A}^T\tilde{A}d = z^T Pz = d^T\tilde{y} \quad \Rightarrow \quad \lambda = \pm\sqrt{\frac{d^T\tilde{y}}{d^T\tilde{A}^T\tilde{A}d}},$$

což po dosazení do $Pz = \lambda\tilde{A}d$ a potom do (#) dokazuje tvrzení věty.

Poznámka 125 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 (C1) nebo (C2) 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 122: Nechť $C_1 = 0$. Jestliže $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme $C_{i+1} = 0$. Jestliže $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, aktualizujeme matici C_i pomocí(C1) nebo (C2). 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 (LB) nebo ($\overline{\text{LB}}$)).

Poznámka 126 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 ($\overline{\text{B}}$) na matici $\bar{B} = J_+^T J_+ + \|f\|T$. Nakonec položíme $B_+ = J_+^T J_+ + \|f_+\|T_+$. Užitím vztahu ($\overline{\text{B}}$) dostaneme

$$\begin{aligned} T_+ &= T + \frac{1}{\|f\|}\left(\frac{(y - \bar{B}d)v^T + v(y - \bar{B}d)^T}{d^T v} - \frac{(y - \bar{B}d)^T d}{d^T v}\frac{vv^T}{d^T v}\right) \\ &= T + \frac{(\tilde{z} - Td)v^T + v(\tilde{z} - Td)^T}{d^T v} - \frac{(\tilde{z} - Td)^T d}{d^T v}\frac{vv^T}{d^T v}, \end{aligned} \tag{T2}$$

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 (HB)	9343	10853	10853	1.67	1
Gaussova-Newtonova metoda	8615	16302	24914	19.02	8
hybridní metoda (poznámka 122)	3809	6080	9884	8.89	2
strukturovaná metoda BFGS podle (C2)	3158	5897	9054	7.34	2
strukturovaná metoda BFGS podle (T2)	3262	6085	9345	6.97	1
metody s lokálně omezeným krokem	NIT	NFV	NFG	Čas	selhání
metoda BFGS podle (BB)	10684	11860	10764	3.39	1
Gaussova-Newtonova metoda	4321	4694	4402	12.84	1
hybridní metoda (poznámka 122)	3450	4013	3531	9.67	-
strukturovaná metoda BFGS podle (C2)	2766	3130	2847	7.61	-
strukturovaná metoda BFGS podle (T2)	2771	3239	2849	7.66	-

Poznámka 127 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 122) nebo pomocí strukturovaných aktualizací (C2) a (T2). 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 Metody pro rozsáhlé řídké a separovatelné úlohy

Rozsáhlé úlohy nemůžeme řešit metodami, které vyžadují uchovávání velkých hustých matic. Nejčastěji se pro tento účel používají některé speciální metody:

- Metody s proměnnou metrikou s omezenou pamětí.
- Diferenční verze nepřesné Newtonovy metody.
- Metody pro řídké úlohy (N, VM).
- Metody pro separovatelné úlohy (N, VM).
- Řídké modifikace Gaussovy-Newtonovy metody pro součet čtverců.

7.1 Metody s proměnnou metrikou s omezenou pamětí

Metody s proměnnou metrikou s omezenou pamětí používají pouze omezený počet aktualizací.

Definice 34 *Nechť $\overline{m} > 0$ a $m = \min(\overline{m}, i-1)$. Řekneme, že základní optimalizační metoda je $\overline{m}$-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 $H_{i-m}^i = H$ pomocí m aktualizací

$$H_{j+1}^i = \gamma_j^i \left(H_j^i + U_j^i M_j^i (U_j^i)^T \right),$$

$i - m \leq j \leq i-1$, kde matice $U_j^i = [d_j, H_j^i y_j]$ a M_j^i jsou voleny tak, aby byly splněny kvazinewtonovské podmínky $H_{j+1}^i y_j = \rho_j^i s_j$, $i - m \leq j \leq i-1$.

Poznámka 128 Škálovací parametry se obvykle vybírají tak, že $\gamma_{i-m}^i = b_{i-1}/a_{i-1}$ a $\gamma_j^i = 1$ pro $i - m < j \leq i - 1$.

Tvrzení 3 *Nechť x_i, $i \in N$, je posloupnost generovaná $\overline{m}$-krokovou metodou s proměnnou metrikou s omezenou pamětí s přesným výběrem délky kroku (platí $s_i^T g_{i+1} = 0 \ \forall i \in N$) aplikovaná na ryze konvexní kvadratickou funkci (Q). Pak:*
(a) Směrové vektory s_i, $i \in N$, jsou rovnoběžné se směrovými vektory generovanými předpodmíněnou metodou sdružených gradientů. Platí

$$s_i = \left(\prod_{k=i-m}^{i-1} \gamma_k^i \right) \left(H g_i - \frac{y_{i-1}^T H g_i}{y_{i-1}^T d_{i-1}} d_{i-1} \right)$$

(předpokládáme že $H_{i-m}^i = H \ \forall i \in N$.
(b) Je splněno m kvazinewtonovských podmínek. Pro $i - m \leq j \leq i-1$ platí

$$H_i^i y_j = \left(\prod_{k=i-m}^{j} \gamma_k^i \right) \frac{\rho_j^i}{\gamma_j^i} d_j.$$

Věta 75 *(Kvadratické ukončení). Nechť jsou splněny předpoklady tvrzení 3. Pak existuje index $k \leq n$ tak, že $g_{k+1} = 0$ a $x_{k+1} = x^*$.*

Důkaz Podle tvrzení 3 jsou směrové vektory generované m-krokovou metodou s proměnnou metrikou s omezenou pamětí rovnoběžné s vektory generovanými metodou sdružených gradientů. Podle věty 20 tedy existuje index $k \leq n$ tak, že $g_{k+1} = 0$ a $x_{k+1} = x^*$ (při přesném výběru délky kroku nezáleží na normě směrového vektoru).

Je zřejmé, že u metod s omezenou pamětí není možné uchovávat matice H^i_{j+1}, $i-m \leq j \leq i-1$, neboť ty jsou obecně husté. Proto je nutné pracovat s jejich vektorovými reprezentacemi nebo s reprezentacemi používajími matice menších rozměrů. Jelikož obvykle $\overline{m} \leq 10$, jsou takové reprezentace velmi výhodné.

Nejprve se budeme zabývat vektorovou reprezentací metody BFGS. Tato reprezentace je založená na pseudosoučinovém tvaru (poznámka 61), který má pro metodu BFGS s $\eta = 1$ tvar

$$H_+ = \gamma V^T H V + \frac{\rho}{b} d d^T, \qquad V = I - \frac{1}{b} y d^T,$$

kde $y = g_+ - g$, $d = x_+ - x$ a $a = y^T H y$, $b = y^T d$. Abychom se vyhnuli dvojímu indexování, budeme bez újmy na obecnosti předpokládat, že $i \leq m$.

Věta 76 *Nechť H_{i+1} je matice získaná v i-tém kroku metody BFGS. Pak platí*

$$H_{i+1} = \left(\prod_{k=1}^{i} \gamma_k V_k\right)^T H_1 \left(\prod_{k=1}^{i} V_k\right) + \sum_{l=1}^{i} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{i} \gamma_k V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{i} V_k\right)$$

Důkaz (Indukcí) Pro $i = 1$ to bezprostředně plyne z pseudosoučinového tvaru pro metodu BFGS. Indukční krok:

$$\begin{aligned} H_{i+1} &= \gamma_i V_i^T H_i V_i + \frac{\rho_i}{b_i} d_i d_i^T = \gamma_i V_i^T \left(\prod_{k=1}^{i-1} \gamma_k V_k\right)^T H_1 \left(\prod_{k=1}^{i-1} V_k\right) V_i + \\ &\quad + \sum_{l=1}^{i-1} \frac{\rho_l}{b_l} \gamma_i V_i^T \left(\prod_{k=l+1}^{i-1} \gamma_k V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{i-1} V_k\right) V_i + \frac{\rho_i}{b_i} d_i d_i^T \\ &= \left(\prod_{k=1}^{i} \gamma_k V_k\right)^T H_1 \left(\prod_{k=1}^{i} V_k\right) + \sum_{l=1}^{i} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{i} \gamma_k V_k\right)^T d_l d_l^T \left(\prod_{k=l+1}^{i} V_k\right) \end{aligned}$$

Tvrzení věty 76 ukazuje, že matici H_i můžeme určit z matice H_1 (která je řídká) pomocí vektorů d_j, y_j, $1 \leq j \leq i-1$. Matici H_i nemusíme konstruovat explicitně, stačí počítat vektor $s_i = -H_i g_i$, což se provádí pomocí dvou rekurentních vztahů (Strangova formule). Nejprve se počítají zpětnou rekurzí vektory $u_i = -g_i$ a

$$u_j = -\left(\prod_{k=j}^{i-1} V_k\right) g_i, \quad i-1 \geq j \geq 1$$

Protože

$$u_j = V_j u_{j+1} = \left(I - \frac{1}{b_j} y_j d_j^T\right) u_{j+1} = u_{j+1} - \frac{d_j^T u_{j+1}}{b_j} y_j$$

můžeme psát

$$u_i = -g_i$$

a

$$\sigma_j = d_j^T u_{j+1}/b_j$$

$$u_j = u_{j+1} - \sigma_j y_j \tag{R1}$$

pro $i-1 \geq j \geq 1$. Potom počítáme přímou rekurzí vektory $v_1 = \gamma_1 H_1 u_1$ a

$$v_{j+1} = \left(\prod_{k=1}^{j} \gamma_k V_k\right)^T H_1 u_1 + \sum_{l=1}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} \gamma_k V_k\right)^T d_l d_l^T u_{l+1}, \quad 1 \leq j \leq i-1$$

Protože

$$v_{j+1} = \gamma_j V_j^T v_j + \frac{\rho_j}{b_j} d_j d_j^T u_{j+1} = \gamma_j \left(I - \frac{1}{b_j} d_j y_j^T\right) v_j + \rho_j \sigma_j d_j = \gamma_j v_j + (\rho_j \sigma_j - y_j^T(\gamma_j v_j)/b_j) d_j$$

můžeme psát

$$v_1 = \gamma_1 H_1 u_1$$

a

$$v_{j+1} = \gamma_j v_j + (\rho_j \sigma_j - y_j^T(\gamma_j v_j)/b_j) d_j \tag{R2}$$

pro $i-m \leq j \leq i-1$. Nakonec položíme $s_i = v_i$.

Poznámka 129 V rekurentních vztazích je třeba uchovávat čísla σ_j, $1 \leq j \leq i-1$. Vektory u_j, v_j, $1 \leq j \leq i-1$ mohou být uloženy v paměti počítače na stejném místě jako vektor $s_i = -H_i g_i$. Pro $m = \overline{m}$, což je maximální možná hodnota, potřebujeme uchovávat $2\overline{m}+3$ vektorů (d_j, y_j, $1 \leq j \leq \overline{m}$, a 3 vektory pro základní optimalizační metodu) a použijeme $O(mn)$ numerických operací.

Strangova formule (R1) a (R2) je nejstarší a nejjednodušší realizací metody BFGS s omezenou pamětí. Pro některé aplikace jsou výhodnější maticové reprezentace, které nyní odvodíme.

Lemma 19 *Nechť $N = -M^{-1}$, kde M je matice vystupující ve větě 33 s $\gamma = 1$. Pak platí*

$$N = \begin{bmatrix} \dfrac{(\eta-1)b^2}{\eta a + (1-\eta)\rho b}, & \dfrac{\eta a b}{\eta a + (1-\eta)\rho b} \\ \dfrac{\eta a b}{\eta a + (1-\eta)\rho b}, & a + \dfrac{\eta \rho a b}{\eta a + (1-\eta)\rho b} \end{bmatrix} \tag{N}.$$

Důkaz Z vyjádření matice M (věta 33) plyne

$$N = -M^{-1} = -\frac{1}{\det M} \begin{bmatrix} \dfrac{\eta-1}{a}, & \dfrac{\eta}{b} \\ \dfrac{\eta}{b}, & \dfrac{1}{b}(\eta \dfrac{a}{b} + \rho) \end{bmatrix}.$$

Dosadíme-li za $-\det M$ vztah μ definovaný v poznámce 59 (s $\gamma = 1$), dostaneme po úpravě tvrzení lemmatu.

Poznámka 130 Pro metodu DFP je $\eta = 0$, takže

$$N = \begin{bmatrix} -\frac{1}{\rho} d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}. \tag{ND}$$

Pro metodu BFGS je $\eta = 1$, takže

$$N = \begin{bmatrix} 0, & d^T y \\ d^T y, & \rho\, d^T y + y^T H y \end{bmatrix}. \tag{NB}$$

Lemma 20 *Nechť B a $\beta - b^T B^{-1} b$ jsou čtvercové regulární matice. Pak platí*

$$[A, a] \begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} [A, a]^T = AB^{-1}A^T + (a - AB^{-1}b)(\beta - b^T B^{-1} b)^{-1}(a - AB^{-1}b)^T.$$

Důkaz Vynásobením se snadno přesvědčíme, že platí

$$\begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} = \begin{bmatrix} B^{-1} + B^{-1}b(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & -B^{-1}b(\beta - b^T B^{-1} b)^{-1} \\ -(\beta - b^T B^{-1} b)^{-1} b^T B^{-1}, & (\beta - b^T B^{-1} b)^{-1} \end{bmatrix}.$$

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 $1 \le k \le m$ platí

$$H_{k+1} = H_k - [d_k,\ H_k y_k]\, N_k^{-1}\, [d_k,\ H_k y_k]^T, \tag{H}$$

kde N_k je matice specifikující konkrétní metodu s proměnnou metrikou. Budeme se snažit nalézt vyjádření

$$H_{k+1} = H_1 - [D_k,\ H_1 Y_k]\, \overline{N}_k^{-1}\, [D_k,\ H_1 Y_k]^T, \tag{$\overline{\text{H}}$}$$

kde $D_k = [d_1, \ldots, d_k]$, $Y_k = [y_1, \ldots, y_k]$ a kde $\overline{N}_k$ je symetrická matice řádu $2k$. Budeme přitom používat označení R_k pro horní trojúhelníkovou matici řádu k takovou, že $(R_k)_{ij} = d_i^T y_j$, $i \le j$, a $(R_k)_{ij} = 0$, $i > j$, C_k pro diagonální matici řádu k takovou, že $(C_k)_{ii} = d_i^T y_i$ a P_k pro diagonální matici řádu k takovou, že $(P_k)_{ii} = \rho_i$. Abychom zjednodušili zápis budeme v důkazech často indexy $k-1$ a k vynechávat a index $k+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $D = [d_1, \ldots, d_{k-1}]$, $Y = [y_1, \ldots, y_{k-1}]$ a $R = R_{k-1}$, $C = C_{k-1}$ a $P = P_{k-1}$, takže $D_k = [D, d]$, $Y_k = [Y, y]$ a

$$R_k = \begin{bmatrix} R, & D^T y \\ 0, & d^T y \end{bmatrix}, \qquad R_k - C_k = \begin{bmatrix} R - C, & D^T y \\ 0, & 0 \end{bmatrix}.$$

Poznamenejme, že pomocí (H) a ($\overline{\text{H}}$) můžeme indukční krok, používaný v důkazech, zapsat ve tvaru

$$\begin{aligned} H_+ &= H - \Big[d,\ H_1 y - [D,\ H_1 Y]\, \overline{N}^{-1}\, [D,\ H_1 Y]^T\, y\Big] \cdot \\ &\quad N^{-1} \Big[d,\ H_1 y - [D,\ H_1 Y]\, \overline{N}^{-1}\, [D,\ H_1 Y]^T\, y\Big]^T \\ &= H - \left([d,\ H_1 y] - [D,\ H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right) \cdot \\ &\quad N^{-1} \left([d,\ H_1 y] - [D,\ H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^T \end{aligned}$$

, (∗)

kde $\overline{N} = \overline{N}_{k-1}$ a $N = N_k$.

Věta 77 *Nechť H_1 je SPD matice a nechť pro libovolný index $1 \le k \le m$ platí (H), kde matice N_k je určená vztahem (ND) (metoda DFP). Pak lze psát*

$$H_{k+1} = H_1 - [D_k,\ H_1Y_k] \begin{bmatrix} -P_k^{-1}C_k, & R_k - C_k \\ (R_k - C_k)^T, & Y_k^T H_1 Y_k \end{bmatrix}^{-1} [D_k,\ H_1Y_k]^T . \qquad (\overline{\text{HD}})$$

Důkaz Pro $k = 1$ je ($\overline{\text{HD}}$) ekvivalentní s (HD) (s (H) kde matice N je určena pomocí (ND)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že ($\overline{\text{HD}}$) platí pro všechny indexy menší než k. Pro index k můžeme ($\overline{\text{HD}}$) zapsat (po permutaci) ve tvaru

$$H_+ = H_1 - [D,\ H_1Y,\ d,\ H_1y] \begin{bmatrix} -P_k^{-1}C, & R - C, & 0, & D^Ty \\ (R-C)^T, & Y^TH_1Y, & 0, & Y^TH_1y \\ 0, & 0, & -d^Ty/\rho, & 0 \\ y^TD, & y^TH_1Y, & 0, & y^TH_1y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^TH_1 \\ d^T \\ y^TH_1 \end{bmatrix}.$$

Použijeme-li lemma 20 a označíme-li

$$\overline{N} = \begin{bmatrix} -P_k^{-1}C, & R - C \\ (R-C)^T, & Y^TH_1Y \end{bmatrix},$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D,\ H_1Y]\,\overline{N}^{-1}\,[D,\ H_1Y]^T - \\ &\quad \left([d,\ H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \left(\begin{bmatrix} -d^Ty/\rho, & 0 \\ 0, & y^TH_1y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^TD, & y^TH_1Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^{-1} \cdot \\ &\quad \left([d,\ H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T \\ &= H - \left([d, H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \begin{bmatrix} -d^Ty/\rho, & 0 \\ 0, & y^THy \end{bmatrix}^{-1} \cdot \left([d, H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T, \end{aligned}$$

což je právě vztah (∗) s maticí N určenou pomocí (ND) (poznámka 130)

Věta 78 *Nechť H_1 je SPD matice a nechť pro libovolný index $1 \le k \le m$ platí (H), kde matice N_k je určená vztahem (NB) (metoda BFGS). Pak lze psát*

$$H_{k+1} = H_1 - [D_k,\ H_1Y_k] \begin{bmatrix} 0, & R_k \\ R_k^T, & P_kC_k + Y_k^TH_1Y_k \end{bmatrix}^{-1} [D_k,\ H_1Y_k]^T \qquad (\overline{\text{HB}}).$$

Důkaz Pro $k = 1$ je ($\overline{\text{HB}}$) ekvivalentní s (HB) (s (H) kde matice N je určena pomocí (NB)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že ($\overline{\text{HB}}$) platí pro všechny indexy menší než k. Pro index k můžeme ($\overline{\text{HB}}$) zapsat (po permutaci) ve tvaru

$$H_+ = H_1 - [D,\ H_1Y,\ d,\ H_1y] \begin{bmatrix} 0, & R, & 0, & D^Ty \\ R^T, & PC + Y^TH_1Y, & 0, & Y^TH_1y \\ 0, & 0, & 0, & d^Ty \\ y^TD, & y^TH_1Y, & d^Ty, & \rho\, d^Ty + y^TH_1y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^TH_1 \\ d^T \\ y^TH_1 \end{bmatrix}.$$

Použijeme-li lemma 20 a označíme-li

$$\overline{N} = \begin{bmatrix} 0, & R \\ R^T, & PC + Y^TH_1Y \end{bmatrix},$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D,\ H_1Y]\,\overline{N}^{-1}\,[D,\ H_1Y]^T - \\ &\quad \left([d,\ H_1y] - [D, H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \left(\begin{bmatrix} 0, & d^Ty \\ d^Ty, & \rho\, d^Ty + y^TH_1y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^TD, & y^TH_1Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^{-1} \cdot \\ &\quad \left([d,\ H_1y] - [D, H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T \\ &= H - \left([d, H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right) \cdot \\ &\quad \begin{bmatrix} 0, & d^Ty \\ d^Ty, & \rho d^Ty + y^THy \end{bmatrix}^{-1} \left([d, H_1y] - [D,\ H_1Y]\,\overline{N}^{-1} \begin{bmatrix} 0, & D^Ty \\ 0, & Y^TH_1y \end{bmatrix}\right)^T, \end{aligned}$$

což je právě vztah $(*)$ s maticí N určenou pomocí (NB) (poznámka 130).

Věta 79 *Nechť H_1 je SPD matice a nechť pro libovolný index $1 \le k \le m$ platí*

$$H_{k+1} = H_k + (d_k - H_ky_k)(d_k^Ty_k - y_k^TH_ky_k)^{-1}(d_k - H_ky_k)^T \tag{HR}$$

(metoda hodnosti 1). Pak lze psát

$$H_{k+1} = H_1 + (D_k - H_1Y_k)(R_k + R_k^T - C_k - Y_k^TH_1Y_k)^{-1}(D_k - H_1Y_k)^T. \tag{$\overline{\text{HR}}$}$$

Důkaz Vztah (HR) je pro $k = 1$ ekvivalentní se vztahem ($\overline{\text{HR}}$). Dále budeme postupovat matematickou indukcí. Předpokládejme, že ($\overline{\text{HR}}$) platí pro všechny indexy menší než k. Pro index k můžeme ($\overline{\text{HR}}$) 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 20 a označíme-li

$$\overline{N} = R + R^T - C - Y^TH_1Y$$

dostaneme

$$
\begin{aligned}
H_+ &= H_1 + (D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T + \\
&\quad \Big((D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T\, y - (d - H_1 y)\Big)\cdot \\
&\quad \Big(d^T y - y^T H_1 y - y^T\,(D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T\, y\Big)^{-1}\cdot \\
&\quad \Big((D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T\, y - (d - H_1 y)\Big)^T \\
&= H + (d - Hy)\left(d^T y - y^T H y\right)^{-1}(d - Hy)^T
\end{aligned}
$$

což je právě vztah (HR).

Poznámka 131 Podobná kompaktní schemata můžeme odvodit pro matici $B = H^{-1}$. Lze k tomu použít dualitu (poznámka 64). Jelikož přitom dojde k výměně $D_k \to Y_k$, $Y_k \to D_k$, je třeba horní polovinu matice $D_k^T Y_k$ nahradit horní polovinou matice $Y_k^T D_k$ neboli transponovanou dolní polovinou matice $D_k^T Y_k$. Proto místo horní trojúhelníkové matice R_k použijeme dolní trojúhelníkovou matici L_k takovou, že $(L_k)_{ij} = 0$, $i < j$, a $(L_k)_{ij} = d_i^T y_j$, $i \geq j$. Pro metodu DFP dostaneme

$$
B_{k+1} = B_1 - [Y_k, B_1 D_k] \begin{bmatrix} 0, & L_k^T \\ L_k, & C_k + D_k^T B_1 D_k \end{bmatrix}^{-1} [Y_k, B_1 D_k]^T \qquad (\overline{\mathrm{BD}})
$$

Pro metodu BFGS dostaneme

$$
B_{k+1} = B_1 - [Y_k, B_1 D_k] \begin{bmatrix} -C_k, & (L_k - C_k)^T \\ L_k - C_k, & D_k^T B_1 D_k \end{bmatrix}^{-1} [Y_k, B_1 D_k]^T \qquad (\overline{\mathrm{BB}})
$$

Pro metodu hodnosti 1 dostaneme

$$
B_{k+1} = B_1 + (Y_k - B_1 D_k)\left(L_k + L_k^T - C_k - D_k^T B_1 D_k\right)^{-1}(Y_k - B_1 D_k)^T \qquad (\overline{\mathrm{BR}})
$$

Nyní ukážeme, jak lze kompaktní schémata použít v souvislosti s metodami s proměnnou metrikou s omezenou pamětí. Omezíme se přitom na metodu BFGS, která je z popsaných metod obecně nejefektivnější. Matici $(\overline{\mathrm{HB}})$ lze po dosazení $H_1 = \gamma_k I$, kde $\gamma_k = d_k^T y_k / y_k^T y_k$, zapsat ve tvaru

$$
H_{k+1} = \gamma_k I + [D_k, \gamma_k Y_k] \begin{bmatrix} (R_k^{-1})^T (C_k + \gamma_k Y_k^T Y_k) R_k^{-1}, & -(R_k^{-1})^T \\ -R_k^{-1}, & 0 \end{bmatrix} [D_k, \gamma_k Y_k]^T
$$

Lze se o tom přesvědčit explicitním invertováním tak, jako v důkazu lematu 20. Nyní pokládáme $D_k = [d_{k-m}, \ldots, d_{k-1}]$, $Y_k = [y_{k-m}, \ldots, y_{k-1}]$. Matice C_k obsahuje diagonálu matice $D_k^T Y_k$ a matice R_k obsahuje horní polovinu matice $D_k^T Y_k$. Matice D_{k+1}, Y_{k+1} se získají z matic D_k, Y_k jednoduše ubráním prvního a přidáním posledního sloupce. Podobně jednoduše se získají matice $D_{k+1}^T Y_{k+1}$, $Y_{k+1}^T Y_{k+1}$ z matic $D_k^T Y_k$, $Y_k^T Y_k$ a tudíž i matice C_{k+1}, R_{k+1} z matic C_k, R_k. Tím máme k dispozici všechny matice potřebné k výpočtu matice H_{k+1}.

Matici $(\overline{\mathrm{BB}})$ můžeme po dosazení $B_1 = (1/\gamma_k) I$ zapsat ve tvaru

$$
B_{k+1} = \frac{1}{\gamma_k} I - \left[Y_k, \frac{1}{\gamma_k} D_k\right] \begin{bmatrix} -C_k^{-\frac{1}{2}}, & (\overline{L}_k^{-1}(L_k - C_k))^T \\ 0, & (\overline{L}_k^{-1})^T \end{bmatrix} \cdot \begin{bmatrix} C_k^{-\frac{1}{2}}, & 0 \\ -\overline{L}_k^{-1}(L_k - C_k), & \overline{L}_k^{-1} \end{bmatrix} \left[Y_k, \frac{1}{\gamma_k} D_k\right]^T
$$

kde

$$
\overline{L}_k \overline{L}_k^T = (L_k - C_k)^T C_k^{-1} (L_k - C_k) + \frac{1}{\gamma_k} D_k^T D_k
$$

Lze se o tom přesvědčit explicitním invertováním a násobením. Je vidět, že potřebujeme rozkládat pouze matici řádu k (k získání matice $\overline{L}_k\overline{L}_k^T$). Zatímco vzorec ($\overline{\mathrm{HB}}$) nepřináší příliš mnoho výhod ve srovnání se Strangovou formulí, je vzorec ($\overline{\mathrm{BB}}$) velmi užitečný, neboť ho lze použít tam, kde je nutné pracovat s maticí B.

7.2 Diferenční verze nepřesné Newtonovy metody

Diferenční verze nepřesné Newtonovy metody jsou v podstatě nepřesné metody s lokálně omezeným krokem (algoritmus 6), kde se nepoužívá matice $B = G$ a násobení $q = Bp = Gp$ se nahražuje numerickým derivováním

$$G(x)p \approx \frac{g(x+\delta p) - g(x)}{\delta}$$

kde δ je malá diference ($\delta = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost). Jinak se algoritmus 6 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.

7.3 Diferenční verze Newtonovy metody pro řídké úlohy

Diferenční verze Newtonovy metody jsou založeny na aproximaci sloupců Ge_i, $1 \le i \le n$, Hessovy matice G pomocí diferenčních vzorců

$$G(x)e_i \approx \frac{g(x+\delta e_i) - g(x)}{\delta}, \qquad 1 \le i \le n$$

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:

$$G = \begin{bmatrix} G_{11}, & G_{12}, & 0, & 0, & 0 \\ G_{21}, & G_{22}, & G_{23}, & 0, & 0 \\ 0, & G_{32}, & G_{33}, & G_{34}, & 0 \\ 0, & 0, & G_{43}, & G_{44}, & G_{45} \\ 0, & 0, & 0, & G_{54}, & G_{55} \end{bmatrix} \tag{G1}$$

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

$$G = \begin{bmatrix} G_{11}, & G_{12}, & G_{13}, & G_{14}, & G_{15} \\ G_{21}, & G_{22}, & 0, & 0, & 0 \\ G_{31}, & 0, & G_{33}, & 0, & 0 \\ G_{41}, & 0, & 0, & G_{44}, & 0 \\ G_{51}, & 0, & 0, & 0, & G_{55} \end{bmatrix}. \tag{G2}$$

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 $\mathcal{S}_i$, $1 \le i \le k$. Při určování těchto skupin však nebudeme pracovat s celou maticí G, ale pouze s jejími submaticemi, které dostaneme vyškrtnutím známých řádků a sloupců. Nechť G_i je submatice matice G, kterou dostaneme, vyškrtneme-li v matici G řádky a sloupce s indexy $j \in S_1 \cup \ldots \cup S_{j-1}$, 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

$$\frac{g(x+\delta v_i)-g(x)}{\delta} \approx Gv_i, \qquad 1 \le i \le k,$$

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 $G_i(\mathcal{S}_i) = G_i \cap G(\mathcal{S}_i)$, ostatní prvky submatice $G(\mathcal{S}_i)$ jsou určeny symetrií matice G).

Zatím jsme se nezabývali určováním skupin $\mathcal{S}_i$, $1 \le i \le k$. 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 $\mathcal{S}_i$, $1 \le i \le k$, 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 difference 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 (G1) a položme

$$\begin{aligned} v_1 &= [1,0,1,0,1]^T, \\ v_2 &= [0,1,0,1,0]^T. \end{aligned}$$

Pak platí

$$\frac{g(x+\delta v_1) - g(x)}{\delta} \approx G v_1 = \begin{bmatrix} G_{11} \\ G_{21} + G_{23} \\ G_{33} \\ G_{43} + G_{45} \\ G_{55} \end{bmatrix}$$

$$\frac{g(x+\delta v_2) - g(x)}{\delta} \approx G v_2 = \begin{bmatrix} G_{12} \\ G_{22} \\ G_{32} + G_{34} \\ G_{44} \\ G_{54} \end{bmatrix}$$

Z těchto rovnic určíme přímo hodnoty G_{11}, G_{33}, G_{55}, G_{12}, G_{22}, G_{44}, G_{54} a protože matice G je symetrická i hodnoty G_{21}, G_{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 difference gradientů (přímá metoda používá tři difference 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

$$\frac{g(x+\delta v_i) - g(x)}{\delta} \approx G v_i, \qquad 1 \le i \le k$$

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 $G_U(\mathcal{S}_i)$, 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 $\lambda \geq 0$). Používá se přitom řídký Choleského rozklad

$$R^T R = P(G + \lambda I)P^T$$

kde R je regulární horní trojúhelníková matice a P je permutační matice, jejíž jediným účelem je přerovnat řádky a sloupce matice $G + \lambda I$ tak, aby počet nově vzniklých nenulových prvků byl co nejmenší. Nalezení permutační matice P a následné určení struktury horní trojúhelníkové matice R se nazývá symbolickou faktorizací. Symbolická faktorizace se provádí pouze jednou (na začátku iteračního procesu) a proto je možné používat časově náročnější složité postupy, které minimalizují počet nově vzniklých nenulových prvků. Tyto postupy mají kombinatorický charakter a jejich popis se vymyká rozsahu tohoto textu (jsou jim věnovány samostatné monografie). Výpočet prvků horní trojúhelníkové matice R (numerická faktorizace) se provádí podle vzorců uvedených v oddílu 5.8.

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 předpomiňování 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 35 *Řídkou reprezentací Hessovy matice G nazveme trojici vektorů num $(G_U) \in R^{m_U}$, ind $(G_U) \in R^{m_U}$, ord $(G_U) \in R^{n+1}$, 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 ord (G_U) obsahuje ukazatele umístění diagonálních prvků matice G_U ve vektorech num (G_U) a ind (G_U). Poslední prvek vektoru ord (G_U) (i indexem $n + 1$) má hodnoty m_{U+1}.*

Pro matici (G1) platí

$$\begin{aligned} num(G_U) &= [G_{11}, G_{12}, G_{22}, G_{23}, G_{33}, G_{34}, G_{44}, G_{45}, G_{55}]^T \\ ind(G_U) &= [1, 2, 2, 3, 3, 4, 4, 5, 5]^T \\ ord(G_U) &= [1, 3, 5, 7, 9, 10]^T \end{aligned}$$

Pro matici (G2) platí

$$
\begin{aligned}
num(G_U) &= [G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{22}, G_{33}, G_{44}, G_{55}]^T \\
ind(G_U) &= [1, 2, 3, 4, 5, 2, 3, 4, 5]^T \\
ord(G_U) &= [1, 6, 7, 8, 9, 10]^T
\end{aligned}
$$

7.4 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

$$
\begin{aligned}
\mathcal{P}_Q B &= \arg \min_{B^+ \in \mathcal{V}_Q} \| B^+ - B \|_F \\
\mathcal{P}_S B &= \arg \min_{B^+ \in \mathcal{V}_S} \| B^+ - B \|_F \\
\mathcal{P}_G B &= \arg \min_{B^+ \in \mathcal{V}_G} \| B^+ - B \|_F
\end{aligned}
$$

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 dalším textu ukážeme, že i jednoduché aktualizace založené na skládání projekcí mohou vést k superlineárně konvergentním metodám.

Věta 80 *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 platí*

$$
\begin{aligned}
\mathcal{P}_Q B &= B + \frac{(y - Bd)d^T}{d^T d} \\
\mathcal{P}_S B &= \frac{1}{2}(B + B^T)
\end{aligned}
$$

a

$$
\begin{aligned}
(\mathcal{P}_G B)_{ij} &= B_{ij}\ ,\ G_{ij} \neq 0 \\
(\mathcal{P}_G B)_{ij} &= 0\quad\ ,\ G_{ij} = 0
\end{aligned}
$$

Důkaz Budeme postupovat podobně jako v důkazu věty 39. Vztah pro $\mathcal{P}_Q B$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2} \parallel B^+ - B \parallel_F^2 + u^T(w - (B^+ - B)d) = \\ &= \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(B_{ij}^+ - B_{ij})^2 + \sum_{i=1}^{n} u_i \left(y_i - \sum_{j=1}^{n} B_{ij}^+ d_j \right) \end{aligned}$$

Derivováním Lagrangeovy funkce podle prvků matice B^+ dostaneme

$$\frac{\partial L}{\partial B_{ij}^+} = (B_{ij}^+ - B_{ij}) - u_i d_j$$

$\forall 1 \leq i \leq n$, $\forall 1 \leq j \leq n$. Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $B^+ = B + ud^T$, což po dosazení do kvazinewtonovské podmínky $B^+ d = y$ dává $ud^T d = y - Bd$, neboli

$$u = \frac{y - Bd}{d^T d}$$

Dosadíme-li tento výraz do vzorce $B^+ = B + ud^T$, dostaneme vztah pro $\mathcal{P}_Q B$. Vztah pro $\mathcal{P}_S B$ odvodíme minimalizací funkce

$$\frac{1}{2} \parallel B^+ - B \parallel_F^2 = \frac{1}{2}\sum_{i=1}^{n}(B_{ii}^+ - B_{ii})^2 + \frac{1}{2}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}((B_{ij}^+ - B_{ij})^2 + (B_{ij}^+ - B_{ji})^2)$$

Derivujeme-li tuto funkci podle prvků matice B^+, a položíme-li derivace rovny nule dostaneme podmínky

$$\begin{aligned} B_{ij}^+ - B_{ij} &= 0 \ , \ i = j \\ (B_{ij}^+ - B_{ij}) + (B_{ij}^+ - B_{ji}) &= 0 \ , \ i \neq j \end{aligned}$$

což dává $B^+ = (B + B^T)/2$. Vztah pro $\mathcal{P}_G B$ odvodíme minimalizací funkce

$$\frac{1}{2} \parallel B^+ - B \parallel_F^2 = \frac{1}{2}\sum_{G_{ij} \neq 0}(B_{ij}^+ - B_{ij})^2 + \frac{1}{2}\sum_{G_{ij}=0} B_{ij}^2$$

neboť podle předpokladu $B_{ij}^+ = 0$ pokud $G_{ij} = 0$. Derivujeme-li tuto funkci podle prvků matice B^+ a položíme-li derivace rovny nule dostaneme

$$\begin{aligned} B_{ij}^+ - B_{ij} &= 0 \ , \ G_{ij} \neq 0 \\ B_{ij}^+ &= 0 \ , \ G_{ij} = 0 \end{aligned}$$

což jsme měli dokázat.

V dalším textu zavedeme vektory $d^i \in R^n$, $1 \leq i \leq n$ takové, že

$$\begin{aligned} d_j^i &= d_j \ , \ G_{ij} \neq 0 \\ d_j^i &= 0 \ , \ G_{ij} = 0 \end{aligned}$$

a místo standardní kvazinewtonovské podmínky $B^+ d = y$ použijeme řídkou kvazinewtonovskou podmínku

$$\sum_{j=1}^{n}(B_{ij}^{+}-(\mathcal{P}_G B)_{ij})d_j^i = w_i, \qquad 1 \le i \le n$$

kde $w = y - (\mathcal{P}_G B)d$.

Věta 81 *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 platí*

$$\begin{aligned}
\mathcal{P}_{QS}B &= B + \frac{(y-Bd)d^T + d(y-Bd)^T}{d^T d} - \frac{(y-Bd)^T d}{d^T d}\,\frac{dd^T}{d^T d} \\
\mathcal{P}_{QG}B &= \mathcal{P}_G(B + ud^T) \\
\mathcal{P}_{SG}B &= \mathcal{P}_S\mathcal{P}_G B = \mathcal{P}_G\mathcal{P}_S B
\end{aligned}$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = w$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \parallel d^i \parallel^2 e_i e_i^T$$

Důkaz Vztah pro $\mathcal{P}_{QS}B$ plyne bezprostředně z věty 39 (metoda PSB). Stačí dosadit $W = I$. Zřejmě platí $\mathcal{P}_{QG}B = \mathcal{P}_{QG}\mathcal{P}_G B$. Vztah pro $\mathcal{P}_{QG}B = \mathcal{P}_{QG}\mathcal{P}_G B$ odvodíme pomocí Lagrangeovy funkce

$$L = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(B_{ij}^{+}-(\mathcal{P}_G B)_{ij})^2 + \sum_{i=1}^{n} u_i \left(w_i - \sum_{j=1}^{n}(B_{ij}^{+}-(\mathcal{P}_G B)_{ij})d_j^i \right)$$

obsahující řídkou kvazinewtonovskou podmínku. Derivováním Lagrangeovy funkce podle prvků matice B^+ dostaneme

$$\frac{\partial L}{\partial B_{ij}^{+}} = (B_{ij}^{+}-(\mathcal{P}_G B)_{ij}) - u_i d_j^i$$

$\forall 1 \le i \le n$, $\forall 1 \le j \le n$. Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $B_{ij}^{+}-(\mathcal{P}_G B)_{ij} = u_i d_j^i$, $1 \le i \le n$, $1 \le j \le n$, což jsme měli dokázat. Dosadíme-li toto vyjádření do řídké kvazinewtonovské podmínky, dostaneme

$$\sum_{j=1}^{n} u_i d_j^i d_j^i = w_i, \qquad 1 \le i \le n$$

neboli $Qu = w$, 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 82 *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 platí*

$$\mathcal{P}_{QSG}B = \mathcal{P}_G(B + ud^T + du^T)$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = w$ se symetrickou pozitivně semidefinitní maticí

$$Q = \mathcal{P}_G(dd^T) + \sum_{i=1}^{n} \parallel d^i \parallel^2 e_i e_i^T$$

která má stejnou strukturu jako matice G.

Důkaz Zřejmě platí $\mathcal{P}_{QSG}B = \mathcal{P}_{QSG}\mathcal{P}_G B$. Jelikož matice $B^+ - \mathcal{P}_G B$ je symetrická, můžeme položit $B^+ - \mathcal{P}_G B = X + X^T$, kde X je zatím neznámá čtvercová matice. Použijeme Lagrangeovu funkci

$$L = \frac{1}{4}\sum_{i=1}^{n}\sum_{j=1}^{n}(X_{ij}+X_{ji})^2 + \sum_{i=1}^{n} u_i \left(w_i - \sum_{j=1}^{n}(X_{ij}+X_{ji})d_j^i \right)$$

obsahující řídkou kvazinewtonovskou podmínku. Derivováním Lagrangeovy funkce podle prvků matice X dostaneme

$$\frac{\partial L}{\partial X_{ij}} = (X_{ij}+X_{ji}) - u_i d_j^i - u_j d_i^j$$

$\forall 1 \le i \le n$, $\forall 1 \le j \le n$. Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $B^+ - (\mathcal{P}_G B)_{ij} = u_i d_j^i + u_j d_i^j$, což jsme měli dokázat. Dosadíme-li toto vyjádření do řídké kvazinewtonovské podmínky, dostaneme

$$w_i = \sum_{j=1}^{n}(u_i d_j^i + u_j d_i^j)d_j^i = \| d^i \|^2 u_i + \sum_{j=1}^{n} d_i^j d_j^i u_j$$

neboli $Qu = w$, kde Q je symetrická matice vystupující v tvrzení věty. Matice Q má zřejmě stejnou strukturu jako matice G. Nechť $v \in R^n$ je libovolný vektor. Pak platí

$$\begin{aligned} v^T Q v &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_i^j d_j^i v_i v_j + \sum_{i=1}^{n} \| d^i \|^2 v_i^2 = \sum_{G_{ij}\neq 0} d_i d_j v_i v_j + \sum_{G_{ij}\neq 0} d_j^2 v_i^2 = \\ &= \frac{1}{2}\sum_{G_{ij}\neq 0}(d_i v_j + d_j v_i)^2 \ge 0 \end{aligned} \qquad (*)$$

(matice G_{ij} je symetrická), takže matice Q je pozitivně semidefinitní. Zbývá dokázat, že rovnice $Qu = w$ 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. Pokud by matice Q nebyla pozitivně definitní, existoval by vektor $v \neq 0$ takový, že $v^T Q v = 0$. Pak by podle vyjádření $(*)$ musel existovat index $1 \le i \le n$ takový, že $v_i \neq 0$ a

$$d_i v_j + d_j v_i = 0 \qquad \forall G_{ij} \neq 0$$

Jelikož předpokládáme, že $G_{ii} \neq 0$ musí nutně platit $d_i v_i = 0$, neboli $d_i = 0$, což po dosazení do poslední rovnosti dává $d_j v_i = 0$ $\forall G_{ij} \neq 0$, neboli $d_j = 0$ $\forall G_{ij} \neq 0$. To je ale ve sporu s předpokladem, že

$$\| d^i \|^2 = \sum_{j=1}^{n}(d_j^i)^2 = \sum_{G_{ij}\neq 0} d_j^2 \neq 0$$

Předpokládejme nyní, že pro nějaký index $1 \le i \le n$ platí $\| d^i \| = 0$. Pak matice Q má nulový i-tý řádek a i-tý sloupec a platí

$$w_i = y_i - \sum_{G_{ij}\neq 0} B_{ij} d_j = \sum_{G_{ij}\neq 0} (\tilde{G}_{ij} - B_{ij}) d_j^i = 0$$

(Matice $\tilde{G}$ je definovaná vztahem (a) v důkazu lemmatu 13). 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

$$B^+ = \mathcal{P}_{QSG} B \tag{BT}$$

se nazývá Tointovou metodou. Její realizace je poměrně pracná, neboť je třeba řešit dodatečnou soustavu rovnic $Qu = w$ (Tointův systém). V hustém případě 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 tomto textu se budeme zabývat Marwilovou metodou s aktualizací

$$B^+ = \mathcal{P}_S \mathcal{P}_{QG} B \tag{BM}$$

Powellovou metodou s aktualizací

$$B^+ = \mathcal{P}_G \mathcal{P}_{QS} B \tag{BP}$$

a Steihaugovou metodou s aktualizací

$$B^+ = \mathcal{P}_{SG} \mathcal{P}_Q B \tag{BS}$$

Lemma 21 *Nechť B^+ je matice určená pomocí některé z aktualizací (BT), (BM), (BP), (BS). Pak platí*

$$B^+ \in \mathcal{V}_S \cap \mathcal{V}_G$$

Důkaz Pro aktualizaci (BT) a (BS) je toto tvrzení zřejmé. V případě aktualizace (BM) tvrzení plyne z toho, že projekce $\mathcal{P}_S$, určená symetrií matice, neovlivní symetrickou řídkou strukturu. V případě aktualizace (BP) tvrzení plyne z toho, že projekce $\mathcal{P}_G$, určená symetrickou řídkou strukturou, neovlivní symetrii matice.

Ve vzorcích (BT), (BM), (BP), (BS) 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 22 *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 83 *Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s lokálně omezeným krokem (definice 28) s $\overline{\gamma} < \infty$. 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 (BT), (BM), (BP), (BS) a množiny N_1, N_2, N_3 jsou definovány v poznámce 95). Pak jestliže funkce $F : R^n \to R$ splňuje podmínky (F1), (F2), (F3) a (F5), platí*

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

Důkaz (a) nejprve ukážeme, že matice B_i, $i \in N$, jsou dostatečně omezené, neboli že platí $\| B_i \| \leq C_i$, $i \in N$, kde C_i, $i \in N$, jsou čísla splňující rekurentní nerovnosti

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

Nechť $i \in N_2$ a nechť $\tilde{G}_i$ je matice definovaná vztahem (a) v důkazu lemmatu 13. Pak platí

$$\begin{aligned} \| \tilde{G}_i - G_i \|_F &= \| \int_0^1 (G(x_i + \lambda d_i) - G(x_i))d\lambda \|_F \leq \sqrt{n} \int_0^1 \| G(x_i + \lambda d_i) - G(x_i) \| d\lambda \leq \\ &\leq \overline{L}\sqrt{n} \| d_i \| \int_0^1 \lambda d\lambda = \frac{1}{2}\overline{L}\sqrt{n} \| d_i \| \end{aligned}$$

(používáme předpoklad (F5) a skutečnost, že Frobeniova norma není větší než $\sqrt{n}$ násobek spektrální normy). Podobným způsobem dostaneme

$$\| \tilde{G}_i - G_{i+1} \|_F \leq \frac{1}{2}\overline{L}\sqrt{n} \| d_i \|$$

Použijeme-li nerovnost $\| B_{i+1} - \tilde{G}_i \|_F \leq \| B_i - \tilde{G}_i \|_F$, která plyne z lemmatu 22, můžeme psát

$$\begin{aligned} \| B_{i+1} - G_{i+1} \|_F &\leq \| B_{i+1} - \tilde{G}_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \leq \| B_i - \tilde{G}_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \leq \\ &\leq \| B_i - G_i \|_F + \| \tilde{G}_i - G_i \|_F + \| \tilde{G}_i - G_{i+1} \|_F \leq \\ &\leq \| B_i - G_i \|_F + \overline{L}\sqrt{n} \| d_i \| \end{aligned}$$

Tato nerovnost je splněna i pro $i \notin N_2$, neboť pro $i \notin N_2$ platí $B_{i+1} = B_i$, $G_{i+1} = G_i$ a $d_i = 0$, a použijeme-li ji několikrát po sobě, dostaneme

$$\| B_{i+1} - G_{i+1} \|_F \leq \| B_1 - G_1 \|_F + \overline{L}\sqrt{n} \sum_{j=1}^{i} \| d_j \|$$

Použijeme-li předpoklad (F3) a položíme-li $C_1 = 2\overline{G}\sqrt{n} + \| B_1 \|_F$ a $\overline{C} = \overline{L}\sqrt{n}$, dostaneme nerovnosti $\| B_i \| \leq C_i$, $i \in N$, kde čísla C_i, $i \in N$, splňují nerovnosti (C).
(b) Označíme-li

$$M_i = \max_{1\leq j\leq i} \| B_j \|$$

můžeme psát $M_i \leq C_i$, $i \in N$, a podle poznámky 102 dostaneme

$$\sum_{i=1}^{\infty} \frac{1}{M_i} = \infty$$

a můžeme použít větu 50.

Věta 84 *Nechť jsou splněny předpoklady věty 83 a nechť $x_i \to x^*$ a $\| \omega_i(s_i) \| \to 0$. Pak jestliže funkce $F : R^n \to R$ splňuje podmínky (F3), (F4), (F5), $x_i \to x^*$ Q-superlineárně.*

Důkaz Nechť $i \in N_2$. Použijeme-li lemma 22 a první dvě nerovnosti z důkazu věty 83, můžeme psát

$$\begin{aligned} \| B_{i+1} - G_{i+1} \|_F^2 &\leq \left(\| B_{i+1} - \tilde{G}_i \|_F + \| G_{i+1} - \tilde{G}_i \|_F \right)^2 \leq \\ &\leq \| B_{i+1} - \tilde{G}_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_{i+1} - \tilde{G}_i \|_F \| d_i \| \leq \\ &\leq \| B_i - \tilde{G}_i \|_F^2 - \frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}n\left(\overline{B} + \overline{G}\right) \right) \| d_i \| \end{aligned}$$

a

$$\begin{aligned} \| B_i - \tilde{G}_i \|_F^2 &\leq \left(\| B_i - G_i \|_F + \| G_i - \tilde{G}_i \|_F \right)^2 \leq \\ &\leq \| B_i - G_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_i - G_i \|_F \| d_i \| \leq \\ &\leq \| B_i - G_i \|_F^2 + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}n\left(\overline{B} + \overline{G}\right) \right) \| d_i \| \end{aligned}$$

(existence konstanty $\overline{\Delta}$ plyne z (T3), existence konstanty $\overline{B}$ plyne z věty 52 a existence konstanty $\overline{G}$ plyne z (F3)). 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 $\overline{M} = \frac{1}{2}\overline{L}^2 n\overline{\Delta} + 2\overline{L}n(\overline{B} + \overline{G})$. Tato nerovnost platí formálně i pro $i \notin N_2$ (pro $i \notin N_2$, kdy $d_i = 0$ a $y_i = 0$, můžeme výraz na levé straně nahradit nulou). Použijeme-li tuto nerovnost a větu 52, dostaneme

$$\begin{aligned} \sum_{i=1}^{\infty} \frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} &\leq \sum_{i=1}^{\infty} \left(\| B_i - G_i \|_F^2 - \| B_{i+1} - G_{i+1} \|_F^2 \right) + \overline{M} \sum_{i=1}^{\infty} \| d_i \| \leq \\ &\leq \| B_1 - G_1 \|_F^2 + \overline{M} \sum_{i=1}^{\infty} \| d_i \| < \infty \end{aligned} \qquad (*)$$

Dále podle (F5) platí

$$\begin{aligned} \frac{\| (G^* - B_i) d_i \|}{\| d_i \|} &\leq \frac{\| G^* d_i - y_i \|}{\| d_i \|} + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\ &\leq \| G^* - G_i \| + \| \tilde{G}_i - G_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\ &\leq \overline{L} \| x^* - x_i \| + \frac{1}{2}\overline{L}\sqrt{n} \| d_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \end{aligned}$$

(viz důkaz věty 83), takže

$$\frac{\| (G^* - B_i) d_i \|}{\| d_i \|} \to 0$$

neboť $x_i \to x^*$ podle předpokladu, $\| d_i \| \to 0$ podle věty 52 a $\| y_i - B_i d_i \| \, / \, \| d_i \| \to 0$ podle $(*)$. Jelikož $\| \omega_i(s_i) \| \to 0$ jsou splněny předpoklady věty 54 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 30). 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.

Následující tabulka ukazuje srovnání několika metod pro řídké úlohy (CG - metoda sdružených gradientů, 5-BFGS - pětikroková metoda BFGS s omezenou pamětí, diferenční verze nepřesné Newtonovy metody, diferenční verze řídké Newtonovy metody s iteračním CG nebo s přesným GM řešením soustavy lineárních rovnic, řídká VM metoda (Marwilova projekce) s iteračním CG nebo s přesným GM řešením soustavy lineárních rovnic, hustá BFGS metoda) při minimalizaci 22 testovacích funkcí se 1000 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG, jakož i celkový čas výpočtu).

Metoda	NIT - NFV - NFG	čas
CG	2066 - 3903 - 3903	7.19
5-BFGS	1965 - 2149 - 2149	7.63
Nepřesná Newtonova (dif. verze) + CG	702 - 822 - 6188	7.25
Řídká Newtonova (dif. verze) + CG	518 - 567 - 2461	7.69
Řídká Newtonova (dif. verze) + GM	377 - 405 - 1722	7.03
Řídká VM + CG (Marwilova projekce)	1318 - 2009 - 2009	12.58
Řídká VM + GM (Marwilova projekce)	2440 - 4841 - 4841	36.41
Hustá BFGS	1498 - 1656 - 1656	23.73

7.5 Diferenční verze Newtonovy metody pro separovatelné úlohy

Rozsáhlé úlohy jsou často formulovány tak, že platí

$$F(x) = \sum_{k=1}^{m} f_k(x)$$

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, $1 \leq k \leq m$). 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 36 *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

$$ord(J)_k = 1 + \sum_{i=1}^{k-1} n_k, \qquad 1 \leq k \leq m+1$$

V dalším výkladu budeme používat redukované gradienty $\hat{g}_k(x) \in R^{n_k}$, které obsahují pouze nenulové prvky gradientů $g_k(x) \in R^n$, $1 \leq k \leq m$, a redukované Hessovy matice $\hat{G}_k(x) \in R^{n_k \times n_k}$, které obsahují pouze nenulové prvky Hessových matic $G_k(x) \in R^{n \times n}$, $1 \leq k \leq m$.

Definice 37 *Nechť N_k, $1 \leq k \leq m$, jsou množiny indexů proměnných vystupujících ve funkcích $f_k(x)$, $1 \leq k \leq m$, a nechť $Z_k \in R^{n \times n_k}$ jsou matice, jejichž sloupce tvoří ortonormální báze v podprostorech určených proměnnými z N_k (jsou to sloupce jednotkové matice s indexy z N_k). Pak vektory $\hat{g}_k(x) = Z_k^T g_k(x)$, $1 \leq k \leq m$, nazveme redukovanými gradienty a matice $\hat{G}_k(x) = Z_k^T G_k(x) Z_k$, $1 \leq k \leq m$, nazveme redukovanými Hessovými maticemi funkcí $f_k(x)$, $1 \leq k \leq m$.*

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

$$\hat{G}_k(x)\hat{e}_j \approx \frac{\hat{g}_k(x + \delta\hat{e}_j) - \hat{g}_k(x)}{\delta}$$

kde $\hat{e}_j$, $1 \leq j \leq n_k$, 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 132 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 36) a numerické hodnoty redukovaných Hessových matic, můžeme snadno určit řídkou reprezentaci Hessovy matice (definice 35). 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 5.3) nebo realizaci ve formě nepřesné metody s lokálně omezeným krokem (oddíl 5.5).

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.

7.6 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 kde $\hat{d}_k \in R^{n_k}$ je vektor dimenze n_k, který obsahuje prvky vektoru d s indexy z N_k (vše pro $1 \leq k \leq m$). 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 20) a aktualizace se provádí podle vzorců

$$\begin{aligned}
\hat{B}_k^+ &= \hat{B}_k + \frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T, \quad \hat{y}_k^T\hat{d}_k > 0 \\
\hat{B}_k^+ &= \hat{B}_k \qquad\qquad\qquad\qquad\qquad\qquad\quad , \quad \hat{y}_k^T\hat{d}_k \leq 0
\end{aligned}$$

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 34.

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 132 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).

Následující tabulka ukazuje srovnání několika metod pro separovatelné úlohy (CG - metoda sdružených gradientů, 5-BFGS - pětikroková metoda BFGS s omezenou pamětí, diferenční verze nepřesné Newtonovy metody, diferenční verze separovatelné Newtonovy metody s iteračním CG nebo s přesným GM řešením soustavy lineárních rovnic, separovatelhá BFGS metoda s iteračním CG nebo s přesným GM řešením soustavy lineárních rovnic) při minimalizaci 10 testovacích funkcí se 100 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG, jakož i celkový čas výpočtu).

Metoda	NIT - NFV - NFG	čas
CG	2390 - 4501 - 4501	12.75
5-BFGS	2389 - 2586 - 2586	11.91
Nepřesná Newtonova (dif. verze) + CG	644 - 813 - 6280	11.09
Separovatelná Newtonova (dif. verze) + CG	556 - 607 - 2328	16.64
Separovatelná Newtonova (dif. verze) + GM	416 - 446 - 1635	14.28
Separovatelná BFGS + CG	841 - 995 - 995	12.96
Separovatelná BFGS + GM	753 - 882 - 882	10.54

7.7 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 6.2). 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

$$B_+ = J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F$$
$$B_+ = \mathcal{P}_S \mathcal{P}_{QS} B \quad , \quad F - F_+ \leq \underline{\vartheta} F$$

(viz (BM) v oddílu 7.4), kde $J_+ = J(x_+)$. Globální konvergence této metody plyne z věty 71 a věty 83. Superlineární konvergence této metody plyne z věty 71 a věty 84.

2) Kombinace GN metody s diferenční verzí Newtonovy metody. V prvním iteračním kroku pokládáme $B = J^T J$. Po skončení každého iteračního kroku pokládáme

$$B_+ = J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F$$
$$B_+ = J_+^T J_+ + \sum_{k=1}^{m} f_k^+ G_k^+ \quad , \quad F - F_+ \leq \underline{\vartheta} F$$

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 69 a věty 71.

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

$$B_k^+ = B_k + \frac{w_k w_k^T}{d_k^T w_k} \quad , \quad |d_k^T w_k| > \underline{\delta}$$
$$B_k^+ = B_k \quad , \quad |d_k^T w_k| \leq \underline{\delta}$$

pro $1 \leq k \leq m$, a

$$B_+ = J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F$$
$$B_+ = J_+^T J_+ + \sum_{k=1}^{m} f_k^+ B_k^+ \quad , \quad F - F_+ \leq \underline{\vartheta} F$$

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 100 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
GN	1584 -1819 - 1603	–	31.47
GN + VM (řídké)	1039 - 1136 - 1061	–	20.92
GN + VM (separovatelné)	980 - 1064 - 1002	–	28.24
GN + Newton	947 -1042 - 1247	–	25.92
Newton	1139 - 1269 - 3732	–	1:20.02
VM (řídké)	3411 - 5131 - 5131	1	1:07.17
VM (separovatelné)	3014 - 3793 - 3793	1	1:23.43
CG	8024 - 15670 - 15670	3	2:23.68
5 - BFGS	7106 - 7762 - 7762	3	1:47.00
Nepřesná Newtonova + CG	1486 - 16608 - 16383	–	2:34.34

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.

7.8 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 G hustá (stejnou strukturu má matice $J^T J$) 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 38 *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.

Definice 39 *Nechť $J \in R^{m\times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$\delta_1 u_1 = f, \qquad \gamma_1 q_1 = J^T u_1$$

a

$$\delta_{i+1} u_{i+1} = J q_i - \gamma_i u_i \tag{BL}$$

$$\gamma_{i+1} q_{i+1} = J^T u_{i+1} - \delta_{i+1} q_i$$

pro $1 \le i \le n$, kde koeficienty γ_i, δ_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 určeným maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$.

Poznámka 133 Nechť $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \le i \le k$ pro nějaký index $1 \le k \le n$. Pak podle (BL) platí $f = U_{k+1}(\delta_1 e_1)$ a

$$J Q_k = U_{k+1} B_k$$

$$J^T U_{k+1} = Q_k B_k^T + \gamma_{k+1} q_{k+1} e_{k+1}^T \tag{$\overline{\text{BL}}$}$$

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

$$B_k = \begin{bmatrix} \gamma_1, & 0, & \ldots, & 0 \\ \delta_2, & \gamma_2, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & \gamma_k \\ 0, & 0, & \ldots, & \delta_{k+1} \end{bmatrix}$$

(matice $B_k \in R^{(k+1)\times k}$ je bidiagonální).

Věta 85 *Uvažujme bidiagonalizační Lanczosův proces určený maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$. Nechť $\gamma_i \neq 0$, $\delta_i \neq 0$, $1 \le i \le k$, 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 $q_1 = J^T f / \parallel J^T f \parallel$ a $u_1 = f / \parallel f \parallel$. Předpokládejme, že tvrzení platí pro nějaký index $1 \le k < n$ a že $\gamma_{k+1} \neq 0$, $\delta_{k+1} \neq 0$. Použijeme-li ($\overline{\text{BL}}$), dostaneme

$$J^T J Q_k = J^T U_{k+1} B_k = Q_k B_k^T B_k + \gamma_{k+1} q_{k+1} e_{k+1}^T B_k = Q_k T_k + \gamma_{k+1}\delta_{k+1} q_{k+1} e_k^T$$

kde

$$T_k = B_k^T B_k = \begin{bmatrix} \gamma_1^2 + \delta_2^2, & \gamma_2\delta_2, & \ldots, & 0, & 0 \\ \gamma_2\delta_2, & \gamma_2^2 + \delta_3^2, & \ldots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \gamma_{k-1}^2 + \delta_k^2, & \gamma_k\delta_k \\ 0, & 0, & \ldots, & \gamma_k\delta_k, & \gamma_k^2 + \delta_{k+1}^2 \end{bmatrix}$$

Je symetrická tridiagonální matice řádu k. Platí tedy ($\overline{\text{SL}}$), kde $B = J^T J$, $T_k = B_k^T B_k$ a $\alpha_i = \gamma_i^2 + \delta_{i+1}^2$, $\beta_i = \gamma_i\delta_i$, $1 \le i \le k$ a můžeme použít větu 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 první ze vztahů ($\overline{\text{BL}}$) dostaneme

$$U_{k+1}^T J Q_k = U_{k+1}^T U_{k+1} B_k$$

a druhý ze vztahů ($\overline{\mathrm{BL}}$) dává

$$U_{k+1}^T J Q_k = B_k Q_k^T Q_k + \gamma_{k+1} e_{k+1} q_{k+1}^T Q_k = B_k$$

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 134 Z důkazu věty 85 plyne, že symetrický Lanzcosův proces určený SPD 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 $\alpha_i = \gamma_i^2 + \delta_{i+1}^2$, $\beta_i = \gamma_i \delta_i$, $1 \leq i \leq k$, 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 135 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

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \parallel Js + f \parallel$$

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

$$z_i = \arg \min_{z \in R^i} \parallel B_i z + \delta_1 e_1 \parallel$$

(plyne to ze vztahů $f = U_{i+1}(\delta_1 e_1)$, $JQ_i = U_{i+1} B_i$ a $U_{i+1}^T U_{i+1} = I$). Pokud $\gamma_{k+1} \delta_{k+1} = 0$ je vektor $s_{k+1} \in \mathcal{K}_k$, řešením soustavy rovnic $J^T J s + J^T f = 0$ (plyne to z poznámky 116 a poznámky 134).

Poznámka 136 Vektory s_{i+1}, $1 \leq i \leq k$, definované v poznámce 135 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 63 a poznámky 134).

Výhodou bidiagonalizačního Lanczosova procesu je skutečnost, že vektory s_{i+1}, $1 \leq i \leq k$, mohou být určeny pomocí stabilních operací (Givensovy elementární rotace). To tvoří základ metody LSQR. Princip metody LSQR spočívá v tom, že se rekurentně určují rozklady

$$P_i B_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \qquad P_i(\delta_1 e_1) = \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

kde

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \ldots, & 0 \\ 0, & \rho_2, & \ldots, & 0 \\ 0, & 0, & \ldots, & \rho_i \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \ldots \\ \eta_i \end{bmatrix}$$

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. Ukážeme nejprve dva kroky tohoto procesu. Na začátku prvního kroku máme matice

$$B_1 = \begin{bmatrix} \overline{\rho}_1 \\ \delta_2 \end{bmatrix}, \qquad \delta_1 e_1 = \begin{bmatrix} \overline{\eta}_1 \\ 0 \end{bmatrix}$$

kde $\overline{\rho}_1 = \gamma_1$ a $\overline{\eta}_1 = \delta_1$. Položíme-li

$$P_1 = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1, & \delta_2 \\ -\delta_2, & \overline{\rho}_1 \end{bmatrix}$$

dostaneme

$$P_1 B_1 = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1^2 + \delta_2^2 \\ 0 \end{bmatrix} \triangleq \begin{bmatrix} \rho_1 \\ 0 \end{bmatrix}$$

$$P_1(\delta_1 e_1) = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \overline{\rho}_1 \overline{\eta}_1 \\ -\delta_2 \overline{\eta}_1 \end{bmatrix} \triangleq \begin{bmatrix} \eta_1 \\ \overline{\eta}_2 \end{bmatrix}$$

a

$$P_1 \begin{bmatrix} 0 \\ \gamma_2 \end{bmatrix} = \frac{1}{\sqrt{\overline{\rho}_1^2 + \delta_2^2}} \begin{bmatrix} \delta_2 \gamma_2 \\ \overline{\rho}_1 \gamma_2 \end{bmatrix} \triangleq \begin{bmatrix} \sigma_2 \\ \overline{\rho}_2 \end{bmatrix}$$

Na začátku druhého kroku máme matice

$$\begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\rho}_1, & 0 \\ \delta_2, & \gamma_2 \\ 0, & \delta_3 \end{bmatrix} = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \overline{\rho}_2 \\ 0, & \delta_3 \end{bmatrix}, \quad \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\eta}_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \overline{\eta}_2 \\ 0 \end{bmatrix}$$

a můžeme položit

$$P_2 = \begin{bmatrix} 1, & 0 \\ 0, & \overline{P}_2 \end{bmatrix} \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix}, \quad \overline{P}_2 = \frac{1}{\sqrt{\overline{\rho}_2^2 + \delta_3^2}} \begin{bmatrix} \overline{\rho}_2, & \delta_3 \\ -\delta_3, & \overline{\rho}_2 \end{bmatrix}$$

Pokračujeme-li takto dále, dostaneme rekurentní vztahy

$$\overline{\rho}_1 = \gamma_1, \quad \overline{\eta}_1 = \delta_1$$

a

$$\begin{aligned} \rho_i = \sqrt{\overline{\rho}_i^2 + \delta_{i+1}^2} \quad &, \quad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \quad \tau_i = \frac{\delta_{i+1}}{\rho_i} \\ \overline{\rho}_{i+1} = \lambda_i \gamma_{i+1} \quad &, \quad \sigma_{i+1} = \tau_i \gamma_{i+1} \\ \eta_i = \lambda_i \overline{\eta}_i \quad &, \quad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i \end{aligned}$$

pro $1 \leq i \leq k$. Nyní odvodíme rekurentní vztahy pro vektory s_{i+1}, $1 \leq i \leq k$. Jelikož

$$P_i(B_i z + \delta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

a $P_i^T P_i = I$, můžeme položit $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i} \parallel R_i z + h_i \parallel$$

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. Ukážeme nejprve dva kroky tohoto procesu. Na začátku prvního kroku platí

$$R_1 = [\rho_1], \quad h_1 = [\eta_1]$$

a vektor $z_1 = [\zeta_{11}]^T$ můžeme určit ze vztahu

$$R_1 z_1 + h_1 = [\rho_1][\zeta_{11}] + [\eta_1] = 0$$

což dává $\zeta_{11} = -\eta_1/\rho_1$. Platí tedy

$$s_2 = \zeta_{11} q_1 = s_1 + \frac{\eta_1}{\rho_1} p_1$$

kde

$$s_1 = 0, \quad p_1 = -q_1$$

Na začátku druhého kroku platí

$$R_2 = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \rho_2 \end{bmatrix}, \quad h_2 = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

a vektor $z_2 = [\zeta_{21}, \zeta_{22}]$ můžeme určit ze vztahu

$$R_2 z_2 + h_2 = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \rho_2 \end{bmatrix} \begin{bmatrix} \zeta_{21} \\ \zeta_{22} \end{bmatrix} + \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = 0$$

což dává $\zeta_{22} = -\eta_2/\rho_2$ a $\zeta_{21} = -\eta_1/\rho_1 + \zeta_{22}\sigma_2/\rho_1$. Platí tedy

$$s_3 = \zeta_{21} q_1 + \zeta_{22} q_2 = \zeta_{21} q_1 + \zeta_{22} \left(q_2 - \frac{\sigma_2}{\rho_1} q_1 \right) = s_2 + \frac{\eta_2}{\rho_2} p_2$$

kde

$$p_2 = -q_2 + \frac{\sigma_2}{\rho_1} p_1$$

Postupujeme-li takto dále, dostaneme rekurentní vztahy

$$s_1 = 0, \quad p_1 = -q_1$$

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 $1 \leq i \leq k$.

Definice 40 *Nechť $J \in R^{m \times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy*

$$s_1 = 0, \quad \delta_1 u_1 = f, \quad \gamma_1 q_1 = J^T u_1, \quad p_1 = q_1$$

a

$$\delta_{i+1} u_{i+1} = J q_i - \gamma_i u_i$$

$$\gamma_{i+1} q_{i+1} = J^T u_{i+1} - \delta_{i+1} q_i$$

$$\rho_i = \sqrt{\overline{\rho}_i^2 + \delta_{i+1}^2}, \quad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \quad \tau_i = \frac{\delta_{i+1}}{\rho_i}$$

$$\overline{\rho}_{i+1} = \lambda_i \gamma_{i+1}, \quad \sigma_{i+1} = \tau_i \gamma_{i+1}$$

$$\eta_i = \lambda_i \overline{\eta}_i, \quad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i$$

$$s_{i+1} = s_i + \frac{\eta_i}{\rho_i} p_i$$

$$p_{i+1} = -q_{i+1} + \frac{\sigma_{i+1}}{\rho_i} p_i$$

pro $1 \leq i \leq n$, *kde koeficienty* γ_i, δ_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 metodu 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 136 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 86 *Nechť* $s_{i+1} \in R_n$, γ_{i+1}, δ_{i+1}, $\rho_i > 0$, η_i, $1 \leq i \leq k$, *jsou veličiny generované metodou LSQR. Pak pro* $1 \leq i \leq k$ *platí*

$$\| J^T(Js_{i+1} + f) \| = \gamma_{i+1}\delta_{i+1} \frac{|\eta_i|}{\rho_i}$$

Důkaz Nechť $\gamma_{i+1} \neq 0$, $\delta_{i+1} \neq 0$. Pak použitím ($\overline{\text{BL}}$) a poznámky 135 dostaneme

$$\begin{aligned} J^T(Js_{i+1} + f) &= J^T(JQ_i z_i + f) = J^T U_{i+1}(B_i z_i + \delta_1 e_1) = \\ &= (Q_i B_i^T + \gamma_{i+1} q_{i+1} e_{i+1}^T)(B_i z_i + \delta_1 e_1) = \gamma_{i+1} q_{i+1} e_{i+1}^T B_i z_i = \\ &= \gamma_{i+1}\delta_{i+1} q_{i+1} e_i^T z_i \end{aligned}$$

neboť $B_i^T(B_i z_i + \delta_1 e_1) = 0$ podle definice vektoru z_i, $e_{i+1}^T e_1 = 0$ a $e_{i+1}^T B_i = \delta_{i+1} e_i^T$. Ale $Q_i^T Q_i = I$ a tudíž $Q_i^T s_{i+1} = Q_i^T Q_i z_i = z_i$, takže $e_i^T z_i = e_i^T Q_i^T s_{i+1} = q_i^T s_{i+1}$, což spolu s $\| q_{i+1} \| = 1$ dává

$$\| J^T(Js_{i+1} + f) \| = \gamma_{i+1}\delta_{i+1} |q_i^T s_{i+1}|$$

Ale

$$q_i^T s_{i+1} = q_i^T s_i + \frac{\eta_i}{\rho_i} q_i^T p_i = q_i^T Q_{i-1} z_{i-1} - \frac{\eta_i}{\rho_i} q_i^T q_i + \frac{\eta_i \sigma_i}{\rho_i \rho_{i-1}} q_i^T p_{i-1} = -\frac{\eta_i}{\rho_i}$$

neboť $q_i^T Q_{i-1} = 0$, $q_i^T q_i = 1$ a vektor p_{i-1} je lineární kombinací sloupců matice Q_{i-1}, tudíž $q_i^T p_{i-1} = 0$. Jestliže $\gamma_{i+1} = 0$, $\delta_{i+1} = 0$, platí $\| J^T(Js_{i+1} + f) \| = 0$ (poznámka 135).

Větu 86 můžeme využít k zastavení iteračního procesu (není třeba počítat reziduum $\| J^T(Js_{i+1} + f) \|$).

Následující tabulka ukazuje srovnání nepřesné QN metody s lokálně omezeným krokem realizované pomocí řídké reprezentace Hessovy matice a pomocí metody CG se dvěma nepřesnými QN metodami s lokálně omezeným krokem realizovanými pomocí řídké reprezentace Jacobiho matice a pomocí metod CGNE nebo LSQR. Je opět použito 22 testovacích funkcí se 100 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG jakož i celkový počet selhání a celkový čas výpočtu

metoda	MT-NFV-NFG	selhání	čas
GN + CG	1584-1819-1603	–	31.47
GN + CGNE	1602-1835-1621	–	43.67
GN + LSQR	1358-1584-1377	–	51.08

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.

8 Metody pro řešení soustav nelineárních rovnic

8.1 Základní vlastnosti metod pro řešení soustav nelineárních rovnic

Nechť $f : R^n \to R^n$ je spojitě diferencovatelné zobrazení. 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 $f : R^n \to R^n$ je spojitě diferencovatelné. V tomto případě budeme psát $f \in \mathcal{C}^1$ nebo $f \in \mathcal{C}^1 : R^n \to R^n$. Příbuznost metod pro řešení soustav nelineárních rovnic s optimalizačními metodami plyne z toho, že:

- Optimalizační metody můžeme chápat jako metody pro řešení soustavy rovnic $g(x) = 0$, kde $g : R^n \to R^n$ je gradient minimalizované funkce $F : R^n \to R$. V tomto případě jde o speciální soustavu rovnic, neboť Jacobiova matice zobrazení $g : R^n \to R^n$ je Hessovou maticí funkce $F : R^n \to R$ a je tedy symetrická (za standardních podmínek kladených na funkci F). Řešením soustavy rovnic $g(x) = 0$ však můžeme získat nejen lokální minimum, ale i sedlový bod nebo dokonce maximum funkce F.
- Řešení soustavy rovnic můžeme převést na minimalizaci funkce $F : R^n \to R$ definované vztahem $F(x) = (1/2)\|f(x)\|^2$ (součet čtverců). V tomto případě však můžeme získat lokální minimum funkce $F : R^n \to R$, 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 87 *Nechť $f \in \mathcal{C}^1 : R^n \to R^n$ a nechť bod $x^* \in R^n$ je lokálním minimem funkce $F(x) = (1/2)\| f(x) \|^2$, přičemž Jacobiova matice $J(x^*)$ zobrazení $f \in \mathcal{C}^1 : R^n \to R^n$ v bodě $x^* \in R^n$ je regulární. Pak platí $f(x^*) = 0$.*

Důkaz Gradient funkce $F : R^n \to R$ v bodě $x^* \in R^n$ 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 podmínka pro lokální extrém funkce $F : R^n \to R$.

Při vyšetřování konvergence metod pro řešení soustav nelineárních rovnic budeme často používat předpoklady (J3)-(J5):

Definice 41 *Řekneme, že zobrazení $f \in \mathcal{C}^1 : R^n \to R^n$ má omezené derivace, jestliže existuje konstanta $\overline{J} > 0$ taková, že platí*

$$\|J(x)d\| \leq \overline{J}\|d\| \quad \forall x \in R^n \quad \forall d \in R^n. \tag{J3}$$

Podmínka (J3) je ekvivalentní podmínce $\|J(x)\| \leq \overline{J}\ \forall x \in R^n$.

Definice 42 *Řekneme, že zobrazení $f \in \mathcal{C}^1 : R^n \to R^n$ je stejnoměrně ragulární, jestliže existuje konstanta $\underline{J} > 0$ taková, že platí*

$$\|J(x)d\| \geq \underline{J}\|d\| \quad \forall x \in R^n \quad \forall d \in R^n. \tag{J4}$$

Podmínka (J4) je ekvivalentní podmínce $\|J^{-1}(x)\| \leq 1/\underline{J}\ \forall x \in R^n$.

Definice 43 *Řekneme, že zobrazení $f \in \mathcal{C}^1 : R^n \to R^n$ má lipschitzovské derivace, jestliže existuje konstanta $\overline{L} > 0$ taková, že platí*

$$\|J(x+d) - J(x)\| \leq \overline{L}\|d\| \quad \forall x \in R^n \quad \forall d \in R^n. \tag{F5}$$

Uvedené podmínky mají podobný význam jako podmínky (F3)-(F5) kladené na funkci $F : R^n \to R$. Je-li splněna podmínka (J3), můžeme podmínku (J4) nahradit ekvivalentní podmínkou

$$\kappa(J(x)) \leq \overline{J}/\underline{J} \tag{$\overline{\text{J4}}$}$$

$\forall x \in R^n$, kde $\kappa(J(x))$ je spektrální číslo podmíněnosti matice $J(x)$.

Podobně jako jsme definovali základní optimalizační metodu (oddíl 1.4), můžeme definovat základní metodu pro řešení soustav nelineárních rovnic jako 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 44 *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

$$\begin{aligned} s_i &= -J^{-1}(x_i)f(x_i), \\ \alpha_i &= 1. \end{aligned}$$

Směrový vektor Newtonovy metody pro řešení soustav nelineárních rovnic je shodný se směrovým vektorem Gaussovy-Newtonovy metody pro minimalizaci součtu čtverců $F(x) = (1/2)\|f(x)\|^2$, neboť (je-li splněna podmínka (J4)) 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 5.9) .

V dalším textu se budeme zabývat metodami, které místo Jacobiových matic $J_i = J(x_i)$, $i \in N$, používají jejich aproximace A_i, $i \in N$, splňující podmínky

$$\|A_i s\| \leq \overline{A}s \qquad \forall s \in R^n, \tag{A3}$$

$$\|A_i s\| \geq \underline{A}s \qquad \forall s \in R^n, \tag{A4}$$

$$\|A_i - J_i\| \leq \overline{\vartheta}. \tag{A5}$$

Podmínka (A3) je ekvivalentní podmínce $\|A\| \leq \overline{A}$. Podmínka (A4) je ekvivalentní podmínce $\|A^{-1}\| \leq 1/\underline{A}$. Poznamenejme, že z (J3) a (A4)-(A5) plyne (A3) s $\overline{A} = \overline{J} + \overline{\vartheta}$.

V důkazech globální konvergence metod pro řešení nelineárních rovnic se vyžaduje, aby čísla $\overline{\vartheta} > 0$ a $\underline{A} > 0$ splňovala nerovnost $\overline{\vartheta} \leq \gamma \underline{A}$, kde $0 < \gamma < 1$ je konstanta závislá na zvolené metodě. Nejprve je třeba dokázat existenci takovýchto čísel.

Lemma 23 *Nechť jsou splněny předpoklady (J3)-(J4), nechť* $0 < \gamma < 1$ *a*

$$\underline{A} \leq \underline{J}(\sqrt{\gamma^2\kappa^2 + 1} - \gamma\kappa),$$

kde $\kappa = \overline{J}/\underline{J}$*. Pak, vyhovují-li matice* A_i, $i \in N$*, podmínce (A5) s* $\overline{\vartheta} \leq \gamma \underline{A}$*, vyhovují i podmínce (A4).*

Důkaz Nechť $\overline{J}$ a $\underline{J}$ jsou konstanty z (J3)-(J4), $\kappa = \overline{J}/\underline{J}$ a $0 < \gamma < 1$. Nechť A_i je matice vyhovující podmínce (A5). Protože $A_i s = J_i s + (A_i - J_i)s$, můžeme psát

$$\|A_i s\|^2 = s^T J_i^T J_i s + 2s^T (A_i - J_i)^T J_i s + \|(A_i - J_i)s\|^2 \geq \underline{J}^2 \|s\|^2 - 2\overline{\vartheta}\overline{J}\|s\|^2$$

$\forall s \in R^n$, takže (A4) platí pro libovolné číslo $\underline{A}$ takové, že $\underline{A}^2 \leq \underline{J}^2 - 2\overline{\vartheta}\overline{J}$, neboli

$$2\overline{\vartheta}\overline{J} \leq \underline{J}^2 - \underline{A}^2.$$

Nerovnost $\overline{\vartheta} \leq \gamma \underline{A}$ můžeme zapsat ve tvaru

$$2\overline{\vartheta}\overline{J} \leq 2\gamma \overline{J} \underline{A}.$$

Jelikož pravá část první nerovnosti klesá a pravá část druhé nerovnosti vzrůstá se vzrůstající hodnotou $\underline{A}$, dostaneme největší hodnotu $\overline{\vartheta}$ pokud $2\gamma\overline{J}\underline{A} = \underline{J}^2 - \underline{A}^2$, což je kvadratická rovnice, jejíž kladný kořen lze vyjádřit ve tvaru

$$\underline{A} = \underline{J}(\sqrt{\gamma^2\kappa^2 + 1} - \gamma\kappa).$$

Největší možná hodnota $\overline{\vartheta}$, která splňuje nerovnost $\overline{\vartheta} \leq \gamma \underline{A}$ a zaručuje platnost vztahu (A4) je tedy

$$\overline{\vartheta} = \gamma \underline{J}(\sqrt{\gamma^2\kappa^2 + 1} - \gamma\kappa).$$

Platí-li (A5) s $\overline{\vartheta} \leq \gamma \underline{J}(\sqrt{\gamma^2\kappa^2 + 1} - \gamma\kappa)$, platí nutně (A4) s $\underline{A} = \underline{J}(\sqrt{\gamma^2\kappa^2 + 1} - \gamma\kappa)$.

Poznámka 137 Vzhledem k nerovnosti uvedené v lemmatu 23 můžeme předpokládat, že číslo $\underline{A}$ použité v (A4) splňuje nerovnost $\underline{A} \leq \underline{J} \leq \overline{J}$.

8.2 Metody spádových směrů

Při výkladu metod spádových směrů 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 45, implikuje nerovnost $h_i^T s_i = f_i^T A_i s_i < 0$.

Definice 45 *Řekněme, 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:*
(1) 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 $0 \leq \overline{\omega}_i \leq \overline{\omega} < 1$.
(2) Délky kroku $\alpha_i > 0$, $i \in N$*, se určují 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

$$F_{i+1} - F_i \le -2\underline{\rho}(1-\overline{\omega})\alpha_i F_i, \qquad (\overline{\text{S2b}})$$

nebo

$$\|f_{i+1}\| - \|f_i\| \le -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|, \qquad (\overline{\text{S2c}})$$

kde $0 < \underline{\beta} \le \overline{\beta} < 1$ *a* $0 < \underline{\rho} < 1$.

Lemma 24 *Nechť funkce* $f : R^n \to R^n$ *vyhovuje předpokladům (J3)-(J5). Nechť matice* A_i, $i \in N$, *splňují podmínky (A3)-(A5) s* $\overline{\vartheta} \le \gamma\underline{A}$, *kde* $\gamma < \lambda(1-\underline{\rho})$, $0 < \lambda = (1-\overline{\omega})/(1+\overline{\omega}) \le 1$ *a* $0 < \underline{\rho} < 1$ *(korektnost těchto podmínek zaručuje lemma 23). Pak lze v každém iteračním kroku nalézt směrový vektor* $s_i \in R^n$ *vyhovující podmínce* $(\overline{\text{S1}})$ *a délku kroku* $\alpha_i > 0$ *vyhovující libovolné z podmínek* $(\overline{\text{S2}})$. *Navíc existuje konstanta* $0 < \underline{\alpha} < 1 - \underline{\rho}$ *taková, že* $\alpha_i \ge \underline{\alpha}$ $\forall i \in N$.

Důkaz Existence směrového vektoru $s_i \in R^n$ vyhovujícího podmínce $(\overline{\text{S1}})$ plyne bezprostředně z (A4) (jelikož matice A_i je podle (A4) regulární, můžeme vektor s_i zvolit tak, že $\|A_i s_i + f_i\| = 0$). Z podmínky $(\overline{\text{S1}})$ z (A5) a z definice vektorů $g_i = J_i^T f_i$, $h_i = A_i^T f_i$ lze jednoduše odvodit nerovnosti

$$(1-\overline{\omega})\|f_i\| \le \|A_i s_i\| \le (1+\overline{\omega})\|f_i\|, \qquad \text{(p)}$$

$$(1-\overline{\omega})\|f_i\|^2 \le -h_i^T s_i \le (1+\overline{\omega})\|f_i\|^2, \qquad \text{(q)}$$

$$|h_i^T s_i - g_i^T s_i| \le \overline{\vartheta}\|f_i\|\|s_i\|. \qquad \text{(r)}$$

Nerovnost (p) spolu s (A3)-(A4) dává

$$\frac{1-\underline{\omega}}{\overline{A}}\|f_i\| \le \|s_i\| \le \frac{1+\overline{\omega}}{\underline{A}}\|f_i\|, \qquad \text{(s)}$$

přičemž předpokládáme, že $\overline{\vartheta} \le \gamma\underline{A}$ pro nějaké, zatím neurčené číslo $\gamma < \lambda$.
Použitím (p) - (s) dostaneme

$$-g_i^T s_i \ge -h_i^T s_i - \overline{\vartheta}\|f_i\|\|s_i\| \ge -h_i^T s_i - \overline{\vartheta}\frac{1+\overline{\omega}}{\underline{A}}\|f_i\|^2 \ge -h_i^T s_i - (1-\overline{\omega})\frac{\gamma}{\lambda}\|f_i\|^2$$

$$\ge -(1-\gamma/\lambda)h_i^T s_i \ge (1-\overline{\omega})(1-\gamma/\lambda)\|f_i\|^2 > 0, \qquad \text{(t)}$$

takže podle lemmatu 1 existuje pro libovolné číslo $0 < \varepsilon_1 < 1$ délka kroku $\alpha_i > 0$ taková, že

$$F_{i+1} - F_i \le \varepsilon_1\alpha_i g_i^T s_i \le \varepsilon_1\alpha_i(1-\gamma/\lambda)h_i^T s_i \le -\varepsilon_1\alpha_i(1-\overline{\omega})(1-\gamma/\lambda)\|f_i\|^2.$$

Položme $\underline{\rho} = \varepsilon_1(1-\gamma/\lambda)$, takže $0 < \underline{\rho} < 1$. Pak

$$F_{i+1} - F_i \le \underline{\rho}\alpha_i h_i^T s_i \le -2\underline{\rho}(1-\overline{\omega})\alpha_i F_i,$$

takže podmínky $(\overline{\text{S2a}})$ a $(\overline{\text{S2b}})$ jsou konzistentní, pokud $0 < \underline{\rho} < 1$ a číslo γ je zvoleno tak, že $0 \ge \gamma < \lambda(1-\underline{\rho})$. Podmínka $(\overline{\text{S2c}})$ je také konzistentní, neboť z

$$2\|f_i\|(\|f_{i+1}\| - \|f_i\|) \le (\|f_{i+1}\| + \|f_i\|)(\|f_{i+1}\| - \|f_i\|) = 2(F_{i+1} - F_i)$$

a z $(\overline{\text{S2b}})$ plyne, že

$$\|f_{i+1}\| - \|f_i\| \le \frac{F_{i+1} - F_i}{\|f_i\|} \le -2\underline{\rho}(1-\overline{\omega})\alpha_i\frac{F_i}{\|f_i\|} = -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|.$$

Poznamenejme, že kromě konzistence podmínek ($\overline{\text{S2a}}$)-($\overline{\text{S2c}}$) jsme též dokázali implikace ($\overline{\text{S2a}}$) $\Rightarrow$ ($\overline{\text{S2b}}$) $\Rightarrow$ ($\overline{\text{S2c}}$). Platí i opačné implikace. Z ($\overline{\text{S2b}}$) a (b) dostaneme $F_{i+1} - F_i \leq -\underline{\rho}(1-\overline{\omega})\alpha_i\|f_i\|^2 \leq \underline{\rho}\lambda\alpha_i h_i^T s_i$, takže platí ($\overline{\text{S2a}}$), kde hodnota $\underline{\rho}$ je nahražena hodnotou $\lambda\underline{\rho}$. 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 je-li splněna podmínka ($\overline{\text{S2c}}$) pro nějakou hodnotu parametru $\underline{\rho}$, je splněna podmínka ($\overline{\text{S2b}}$) pro poloviční hodnotu tohoto parametru. Nyní se omezíme na podmínku ($\overline{\text{S2a}}$) (neboť jsme právě dokázali, že podmínky ($\overline{\text{S2a}}$)-($\overline{\text{S2c}}$) jsou ekvivalentní v tom smyslu, že platnost jedné z nich implikuje platnost libovolné jiné pro nějakou hodnotu parametru ρ). Při výběru délky kroku 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 větu o střední hodnotě (pokládáme $d_i = \mu(\alpha_i/\beta)s_i$, kde $0 \leq \mu \leq 1$) a předpoklady (J3)-(J5), můžeme psát

$$\begin{aligned} F(x_i + \frac{\alpha_i}{\beta}s_i) - F(x_i) &= \frac{\alpha_i}{\beta}g^T(x_i+d_i)s_i \\ &\leq \frac{\alpha_i}{\beta}\left(g_i^T s_i + \|g(x_i+d_i)-g(x_i)\|\|s_i\|\right) \\ &\leq \frac{\alpha_i}{\beta}\left(g_i^T s_i + \frac{\alpha_i}{\beta}(\overline{J}^2 + \overline{LF})\|s_i\|^2\right), \end{aligned}$$

kde $\overline{F}$ je libovolná konstanta taková, že $\overline{F} \geq \|f_1\|$, neboť

$$\begin{aligned} \|g(x_i+d_i)-g(x_i)\| &= \|J^T(x_i+d_i)f(x_i+d_i) - J^T(x_i)f(x_i)\| \\ &\leq \|J^T(x_i+d_i)(f(x_i+d_i)-f(x_i))\| + \|(J^T(x_i+d_i)-J^T(x_i))f(x_i)\| \\ &\leq \overline{J}\|(f(x_i+d_i)-f(x_i)\| + \overline{L}\|d_i\|\|f_i\| \\ &= \overline{J}\left\|\int_0^1 J(x_i+\tau d_i)d_i \mathrm{d}\tau\right\| + \overline{L}\|d_i\|\|f_i\| \\ &\leq \left(\overline{J}^2 + \overline{LF}\right)\|d_i\| \leq \frac{\alpha_i}{\beta}\left(\overline{J}^2+\overline{LF}\right)\|s_i\|. \end{aligned}$$

Spojíme-li obě nerovnosti, dostaneme

$$\underline{\rho}h_i^T s_i \leq g_i^T s_i + \frac{\alpha_i}{\beta}(\overline{J}^2+\overline{LF})\|s_i\|^2$$

a použijeme-li (s) a (t), můžeme psát

$$\frac{\alpha_i}{\beta}\left(\overline{J}^2+\overline{LF}\right)\frac{(1+\overline{\omega})^2}{\underline{A}^2}\|f_i\|^2 \geq \underline{\rho}h_i^T s_i - g_i^T s_i \geq (\underline{\rho}-(1-\gamma/\lambda))h_i^T s_i \geq (1+\overline{\omega})(1-\underline{\rho}-\gamma/\lambda)\|f_i\|^2$$

což spolu s $\beta \geq \underline{\beta}$ dává $\alpha_i \geq \underline{\alpha}$, kde

$$\underline{\alpha} = \frac{\underline{\beta}(1-\underline{\rho}-\gamma/\lambda)\underline{A}^2}{(\overline{J}^2+\overline{LF})(1+\overline{\omega})}.$$

Jelikož $0 < \underline{\beta} < 1$, $0 < (1-\underline{\rho}-\gamma/\lambda) < 1-\underline{\rho}$, $0 < \underline{A} \leq \underline{J} \leq \overline{J}$ (poznámka 137) a $0 \leq \overline{\omega} < 1$, platí $0 < \underline{\alpha} < 1-\underline{\rho}$, takže $\alpha_i \geq \underline{\alpha}$ i v případě, že $\alpha_i = \alpha_i^1 = 1$.

Poznámka 138 Lemma 24 ukazuje, že nestačí dostatečně přesně řešit soustavu lineárních rovnic $A_i s_i + f_i = 0$, tak jako v případě nepodmíněné minimalizace, ale že je též třeba dostatečně přesně aproximovat Jacobiovu matici J_i (nerovnost $\overline{\vartheta} \leq \gamma \underline{A}$). Je to způsobeno tím, že ve vztazích pro gradienty funkce $F : R^n \to R$ vystupují Jacobiovy matice J_i, které jsou různé od matic A_i (nepřesná aproximace Jacobiovy matice implikuje nepřesnost gradientu).

Věta 88 *(globální konvergence). Nechť jsou splněny předpoklady lemmatu 24. Nechť $x_i \in R^n$, $i \in N$ je posloupnost generovaná metodou spádových směrů ($\overline{\text{S1}}$)-($\overline{\text{S2}}$). Potom $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Důkaz provedeme pro ($\overline{\text{S2b}}$), neboť podmínky ($\overline{\text{S2a}}$)-($\overline{\text{S2c}}$) jsou ekvivalentní (jak je ukázáno v důkazu lemmatu 24). Podle ($\overline{\text{S2b}}$) platí

$$F_{i+1} \leq (1 - 2\underline{\rho}(1 - \overline{\omega})\alpha_i)F_i \leq (1 - 2\underline{\rho}(1 - \overline{\omega})\underline{\alpha})F_i \triangleq qF_i,$$

kde $0 < q < 1$, neboť $0 < 1 - \overline{\omega} \leq 1$ a z $0 < \underline{\alpha} < 1 - \underline{\rho}$ plyne, že $0 < 2\underline{\rho}\underline{\alpha} < 2\underline{\rho}(1 - \underline{\rho}) < 1$. Porovnáním s geometrickou řadou dostaneme

$$\sum_{i=1}^{\infty} F_i \leq \frac{1}{1-q} F_1 < \infty,$$

což implikuje $F_i \to 0$ a tedy i $\|f_i\| = \sqrt{2F_i} \to 0$. Z nerovnosti (s) svazující normy $\|s_i\|$ a $\|f_i\|$ dostaneme

$$\sum_{i=1}^{\infty} \|s_i\| \leq \frac{1 + \overline{\omega}}{\underline{A}} \sum_{i=1}^{\infty} \|f_i\| < \infty,$$

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 $f(x^\star) = 0$.

Poznámka 139 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 $0 < \underline{\rho} < 1$. Kdybychom chtěli použít podmínky ($\overline{\text{S2a}}$), ($\overline{\text{S2b}}$), museli bychom volit $\underline{\rho}$ tak, aby platilo $0 < \underline{\rho} < 1/2$ (poznámka 140).

Věta 89 *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná metodou spádových směrů taková, že $x_i \to x^*$. Nechť jsou splněny předpoklady (J3)-(J5). Nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínce ($\overline{\text{S2c}}$). Nechť platí*

$$\lim_{i \to \infty} \frac{\|A_i s_i + f_i\|}{\|f_i\|} = 0 \qquad (\alpha)$$

a

$$\lim_{i \to \infty} \frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0 \qquad (\beta).$$

Pak existuje index $k \in N$ takový, že $\alpha_i = 1$, $\forall i \geq k$ a posloupnost x_i, $i \in N$, konverguje superlineárně k bodu $x^* \in R^n$.

Důkaz Důkaz povedeme poněkud obecněji, neboť získané výsledky použijeme v důkazu věty 99. To znamená, že v částech (a)-(b) budeme předpokládat pouze platnost podmínky ($\overline{\text{S1}}$), takže

$$\limsup_{i \to \infty} \frac{\|A_i s_i + f_i\|}{\|f_i\|} \leq \overline{\omega} < 1.$$

Platí-li (α), můžeme ve všech vzorcích položit $\overline{\omega} = 0$.
(a) Ukážeme, že existuje index $k_1 \in N$ tak, že

$$\|f_i\|(1-\overline{\omega})/\overline{J} \le \|s_i\| \le \|f_i\|(1+\overline{\omega})/\underline{J}$$

$\forall i \ge k_1$, pokud $\|J^*\| < \overline{J}$ a $\|(J^*)^{-1}\| < 1/\underline{J}$. 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 (β)) a $\|J_i\| \to \|J^*\| < \overline{J}$, existuje index $k_0 \in N$ tak, že $\|s_i\| \ge \|f_i\|(1-\overline{\omega})/\overline{J}$ $\forall i \ge k_0$. 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 (β)) a $\|J_i^{-1}\| \to \|(J^*)^{-1}\| < 1/\underline{J}$, existuje index $k_1 \ge k_0$ tak, že $\|s_i\| \le \|f_i\|(1+\overline{\omega})/\underline{J}$ $\forall i \ge k_1$.
(b) Ukážeme, že existuje index $k \ge k_1$ tak, že hodnota $\alpha_i = 1$ vyhovuje podmínce ($\overline{\text{S2b}}$), pokud $\underline{\rho} < 1 - \overline{\omega}$. Použijeme-li větu o střední hodnotě, 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

$$\frac{\|f(x_i+s_i)\|}{\|f_i\|} \le \|\omega_i\| + \|\vartheta_i\|(1+\overline{\omega})/\underline{J} + o(\|f_i\|)/\|f_i\|,$$

takže $\limsup_{i\to\infty}\|f(x_i+s_i)\|/\|f_i\| \le \overline{\omega}$ (podle ($\overline{\text{S1}}$) a (β)). Pokud $\underline{\rho} < 1 - \overline{\omega}$, existuje index $k \ge k_1$ tak, že podmínka ($\overline{\text{S2b}}$) s $\alpha_i = 1$ je splněna $\forall i \ge k$ (platí-li (α), může být číslo $0 < \underline{\rho} < 1$ libovolné, neboť $\|f(x_i+s_i)\|/\|f_i\| \to 0$).
(c) Předpokládejme nyní že platí (α). Pomocí vět 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 (α), (β) a (b) 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 140 Věta 89 zůstane v platnosti, i tehdy používáme-li k výběru délky kroku podmínky ($\overline{\text{S2a}}$)-($\overline{\text{S2b}}$). Abychom mohli používat kroky jednotkové délky, což se předpokládá v důkazu superlineární konvergence, musí v tomto případě platit $\underline{\rho} < 1/2$, neboť z ($\overline{\text{S2b}}$) plyne $2\underline{\rho}\alpha_i \le 1 - F_{i+1}/F_i < 1$ (předpokládáme, že $F_i > 0$ $\forall i \in N$), takže $\alpha_i = 1$ lze volit pouze tehdy, pokud $2\underline{\rho} < 1$.

Poznámka 141 Položíme-li $A = J$, dostaneme Newtonovu metodu. V tomto případě podmínky (J3)-(J4) implikují (A3)-(A4) a (A5) platí s $\overline{\vartheta} = 0$, takže lze položit $\gamma = 0$ ve všech vzorcích uvedených v předchozím textu. Tyto úvahy ukazují, že Newtonova metoda realizovaná jako metoda spádových směrů je globálně konvergentní (platí-li (J3)-(J4)). Lemma 25 ukazuje, jak lze vlastnosti libovolné metody spádových směrů odvodit z vlastností Newtonovy metody.

Lemma 25 *Nechť funkce $f : R^n \to R^n$ vyhovuje předpokladům (J3)-(J5). Nechť matice A_i, $i \in N$, splňují podmínku (A5) s $\overline{\vartheta} < (1/2)(1-\overline{\omega})\underline{J}$. Pak platí*

$$\|J_i s_i + f_i\| \le \overline{\omega}'\|f_i\|,$$

kde $\overline{\omega}' = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$. Jinými slovy, platí-li ($\overline{\text{S1}}$) a (A5) s $\overline{\vartheta} < (1/2)(1-\overline{\omega})\underline{J}$, můžeme vektor s_i považovat za směrový vektor získaný Newtonovou metodou, kde příslušná soustava lineárních rovnic je řešena s přesností $\overline{\omega}' = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$.

Důkaz Použijeme-li (a) a (A5), dostaneme

$$(1+\overline{\omega})\|f_i\| \ge \|A_i s_i\| \ge \|J_i s_i\| - \|(A_i - J_i)s_i\| \ge (\underline{J} - \overline{\vartheta})\|s_i\|,$$

neboli

$$\|s_i\| \le \frac{1+\overline{\omega}}{\underline{J} - \overline{\vartheta}}\|f_i\|.$$

Můžeme tedy psát

$$\|J_i s_i + f_i\| \le \|A_i s_i + f_i\| + \|(J_i - A_i)s_i\| \le \overline{\omega}\|f_i\| + \overline{\vartheta}\|s_i\| \le \frac{\underline{J}\overline{\omega} + \overline{\vartheta}}{\underline{J} - \overline{\vartheta}}\|f_i\| \triangleq \overline{\omega}'\|f_i\|.$$

Přitom $\overline{\omega}' = (\underline{J}\overline{\omega} + \overline{\vartheta})/(\underline{J} - \overline{\vartheta}) < 1$, pokud $\overline{\vartheta} < (1/2)(1-\overline{\omega})\underline{J}$.

Teoretické výsledky shrnuté v lemmatech 24 a 25 vyžadují splnění podmínky (A5) (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 2 Data $0 \le \overline{\omega} < 1$, $0 < \underline{\rho} < 1$, $0 < \underline{\beta} < \overline{\beta} < 1$, $\overline{\varepsilon} > 0$, $0 < \overline{j}_1 \le \overline{j}_2$.

Krok 1 Zvolíme počáteční odhad $x_1 \in R^n$, vypočteme $f_1 = f(x_1)$ a položíme $i = 1$ a $k = 1$.

Krok 2 Pokud $\|f_i\| \le \overline{\varepsilon}$, ukončíme výpočet.

Krok 3 Pokud $k = 1$, vypočteme Jacobiovu matici $J_i = J(x_i)$ a položíme $A_i = J_i$ (restart). Zvolíme přesnost $0 \le \overline{\omega}_i \le \overline{\omega} < 1$ a vypočteme směrový vektor $s_i \in R^n$ vyhovující podmínce ($\overline{\text{S1}}$).

Krok 4a Položíme $\alpha_i^1 = 1$ a $j = 1$.

Krok 4b Položíme $x_{i+1} = x_i + \alpha_i^j s_i$ a vypočteme $f_i = f(x_i)$. Je-li splněna některá (vybraná) podmínka z ($\overline{\text{S2}}$), přejdeme na krok 5.

Krok 4c Pokud $k = 1$ a $j > \overline{j}_2$, ukončíme výpočet (předčasné ukončení způsobené selháním Newtonovy metody). Pokud $k > 1$ a $j > \overline{j}_1$, položíme $k = 1$ a přejdeme na krok 3. V ostatních případech určíme délku kroku α_i^{j+1} tak aby platilo $\underline{\beta}\alpha_i^j \le \alpha_i^{j+1} \le \overline{\beta}\alpha_i^j$ a přejdene na krok 4b.

Krok 5 Určíme novou matici A_{i+1} (například pomocí kvazinewtonovské aktualizace), položíme $i := i + 1$, $k := k + 1$ a přejdeme na krok 2.

8.3 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 funkci, která lokálně aproximuje rozdíl $\|f(x_i + s)\| - \|f(x_i)\|$ a označení

$$\rho_i(s) = (\|f(x_i + s)\| - \|f_i(x_i)\|)/L_i(s)$$

pro podíl skutečného a předpověděného poklesu normy funkce $f : R^n \to R^n$.

Definice 46 *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

$$\|s_i\| \le \Delta_i, \tag{$\overline{\text{T1a}}$}$$

$$\|s_i\| < \Delta_i \Rightarrow \|A_i s_i + f_i\| \le \overline{\omega}_i \|f_i\|, \tag{$\overline{\text{T1b}}$}$$

$$-L_i(s_i) \ge \underline{\sigma} \|A_i s_i\|, \tag{$\overline{\text{T1c}}$}$$

kde $0 \le \overline{\omega}_i \le \overline{\omega} < 1$ a $0 < \underline{\sigma} < 1$.
(2) Délky kroku $\alpha_i \ge 0$, $i \in N$, se určují tak, že

$$\rho_i(s_i) \le 0 \Rightarrow \alpha_i = 0, \tag{$\overline{\text{T2a}}$}$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1. \tag{$\overline{\text{T2b}}$}$$

(3) Meze $0 < \Delta_i \le \overline{\Delta}$, $i \in N$, se určují tak, že

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta}\|s_i\| \le \Delta_{i+1} \le \overline{\beta}\|s_i\|, \tag{$\overline{\text{T3a}}$}$$

$$\rho_i(s_i) \ge \underline{\rho} \Rightarrow \Delta_i \le \Delta_{i+1} \le \overline{\Delta}, \tag{$\overline{\text{T3b}}$}$$

kde $0 < \underline{\beta} < \overline{\beta} < 1$ a $0 < \underline{\rho} < 1/2$.

V dalším textu budeme používat označení N_1, N_2 a N_3 pro množiny indexů takové, že $\|s_i\| < \Delta_i$, $\rho_i(s_i) > 0$ a $\rho_i(s_i) \ge \underline{\rho}$.

Lemma 26 *Nechť funkce $f : R^n \to R^n$ vyhovuje předpokladům (J3)-(J5). Nechť matice A_i, $i \in N$, splňují podmínky (A5)-(A4) s $\overline{\vartheta} < \gamma \underline{A}$, kde $\gamma = (1/2 - \underline{\rho})\underline{\sigma}$ (splnění těchto podmínek zaručuje lemma 23). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem ($\overline{\text{T1}}$)-($\overline{\text{T3}}$). 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 $(1 - \overline{\omega})\|f_i\| \le \|A_i s_i\|$. Z druhé strany podmínka (A5) dává

$$\|A_i s_i\| \le \|J_i s_i\| + \|(A_i - J_i)s_i\| \le (\overline{J} + \overline{\vartheta})\|s_i\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$\|s_i\| \geq \frac{1-\overline{\omega}}{\overline{J}+\overline{\vartheta}}\|f_i\|.$$

(b) Nechť $i \notin N_1$ a $i \notin N_3$. Z ($\overline{\text{T1c}}$) plyne, že $L_i(s_i) \leq 0$, takže

$$\begin{aligned} L_i(s_i)\|f_i\| &= \left(\|A_i s_i + f_i\| - \|f_i\|\right)\|f_i\| \geq \left(\|A_i s_i + f_i\|^2 - \|f_i\|^2\right) \\ &= 2\left(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i\right) \triangleq 2Q_i(s_i). \end{aligned} \qquad (*)$$

Jestliže $\|f(x_i + s_i)\| \leq \|f(x_i)\|$, pak nerovnost $\rho_i(s_i) < \underline{\rho}$ spolu s $(*)$ dává

$$\begin{aligned} F(x_i + s_i) - F(x_i) &= \frac{1}{2}\left(\|f(x_i + s_i)\|^2 - \|f(x_i)\|^2\right) \\ &\geq \left(\|f(x_i + s_i)\| - \|f(x_i)\|\right)\|f(x_i)\| \\ &\geq \underline{\rho} L_i(s_i)\|f_i\| \geq 2\underline{\rho} Q_i(s_i). \end{aligned}$$

Jestliže $\|f(x_i + s_i)\| \geq \|f(x_i)\|$, platí tato nerovnost triviálně. Můžeme tedy psát

$$F(x_i + s_i) - F(x_i) \geq 2\underline{\rho} Q_i(s_i).$$

Z druhé strany, použijeme-li větu o střední hodnotě (pokládáme $d_i = \mu s_i$, kde $0 \leq \mu \leq 1$) a předpoklady (J3)-(J5), můžeme psát

$$\begin{aligned} F(x_i + s_i) - F(x_i) &\leq g_i^T s_i + \|g(x_i + d_i) - g(x_i)\|\|s_i\| \\ &\leq g_i^T s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &= f_i^T A_i s_i + f_i^T (J_i - A_i) s_i + (\overline{J}^2 + \overline{LF})\|s_i\|^2 \\ &\leq Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2, \end{aligned}$$

kde $\overline{F}$ je libovolná konstanta taková, že $\overline{F} \geq \|f_1\|$, neboť tak jako v důkazu lemmatu 24 platí

$$\|g(x_i + d_i) - g(x_i)\| \leq (\overline{J}^2 + \overline{LF})\|d_i\| \leq (\overline{J}^2 + \overline{LF})\|s_i\|.$$

Spojíme-li obě nerovnosti, dostaneme

$$2\underline{\rho} Q_i(s_i) \leq Q_i(s_i) + \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2,$$

neboli

$$-(1 - 2\underline{\rho}) Q_i(s_i) \leq \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2.$$

Podmínky ($\overline{\text{T1c}}$) a (A4) spolu s nerovností $(*)$ dávají

$$-Q_i(s_i) \geq -\frac{1}{2} L_i(s_i)\|f_i\| \geq \frac{\sigma}{2}\|A_i s_i\|\|f_i\| \geq \frac{\sigma}{2}\underline{A}\|s_i\|\|f_i\|.$$

Dosadíme-li tento vztah do předchozí nerovnosti, dostaneme

$$(1 - 2\underline{\rho})\frac{\sigma}{2}\underline{A}\|s_i\|\|f_i\| \leq -(1 - 2\underline{\rho}) Q_i(s_i) \leq \overline{\vartheta}\|s_i\|\|f_i\| + (\overline{J}^2 + \overline{LF})\|s_i\|^2,$$

neboli

$$\|s_i\| \geq \frac{(1/2-\underline{\rho})\underline{\sigma A}-\overline{\vartheta}}{\overline{J}^2+\overline{LF}}\|f_i\| \geq \frac{((1/2-\underline{\rho})\underline{\sigma}-\gamma)\underline{A}}{\overline{J}^2+\overline{LF}}\|f_i\|$$

(čitatel je kladný, neboť $\gamma < (1/2-\underline{\rho})\underline{\sigma}$).
(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

$$\underline{c} = \underline{\beta}\min\left(\frac{1-\overline{\omega}}{\overline{J}+\overline{\vartheta}}, \frac{((1/2-\underline{\rho})\underline{\sigma}-\gamma)\underline{A}}{\overline{J}^2+\overline{LF}}, \frac{\|s_1\|}{\|f_1\|}\right).$$

Věta 90 *(globální konvergence). Nechť jsou splněny předpoklady lemmatu 26. Pak $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz (a) Nejprve ukážeme, že $f_i \to 0$. Předpokládejme,že toto tvrzení neplatí. Protože posloupnost $\|f_i\|$, $i \in N$, je nerostoucí podle ($\overline{\text{T2a}}$)-($\overline{\text{T2b}}$), existuje číslo $\underline{\varepsilon} > 0$ takové, že $\|f_i\| \geq \underline{\varepsilon}$, $\forall i \in N$ a podle lemmatu 26 platí

$$\|s_i\| \geq \underline{c\varepsilon}, \quad \forall i \in N.$$

Předpokládejme nejprve, že množina N_3 je nekonečná. Protože $N_3 \subset N_2$, můžeme psát

$$\begin{aligned}\|f_i\| - \|f_{i+1}\| &= \|f(x_i)\| - \|f(x_i+s_i)\| \geq -\underline{\rho}L_i(s_i)\\ &\geq \underline{\rho\sigma}\|A_i s_i\| \geq \underline{\rho\sigma A c\varepsilon}, \ \forall i \in N_3.\end{aligned}$$

Odtud plyne

$$\begin{aligned}\|f_1\| &\geq \lim_{i\to\infty}(\|f_1\| - \|f_{i+1}\|) = \sum_{i=1}^{\infty}(\|f_i\| - \|f_{i+1}\|)\\ &\geq \sum_{i\in N_3}(\|f_i\| - \|f_{i+1}\|) \geq \sum_{i\in N_3}\underline{\rho\sigma A c\varepsilon} = \infty,\end{aligned}$$

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

$$\underline{A}\|s_i\| \leq \|A_i s_i\| \leq 2\|f_i\|.$$

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

$$\begin{aligned}\|f_1\| &\geq \sum_{i=1}^{\infty}(\|f_i\| - \|f_{i+1}\|) \geq \sum_{i\in N_3}(\|f_i\| - \|f_{i+1}\|)\\ &\geq \underline{\rho\sigma}\sum_{i\in N_3}\|A_i s_i\| \geq \underline{\rho\sigma A}\sum_{i\in N_3}\|s_i\|.\end{aligned}$$

Označme $N_3 = \{l_1, l_2, l_3, \ldots\}$. Použijeme-li lemma 26, dostaneme

$$\|s_{l_j+1}\| \leq \frac{2}{\underline{A}}\|f_{l_j+1}\| \leq \frac{2}{\underline{cA}}\|s_{l_j}\|$$

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

$$\begin{aligned}\sum_{i=1}^{\infty}\|s_i\| &= \sum_{i=1}^{l_1-1}\|s_i\| + \sum_{j=1}^{\infty}\left[\|s_{l_j}\| + \sum_{k=1}^{l_{j+1}-l_j-1}\|s_{l_j+k}\|\right]\\ &\leq (l_1-1)\overline{\Delta} + \sum_{j=1}^{\infty}\|s_{l_j}\|\left[1 + \frac{2}{\underline{cA}}\sum_{k=1}^{l_{j+1}-l_j-1}\overline{\beta}^{k-1}\right]\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{cA}}\frac{1}{1-\overline{\beta}}\right]\sum_{i\in N_3}\|s_i\|\\ &\leq (l_1-1)\overline{\Delta} + \left[1 + \frac{2}{\underline{cA}}\frac{1}{1-\overline{\beta}}\right]\frac{\|f_1\|}{2\underline{\rho\sigma A}} < \infty.\end{aligned}$$

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 91 *(superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem ($\overline{\text{T1}}$)–($\overline{\text{T3}}$) taková, že $x_i \to x^*$. Nechť funkce $f : R^n \to R^n$ splňuje podmínky (J3)-(J5). Nechť*

$$\lim_{i\to\infty}\overline{\omega}_i = 0 \qquad (\alpha)$$

a

$$\lim_{i\to\infty}\frac{\|(A_i - J_i)s_i\|}{\|s_i\|} = 0 \qquad (\beta).$$

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz (a) Ukážeme, že existuje index $k_2 \in N$ takový, že

$$-L_i(s_i) \geq \underline{\sigma J}\|s_i\|$$

a

$$\|f_i\| \geq \frac{1}{2}\underline{J}\|s_i\|$$

$\forall i \geq k_2$, pokud $\underline{J} < 1/\|(J^*)^{-1}\|$. 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 $\|\vartheta_i\| \to 0$, $J_i \to J^*$ a $\underline{J} < 1/\|(J^*)^{-1}\|$, existuje index $k_2 \in N$ takový, že $\|A_i s_i\| \geq \underline{J}\| \ s_i \ \| \ \forall i \geq k_2$. 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 $\|f_i\| \geq (\underline{J}/2)\|s_i\| \ \forall i \geq k_2$.
(b) Ukážeme, že existuje index $k_3 \geq k_2$ takový, že $i \in N_3 \ \forall i \geq k_3$. Použijeme-li větu o střední hodnotě 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 $\underline{\rho} < 1$, existuje index $k_3 \geq k_2$ takový, že $\rho_i(s_i) \geq \underline{\rho} \ \forall i \geq k_3$.
(c) Ukážeme, že existuje index $k \geq k_3$ takový, že $i \in N_1 \ \forall i \geq k$. Poznamenejme nejprve, že množina $N_1 \subset N$ je nekonečná. Kdyby tomu tak nebylo, muselo by platit $\|s_i\| \geq \Delta_i \geq \Delta_{k_3} \ \forall i \geq k_3$, neboť z (b) plyne $i \in N_3 \ \forall i \geq k_3$. 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 $i \geq k_3$, $i \in N_1$ a označme $\omega_i = (A_i s_i + f_i)/\|s_i\|$. Podle ($\alpha$), ($\beta$) a ($\overline{\text{T1b}}$) platí $\|\omega_i\| \to 0$ a $\|\vartheta_i\| \to 0$, takže stejným způsobem jako v důkazu věty 89 (s $\overline{\omega} = 0$) se dá ukázat, že existuje index $k_4 \geq k_3$, $k_4 \in N_1$ takový, že

$$\|f_i\|/\overline{J} \leq \|s_i\| \leq \|f_i\|/\underline{J}$$

$\forall i \geq k_4$, $i \in N_1$. Použijeme-li větu o střední hodnotě, 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

$$\lambda_i = \frac{f_{i+1} - f_i - A_i s_i}{\|f_i\|}.$$

Pak podle předchozích úvah platí $\|\lambda_i\| \leq \|\vartheta_i\|/\overline{J} + o(\|s_i\|)/\|s_i\| \to 0$. Jelikož zároveň $\|\omega_i\| \to 0$, existuje index $k \geq k_4$, $k \in N_1$ takový, že $\|\lambda_i\| < (\underline{J}/\overline{J})/2$ a $\|\omega_i\| < (\underline{J}/\overline{J})/2 \ \forall i \geq k$, $i \in N_1$. Pak můžeme psát

$$\begin{aligned}\|s_{i+1}\| &\leq \frac{1}{\underline{J}}\|f_{i+1}\| \leq \frac{1}{\underline{J}}\left(\|f_{i+1}-f_i-A_is_i\|+\|A_is_i+f_i\|\right) \leq \\ &\leq \frac{\overline{J}}{\underline{J}}\left(\|\lambda_i\|+\|\omega_i\|\right)\|s_i\| < \left(\frac{1}{2}+\frac{1}{2}\right)\|s_i\| = \|s_i\|.\end{aligned}$$

Jelikož $i \in N_3$ 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

$$\frac{\|f_{i+1}\|}{\|f_i\|} \leq \frac{\|f_{i+1}-f_i-A_is_i\|+\|A_is_i+g_i\|}{\|f_i\|} \leq \|\lambda_i\|+\|\omega_i\|,$$

což spolu s $\|\lambda_i\| \to 0$ a $\|\omega_i\| \to 0$ dává

$$\lim_{i\to\infty}\frac{\|x_{i+1}-x^*\|}{\|x_i-x^*\|} \leq \frac{\overline{J}}{\underline{J}}\frac{\|f_{i+1}\|}{\|f_i\|} = 0.$$

8.4 Newtonova metoda

Newtonova metoda používá matice $A_i = J(x_i)$ $\forall i \in N$, takže $\vartheta_i = (A_i - J_i)s_i/\|s_i\| = 0$ $\forall i \in N$ a z (J3)-(J4) plyne platnost podmínek (A3)-(A4).

Věta 92 *Nechť jsou splněny podmínky (J3)-(J5). 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 88 a věty 90. Superlineární konvergence plyne bezprostředně z věty 89 a věty 91, neboť $\vartheta_i = 0$ $\forall i \in N$.

Poznámka 142 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 27 *Nechť je splněn předpoklad (J5) a nechť platí*

$$Ae_j = \frac{f(x+\delta e_j)-f(x)}{\delta} \tag{D}$$

pro $1 \leq j \leq n$, kde e_j, $1 \leq j \leq n$, jsou sloupce jednotkové matice řádu n. Pak platí

$$\|A - J(x)\| \leq \frac{1}{2}\overline{L}\sqrt{n}\delta.$$

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

$$\begin{aligned}\|(A-J(x))e_j\| &= \left\|\frac{f(x+\delta e_j)-f(x)}{\delta}-J(x)e_j\right\| \le \frac{1}{\delta}\left\|\int_0^1 (J(x+\tau\delta e_j)-J(x))\delta e_j \mathrm{d}\tau\right\| \\ &\le \frac{1}{2\delta}\overline{L}\delta^2\|e_j\|^2 = \frac{1}{2}\overline{L}\delta.\end{aligned}$$

Nechť $s \in R^n$ je libovolný vektor s jednotkovou normou. Pak platí

$$\begin{aligned}\|(A-J(x))s\| &= \left\|\sum_{j=1}^n (A-J(x))e_j e_j^T s\right\| \le \sum_{j=1}^n |e_j^T s|\|(A-J(x))e_j\| \le \frac{1}{2}\overline{L}\delta\sum_{j=1}^n |e_j^T s| \\ &\le \frac{1}{2}\overline{L}\sqrt{n}\delta\|s\| = \frac{1}{2}\overline{L}\sqrt{n}\delta.\end{aligned}$$

a jelikož

$$\|A-J(x)\| = \max_{\|s\|=1}\|(A-J(x))s\|,$$

dostaneme tvrzení lemmatu.

Věta 93 *Nechť jsou splněny předpoklady (J3)-(J5). Je-li matice A určena podle vzorce (D), kde*

$$\delta \le \frac{2\gamma\underline{A}}{\overline{L}\sqrt{n}}, \quad \underline{A} = \underline{J}(\sqrt{\gamma^2\kappa^2+1}-\gamma\kappa)$$

a kde γ, κ jsou čísla použitá v lemmatu 23, pak platí $\|A-J(x)\| \le \overline{\vartheta}$, kde $\overline{\vartheta} \le \gamma\underline{A}$. Je-li matice A určena podle vzorce (D), kde

$$\delta \le \frac{(1-\overline{\omega})\underline{J}}{\overline{L}\sqrt{n}}, \quad 0 \le \overline{\omega} < 1,$$

pak $\|As+f\| \le \overline{\omega}$ implikuje $\|Js+f\| \le \overline{\omega}'$, kde $0 \le \overline{\omega}' < 1$.

Důkaz Z lemmatu 27 a z předpokladů věty 93 a z důkazu lemmatu 23 plyne, že

$$\|A-J(x)\| \le \frac{1}{2}\overline{L}\sqrt{n}\delta \le \overline{\vartheta} \triangleq \underline{J}\lambda(2\kappa) \le \gamma\underline{A}.$$

Poznámka 143 Věta 93 ukazuje, že lze zvolit diferenci $\delta > 0$ tak, aby matice určená podle vztahu (D) 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 (J4) a čím větší jsou čísla $\overline{J}$ a $\overline{L}$ v (J3) a (J5). Pro metodu spádových směrů ($\overline{\text{S1}}$)-($\overline{\text{S2}}$) musí platit $\gamma < (1-\underline{\rho})(1-\overline{\omega})/(1+\overline{\omega})$. Pro metodu s lokálně omezeným krokem ($\overline{\text{T1}}$)-($\overline{\text{T3}}$) musí platit $\gamma < (1/2-\underline{\rho})\underline{\sigma}$.

8.5 Kvazinewtonovské metody

Definice 47 *Řekneme, že základní metoda pro řešení systémů nelineárních rovnic je kvazinewtonovskou metodou, jestliže*

$$A_i s_i + f_i = 0 \tag{QN1}$$

kde A_i, $i \in N$, jsou regulární matice konstruované podle rekurentního vztahu

$$A_{i+1} = A_i + u_i v_i^T \tag{QN2}$$

kde $u_i \in R^n$, $v_i \in R^n$, a vyhovující podmínce

$$A_{i+1}d_i = y_i \tag{QN3}$$

kde $y_i = f_{i+1} - f_i$, $d_i = x_{i+1} - x_i$.

Poznámka 144 V tomto oddílu se budeme zabývat pouze přesnými kvazinewtonovskými metodami (podmínka (QN1)), takže $(A_i s_i + f_i)/\|f_i\| = 0 \ \forall i \in N$. Neplatí však $(A_i - J_i)s_i/\|f_i\| = 0 \ \forall i \in N$ (matice A_i se mohou od matic J_i dosti lišit).

Věta 94 *Nechť $A_+ = A + uv^T$ a $Ad \neq y$. Pak $A_+d = y$ právě tehdy, jestliže $v^Td \neq 0$ a $u = (y - Ad)/v^Td$, takže*

$$A_+ = A + \frac{(y - Ad)v^T}{v^Td} \tag{$\overline{\text{A}}$}$$

Jestliže $Ad = y$ stačí položit $u = v = 0$, takže $A_+ = A$.

Důkaz Z podmínky $A_+d = y$ dostaneme $A_+d = Ad + uv^Td = y$. Jestliže $Ad = y$, stačí položit $u = v = 0$, takže $A_+ = A$. Jestliže $Ad \neq y$, musí platit $v^Td \neq 0$ a $u = (y - Ad)/v^Td$.

Poznámka 145 Položíme-li $v = d$ dostaneme Broydenovu dobrou metodu

$$A_+ = A + \frac{(y - Ad)d^T}{d^Td} \tag{$\overline{\text{AG}}$}$$

Položíme-li $v = A^Ty$, dostaneme Broydenovu špatnou metodu

$$A_+ = A + \frac{(y - Ad)y^TA}{y^TAd} \tag{$\overline{\text{AB}}$}$$

Nechť

$$e_k^Td = \max_{1 \leq i \leq n} e_i^Td$$

Položíme-li $v = e_k$, dostaneme přímou metodu aktualizace sloupců

$$A_+ = A + \frac{(y - Ad)e_k^T}{e_k^Td} \tag{$\overline{\text{AD}}$}$$

která aktualizuje vždy pouze jeden sloupec matice A.

Věta 95 *Nechť A je regulární matice a nechť platí ($\overline{\text{A}}$). Pak matice A_+ je regulární právě tehdy, jestliže $v^TA^{-1}y \neq 0$.*

Důkaz Nechť $A_+ = A + uv^T$. Pak podle Shermanova-Morrisonova vzorce platí

$$A_+^{-1} = A^{-1} - \frac{A^{-1}uv^TA^{-1}}{1 + v^TA^{-1}u}$$

takže A_+ je regulární právě tehdy, jestliže $1 + v^TA^{-1}u \neq 0$. 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 146 Věta 95 opodstatňuje použití Broydenovy špatné metody. Jestliže $y \neq 0$ a matice A je regulární, pak volba $v = A^T y$ dává $v^T A^{-1} y = y^T A A^{-1} y = y^T y = \|y\|^2 \neq 0$.

Věta 96 *(Aktualizace matice $S = A^{-1}$). Nechť jsou splněny předpoklady věty 95. Nechť $S = A^{-1}$ a nechť A_+ je matice určená podle aktualizace* ($\overline{\text{A}}$)*, kde $v^T A^{-1} y \neq 0$. Nechť $S_+ = A_+^{-1}$. Pak platí*

$$S_+ = S + \frac{(d - Sy)v^T S}{v^T S y} \tag{$\overline{\text{S}}$}$$

Důkaz Podle Shermanova-Morrisonova vzorce (důkaz věty 95) platí

$$S_+ = S - \frac{Suv^T S}{\delta} = S + \frac{(d - Sy)v^T S}{\delta v^T d}$$

kde δ je zatím neznámé číslo. Z rovnice $S_+ y = d$ však plyne

$$S_+ y = Sy + \frac{v^T S y}{\delta v^T d}(d - Sy) = d$$

takže nutně $\delta = v^T S y / v^T d$.

Poznámka 147 Položíme-li $v = d$, dostaneme Broydenovu dobrou metodu

$$S_+ = S + \frac{(d - Sy)d^T S}{d^T S y} \tag{$\overline{\text{SG}}$}$$

Položíme-li $v = (S^{-1})^T y$, dostaneme Broydenovu špatnou metodu

$$S_+ = S + \frac{(d - Sy)y^T}{y^T y} \tag{$\overline{\text{SB}}$}$$

Nechť

$$e_k^T y = \max_{1 \leq i \leq n} e_i^T y$$

Položíme-li $S^T v = e_k$ dostaneme inverzní metodu aktualizace sloupců

$$S_+ = S + \frac{(d - Sy)e_k}{e_k^T y} \tag{$\overline{\text{SI}}$}$$

Poznámka 148 (Dualita). Vztah ($\overline{\text{S}}$) dostaneme ze vztahu ($\overline{\text{A}}$) 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 149 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.

Kvazinewtonovské metody splňují kvazinewtonovskou podmínku podobně jako metody s proměnnou metrikou (stačí porovnat (QN3) a (VM3)). Metody s proměnnou metrikou s přesným výběrem délky kroku nalezenou minimum kvadratické funkce (Q) po konečném počtu kroků. Ukážeme, že kvazinewtonovské metody s jednotkovým výběrem délky kroku ($\alpha_i = 1 \ \forall i \in N$) naleznou řešení soustavy lineárních rovnic

$$J^*(x - x^*) = 0 \tag{L}$$

s regulární maticí J^* také po konečném počtu kroků. Při důkazu tohoto tvrzení budeme používat vyjádření

$$x_{i+1} = x_i - S_i f_i \qquad (\alpha)$$

a

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} \qquad (\beta)$$

$\forall i \in N$, 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 28 *Uvažujme iterační proces* (α), (β) *aplikovaný na soustavu lineárních rovnic (L) 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 (L) a (α) plyne

$$y_i = f_{i+1} - f_i = J^* d_i = -J^* S_i f_i \qquad (\gamma)$$

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 (β) a (γ), 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 29 *Nechť jsou splněny předpoklady lemmatu 28 a nechť* $i \in N$ *je index takový, že vektor* f_{i+1} *není násobkem vektoru* f_i. *Pak vektory* $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, *jsou lineárně nezávislé pro každé číslo* $l \in N$ *takové, že* $2l \leq i+1$.

Důkaz (indukcí). Předpokládejme, že vektory $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, jsou lineárně nezávislé pro nějaké číslo $l \in N$ takové, že $2l \leq i-1$. Platí to zcela jistě pro $l = 1$, neboť podle (γ) 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 28 je vektor $(J^* S_{i-2l+2})^k f_{i-2l+2}$ lineární kombinací vektorů $(I - J^* S_{i-2l+1})(J^* S_{i-2l+1})^j$ f_{i-2l+1}, $0 \leq j \leq k$. Jelikož $l+1$ lineárně nezávislých vektorů $(J^* S_{i-2l+2})^k f_{i-2l+2}$, $0 \leq k \leq l$, vyjadřujeme pomocí $l+1$ vektorů $(I - J^* S_{i-2l+1})(J^* S_{i-2l+1})^k f_{i-2l+1}$, $0 \leq k \leq l$, musí být tyto vektory také lineárně nezávislé. Odtud bezprostředně plyne, že i vektory $(J^* S_{i-2l+1})^k f_{i-2l+1}$, $0 \leq k \leq l$, jsou lineárně nezávislé.
(b) Použijeme-li (γ), dostaneme

$$y_{i-2l} = -J^* S_{i-2l} f_{i-2l} \neq 0$$

Ukážeme, že vektor y_{i-2l} není lineární kombinací vektorů $(J^* S_{i-2l+1})^k f_{i-2l+1}$, $0 \leq k \leq l$. Použijeme-li kvazinewtonovskou podmínku

$$S_{i-2l+1} y_{i-2l} = d_{i-2l} = (J^*)^{-1} y_{i-2l}$$

můžeme psát

$$(I - J^* S_{i-2l+1}) y_{i-2l} = 0 \qquad (\delta)$$

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 (δ), což je spor s lineární nezávislostí těchto vektorů (viz (a)).
(c) Podle lemmatu 28 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 97 *Nechť jsou splněny předpoklady lemmatu 28. 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 (L).*

Důkaz Předpokládejme, že pro $i = 2n-1$ není vektor f_{i+1} násobkem vektoru f_i. Pak podle lemmatu 29 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 (β) a (γ) 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 (L).

Nevýhodou kvazinewtonovských metod 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 $A_{i+1} = J_{i+1}$. 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 $(\overline{\text{A}})$. Tyto úpravy mají své opodstatnění, neboť platí toto tvrzení.

Tvrzení 4 *Nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ taková, že pokud $\|x_1 - x^*\| \le \overline{\delta}$ a $\|A_1 - J_1\| \le \overline{\vartheta}$, posloupnost x_i, $i \in N$, určená dobrou Broydenovou metodou $(\overline{\text{AG}})$ s jednotkovým výběrem délky kroku $(\alpha_i = 1 \ \forall i \in N)$ konverguje Q-superlineárně k bodu $x^* \in R^n$.*

Tvrzení 4 je speciálním případem věty 99.

Následující tabulka ukazuje srovnání diferenční verze Newtonovy metody s dobrou Broydenovou metodou při minimalizaci 28 testovacích problémů s 2-16 neznámými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a Jacobiových matic NFJ, jakož i celkový čas výpočtu). Obě metody byly realizovány jako metody s lokálně omezeným krokem.

Metoda	NIT-NFV-NFJ	čas
Newtonova (diferenční verze)	504-5890-504	6.37
Broydenova (dobrá)	723-1844-93	2.75

9 Metody pro rozsáhlé řídké systémy nelineárních rovnic

Rozsáhlé řídké systémy nelineárních rovnic nemůžeme řešit metodami, které vyžadují uchovávání velkých hustých matic. Nejčastěji se pro tento účel používají některé speciální metody

- Kvazinewtonovské metody s omezenou pamětí
- Diferenční verze nepřesné Newtonovy metody
- Diferenční verze Newtonovy metody pro řídké úlohy
- Kvazinewtonovské metody pro řídké úlohy
- Metody používající některé speciální aktualizace

9.1 Kvazinewtonovské metody s omezenou pamětí

Kvazinewtonovské metody s omezenou pamětí jsou založeny na použití omezeného počtu kroků Broydenovy dobré metody nebo přímé metody aktualizace sloupců. Nechť $M = \{i \in N : \ i = (j-1)m+1, \ j \in N\}$, kde m je počet kroků kvazinewtonovské metody s omezenou pamětí. Pak pokládáme $S_l = (J_l^{-1})$ pro $l \in M$ a

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) v_i^T S_i}{v_i^T S_i y_i} = (I + w_i v_i^T) S_i$$

pro $l \leq i \leq l+m$ (viz ($\overline{\text{S}}$)), kde $v_i = d_i$ (Broydenova dobrá metoda) nebo $v_i = e_k$ (přímá metoda aktualizace sloupců) a

$$w_i = \frac{d_i - S_i y_i}{v_i^T S_i y_i}$$

vektory $v_i \in R^n$, $w_i \in R^n$, $l \leq i \leq l+m$, se uchovávají v paměti počítače.

Známe-li vektory $v_j \in R^n$, $w_j \in R^n$, $l \leq j \leq l+m$, určíme nejprve vektor $p_l^{i+1} = -S_l f_{i+1}$ (matice S_l je obvykle reprezentována trojúhelníkovým rozkladem $(S_l)^{-1} = L_l U_l$, který je úplným nebo neúplným trojúhelníkovým rozkladem matice J_l). Pak počítáme vektory

$$p_{j+1}^{i+1} = (I + w_j v_j^T) p_j^{i+1}$$

pro $l \leq j \leq i-1$. Nakonec určíme vektory v_i a

$$w_i = \frac{d_i - (p_i^{i+1} + s_i)}{v_i^T (p_i^{i+1} + s_i)}$$

kde $s_i = -S_i f_i$ je směrový vektor z předchozího iteračního kroku (obvykle $s_i = d_i/\alpha_i$) a položíme

$$s_{i+1} = -S_{i+1} f_{i+1} = -(I + w_i v_i^T) p_i^{i+1}$$

Kvazinewtonovské metody s omezenou pamětí můžeme také realizovat pomocí kompaktních schémat. Při odvozování kompaktních schémat budeme používat označení $D_k = [d_1, \ldots, d_k]$, $Y_k = [y_1, \ldots, y_k]$, $V_k = [v_1, \ldots, v_k]$. Dále označíme R_k horní trojúhelníkovou matici řádu k takovou, že $(R_k)_{ij} = v_i^T d_j$, $i \leq j$ a $(R_k)_{ij} = 0$, $i > j$. Abychom zjednodušili zápis budeme v důkazech index k vynechávat a index $k+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $D = [d_1, \ldots, d_{k-1}]$, $Y = [y_1, \ldots, y_{k-1}]$, $V = [v_1, \ldots, v_{k-1}]$ a $R = R_{k-1}$, takže $D_k = [D, d]$, $Y_k = [Y, y]$, $V_k = [V, v]$ a

$$R_k = \begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}$$

Věta 98 *Nechť A_1 je regulární matice a nechť platí* $(\overline{\mathrm{A}})$ *s $v_k^T d_k \neq 0$ pro libovolný index $1 \leq k \leq m$. Pak lze psát*

$$A_{k+1} = A_1 + (Y_k - A_1 D_k) R_k^{-1} V_k^T \qquad (\overline{\mathrm{AA}})$$

Důkaz Pro $k = 1$ je $(\overline{\mathrm{AA}})$ ekvivalentní s $(\overline{\mathrm{A}})$. Dále budeme postupovat matematickou indukcí. Předpokládejme, že $(\overline{\mathrm{AA}})$ platí pro všechny indexy menší než k. Pro index k můžeme $(\overline{\mathrm{AA}})$ 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

$$\begin{bmatrix} R, & V^T d \\ 0, & v^T d \end{bmatrix}^{-1} = \begin{bmatrix} R^{-1}, & -\frac{R^{-1} V^T d}{v^T d} \\ 0, & \frac{1}{v^T d} \end{bmatrix}$$

(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 $(\overline{A})$.

Poznámka 150 Přímou inverzí vztahu $(\overline{AA})$ (použitím Woodburyho věty), dostaneme

$$A_{k+1}^{-1} = A_1^{-1} - A_1^{-1}(Y_k - A_1 D_k)(R_k + V_k^T A_1^{-1}(Y_k - A_1 D_k))^{-1} V_k^T A_1^{-1}$$

neboli

$$S_{k+1} = S_1 + (D_k - S_1 Y_k)(C_k - L_k + V_k^T S_1 Y_k)^{-1} V_k^T S_1 \qquad (\overline{\mathrm{SS}})$$

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 $(C_k)_{ij} = 0$, $i \neq j$.

Kompaktní schémata používáme nejčastěji ve spojení s iteračním řešením soustavy rovnic $A_i s_i + f_i = 0$, $i \in N$. Pokládáme $A_l = J_l$ pro $l \in N$ a

$$A_{i+1} = A_l + (Y_k - A_l D_k) R_k^{-1} V_k^T$$

pro $l \leq i \leq l + m$ (viz $(\overline{\mathrm{AA}})$, kde $D_k = [d_l, \ldots, d_i]$, $Y_k = [y_l, \ldots, y_i]$, $V_k = [v_l, \ldots, v_i]$ a R_k je horní trojúhelníková matice řádu $k = i - l + 1$ taková, že $(R_k)_{ij} = v_{l+i-1}^T d_{l+j-1}$, $i \leq j$, a $(R_k)_{ij} = 0$, $i > j$. 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 $A_{i+1} p$ počítáme podle vzorce $A_{i+1} p = A_l p + U_k R_k^{-1} V_k^T p$.

9.2 Diferenční verze nepřesné Newtonovy metody

Diferenční verze nepřesné Newtonovy metody se vyznačují tím, že se systémy lineárních rovnic řeší nepřesně iteračními metodami. Nepoužívá se přitom matice $A = J$ a násobení $q = Ap = Jp$ se nahražuje numerickým derivováním

$$J(x)p \approx \frac{f(x + \delta p) - f(x)}{\delta}$$

kde δ je malá diference ($\delta = \sqrt{\varepsilon_M}/\|p\|$, kde ε_M je strojová přesnost). 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 9.7).

9.3 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 jsou založeny na aproximaci sloupců Je_j, $1 \leq j \leq n$, Jacobiovy matice J pomocí diferenčních vzorců

$$J(x)e_j \approx \frac{f(x + \delta e_j) - f(x)}{\delta}$$

kde δ je malá diference ($\varepsilon = \sqrt{\varepsilon_M}$). Je-li matice J řídká může nastat případ, kdy pomocí jedné diference vektorů funkčních hodnot určíme více sloupců této matice (podobně jako v oddílu 7.3). Rozdělme sloupce matice J do k disjunktních skupin $\mathcal{S}_i \subset \{1, \ldots, n\}$, $1 \leq i \leq k$, 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. Pak můžeme všechny sloupce matice J určit pomocí k diferencí

$$\frac{f(x + \delta v_i) - f(x)}{\delta} \approx Jv_i \qquad 1 \leq i \leq k$$

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 $J(\mathcal{S}_i)$). Získání rozkladu $\{1, \ldots, n\} = \mathcal{S}_i \cup \ldots \cup \mathcal{S}_k$, takového, aby počet skupin k byl minimální je složitý kombinatorický problém, jehož řešení se vymyká rozsahu tohoto textu.

Řádkové metody určují jednotlivé nenulové prvky Jacobiovy matice podle vzorců

$$(J(x))_{ij} \approx \frac{f_i(x + \delta e_j) - f_i(x)}{\delta}$$

Pro každý řádek $1 \leq i \leq n$, se počítají jen ty diference, které odpovídají nenulovým prvkům $(J(x))_{ij} \neq 0$. Numerickým porovnáním sloupcových a řádkových metod lze zjistit, že oba dva typy metod vyžadují přibližně stejný počet operací ne jednu iteraci. Sloupcové metody jsou algoritmicky náročnější (je třeba hledat rozklady sloupců Jacobiovy 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 řádkových metod.

Použití diferenčních verzí Newtonovy metody je podloženo teorií uvedenou v oddílu 8.3 (lemma 27).

9.4 Kvazinewtonovské metody pro řídké úlohy

Kvazinewtonovské metody pro řídké úlohy používají aktualizace, které zachovávají strukturu řídké Jacobiovy matice. Označme

$$\begin{aligned} \mathcal{V}_Q &= \{A \in R^{n \times n} : Ad = y\} \\ \mathcal{V}_G &= \{A \in R^{n \times n} : J_{ij} = 0 \Rightarrow A_{ij} = 0\} \end{aligned}$$

Podobně jako v oddílu 7.4 můžeme definovat operátory ortogonální projekce $\mathcal{P}_Q$, $\mathcal{P}_G$ do lineárních variet $\mathcal{V}_Q$, $\mathcal{V}_G$ předpisem

$$\begin{aligned}\mathcal{P}_Q A &= \min_{A_+ \in \mathcal{V}_Q} \|A_+ - A\|_F \\ \mathcal{P}_G A &= \min_{A_+ \in \mathcal{V}_G} \|A_+ - A\|_F\end{aligned}$$

Podobně můžeme definovat operátor ortogonální projekce $\mathcal{P}_{QG}$ do $\mathcal{V}_Q \cap \mathcal{V}_G$. Podle věty 81 platí

$$\mathcal{P}_{QG} A = \mathcal{P}_G(A + ud^T)$$

kde vektor $u \in R^n$ je řešením soustavy rovnic $Qu = y - Ad$ s diagonální pozitivně semidefinitní maticí

$$Q = \sum_{i=1}^{n} \|d^i\|^2 e_i e_i^T$$

kde d^i, $1 \leq i \leq n$, jsou vektory takové, že $d^i_j = d_j$, $J_{ij} \neq 0$ a $d^i_j = 0$, $J_{ij} = 0$. Označíme-li $A_+ = \mathcal{P}_{QG} A$, můžeme vzorec $\mathcal{P}_{QG} A = \mathcal{P}_G(A + ud^T)$ zapsat formálně ve tvaru

$$A_+ = A + \sum_{i=1}^{n} \frac{e_i^T (y - Ad) e_i (d^i)^T}{(d^i)^T d^i} \qquad (\overline{\text{AS}})$$

kde členy s $d^i = 0$ odpadnou. Metoda, která používá aktualizaci ($\overline{\text{AS}}$) 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 často nutné iterační proces přerušovat a pokládat $A_+ = J_+$. Je však možné dokázat, že Schubertova metoda konverguje lokálně Q-superlineárně.

Lemma 30 *Nechť A_+ je matice určená podle ($\overline{\text{AS}}$). Pak pro libovolnou matici $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_G$ platí*

$$\|A_+ - \tilde{J}\|_F^2 \leq \|A - \tilde{J}\|_F^2 - \frac{\|y - Ad\|^2}{\|d\|^2}$$

Důkaz Jelikož $\tilde{J} \in \mathcal{V}_Q \cap \mathcal{V}_G$, $\mathcal{P}_{QG}$ je operátor ortogonální projekce do $\mathcal{V}_Q \cap \mathcal{V}_G$ a $A_+ = \mathcal{P}_{QG} A$, můžeme použít Pythagorovu větu

$$\|A_+ - \tilde{J}\|_F^2 = \|A - \tilde{J}\|_F^2 - \|A_+ - A\|_F^2$$

Jelikož $\mathcal{V}_Q \cap \mathcal{V}_G \subset \mathcal{V}_Q$, platí $A_+ d = y$, takže

$$\|y - Ad\| = \|(A_+ - A)d\| \leq \|A_+ - A\| \|d\| \leq \|A_+ - A\|_F \|d\|$$

což po dosazení dává tvrzení lemmatu.

Lemma 31 *Nechť A_+ je matice určená podle ($\overline{\text{AS}}$) a nechť platí (J5). Pak*

$$\|A_+ - J_+\|_F \leq \|A - J\|_F + \overline{L}\sqrt{n}\|d\|$$

Důkaz Označme

$$\tilde{J} = \int_0^1 J(x + \lambda d) d\lambda$$

stejným způsobem jako v části (a) důkazu věty 83 (použitím věty o střední hodnotě) se ukáže, že platí

$$\begin{aligned}\|\tilde{J} - J\|_F &\leq \frac{1}{2}\overline{L}\sqrt{n}\|d\| \\ \|\tilde{J} - J_+\|_F &\leq \frac{1}{2}\overline{L}\sqrt{n}\|d\|\end{aligned}$$

Použijeme-li lemma 30, dostaneme

$$\begin{aligned}\|A_+ - J_+\|_F &\leq \|A_+ - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \leq \|A - \tilde{J}\|_F + \|\tilde{J} - J_+\|_F \leq \\ &\leq \|A - J\|_F + \|\tilde{J} - J\|_F + \|\tilde{J} - J_+\|_F\end{aligned}$$

což po dosazení dává tvrzení lemmatu

Věta 99 *Nechť platí (J5) a nechť $x^* \in R^n$ je bod takový, že $f(x^*) = 0$ a matice $J(x^*)$ je regulární. Pak existují čísla $\overline{\delta} > 0$, $\overline{\lambda} > 0$ taková, že pokud $\|x_1 - x^*\| \leq \overline{\delta}$, $\|A_1 - J_1\| \leq \overline{\lambda}$ a pokud platí*

$$\begin{aligned}\|A_i d_i + f_i\| &\leq \overline{\omega}\|f_i\| \\ x_{i+1} &= x_i + d_i \\ A_{i+1} &= \mathcal{P}_{QG} A_i\end{aligned}$$

$\forall i \in N$, kde $0 \leq \overline{\omega} < 1$ (nepřesná Schubertova metoda), posloupnost x_i, $i \in N$, konverguje k bodu $x^* \in R^n$. Jestliže navíc $\|\omega_i\| = \|A_i d_i + f_i\|/\|f_i\| \to 0$ pak $x_i \to x^*$ Q-superlineárně.

Důkaz Výsledky dosažené v částech (a) - (b) důkazu věty 89 můžeme přeformulovat (pomocí okolí) tak, že existují čísla $\delta > 0$, $\vartheta > 0$ taková, že pokud $\|x - x^*\| \leq \delta$, $\|(A - J(x))d\| \leq \vartheta\|d\|$ a $\|Ad + f\| \leq \overline{\omega}\|f\|$, kde $0 \leq \overline{\omega} < 1$, platí

$$\frac{1 - \overline{\omega}}{\overline{J}}\|f\| \leq \|d\| \leq \frac{1 + \overline{\omega}}{\underline{J}}\|f\|$$

kde $\|J^*\| < \overline{J}$ a $\|(J^*)^{-1}\| < 1/\underline{J}$ a

$$\|f(x + d)\| \leq r\|f\|$$

(kde $\overline{\omega} < r < 1$). Zdůrazněme, že číslo $0 \leq \overline{\omega} < 1$ může být libovolné zatímco čísla $\delta > 0$ a $\vartheta > 0$ mohou vycházet malá.
(a) Zvolme čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ tak, aby platilo

$$\overline{\delta}\left(1 + \frac{\overline{J}}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\right) \leq \delta$$

a

$$\overline{\vartheta} + \overline{L}\frac{\overline{J}}{\underline{J}}\frac{1 + \overline{\omega}}{1 - r}\overline{\delta} \leq \vartheta/\sqrt{n}$$

Nechť $\|x_1 - x^*\| \leq \overline{\delta}$ a $\|(A_1 - J(x_1)\| \leq \overline{\vartheta}$. Dokážeme indukcí, že pro libovolný index $i \in N$ platí $\|x_i - x^*\| \leq \delta$ a $\|A_i - J(x_i)\| \leq \vartheta$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme platnost tohoto tvrzení pro $1 \leq i \leq k$. Pak platí

$$\begin{aligned}\|x_{k+1}-x^*\| &\le \|x_1-x^*\|+\sum_{i=1}^{k}\|d_i\| \le \|x_1-x^*\|+\frac{1+\overline{\omega}}{\underline{J}}\sum_{i=1}^{k}\|f_i\| \le \\ &\le \|x_1-x^*\|+\frac{1+\overline{\omega}}{\underline{J}}\|f_1\|\sum_{i=1}^{k}r^{i-1} \le \|x_1-x^*\|+\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1-x^*\| \le \\ &\le \overline{\delta}\left(1+\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\right) \le \delta\end{aligned}$$

a použijeme-li lemma 30, dostaneme

$$\begin{aligned}\frac{\|(A_{k+1}-J_{k+1})d_{k+1}\|}{\|d_{k+1}\|} &\le \|A_{k+1}-J_{k+1}\| \le \|A_{k+1}-J_{k+1}\|_F \le \\ &\le \|A_1-J_1\|_F+\overline{L}\sqrt{n}\sum_{i=1}^{k}\|d_i\| \le \overline{\vartheta}\sqrt{n}+\overline{L}\sqrt{n}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\overline{\delta} \le \vartheta\end{aligned}$$

(b) Podle (a) platí $\|f_{i+1}\| \le r\|f_i\| \le r^i\|f_1\|\ \forall i \in N$, 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 $x_i \to x^*$.
(c) Podle lemmatu 30 platí

$$\begin{aligned}\frac{\|y-Ad\|^2}{\|d\|^2} &\le \|A-\tilde{J}\|_F^2-\|A_+-\tilde{J}\|_F^2 = \\ &= \left(\|A-\tilde{J}\|_F-\|A_+-\tilde{J}\|_F\right)\left(\|A-\tilde{J}\|_F+\|A_+-\tilde{J}\|_F\right) \le \\ &\le \overline{M}\left(\|A-\tilde{J}\|_F-\|A_+-\tilde{J}\|_F\right)\end{aligned}$$

Existence konstanty $\overline{M}$ plyne z toho, že

$$\begin{aligned}\|A-\tilde{J}\|_F+\|A_+-\tilde{J}\|_F &\le \|A-J\|_F+\|A_+-J_+\|_F+\overline{L}\sqrt{n}\|d\| \le \\ &\le 2\|A-J\|_F+2\overline{L}\sqrt{n}\|d\| \le \\ &\le 2\|A-J\|_F+2\overline{L}\sqrt{n}\left(\|x^+-x^*\|+\|x-x^*\|\right)\end{aligned}$$

takže podle (a) platí

$$\|A-\tilde{J}\|_F+\|A_+-\tilde{J}\|_F \le 2\sqrt{n}\vartheta+4\overline{L}\sqrt{n}\delta \triangleq \overline{M}$$

Dále lze psát

$$\|A_+-J_+\|_F \le \|A_+-\tilde{J}\|_F+\|J_+-\tilde{J}\|_F$$

takže

$$\begin{aligned}\|A-\tilde{J}\|_F-\|A_+-\tilde{J}\|_F &\le \|A-J\|_F+\|J-\tilde{J}\|_F-\|A_+-J_+\|_F+\|J_+-\tilde{J}\|_F \le \\ &\le \|A-J\|_F-\|A_+-J_+\|_F+\overline{L}\sqrt{n}\|d\|\end{aligned}$$

což dává

$$\begin{aligned}\sum_{i=1}^{\infty}\frac{\|y_i - A_i d_i\|^2}{\|d_i\|^2} &\leq \overline{M}\left(\|A_1 - J_1\|_F - \lim_{i\to\infty}\|A_{i+1} - J_{i+1}\|_F\right) + \overline{ML}\sqrt{n}\sum_{i=1}^{\infty}\|d_i\| \leq \\ &\leq \overline{M}\|A_1 - J_1\|_F + \overline{ML}\frac{\overline{J}}{\underline{J}}\frac{1+\overline{\omega}}{1-r}\|x_1 - x^*\| < \infty\end{aligned}$$

Platí tedy nutně $\|y_i - A_i d_i\|/\|d_i\| \to 0$, což spolu s $\|\omega_i\| = \|A_i d_i + f_i\|/\|f_i\| \to 0$ (stejně jako v důkazu věty 84) implikuje, že $x_i \to x^*$ Q-superlinárně.

9.5 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 si 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 7.3) 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

$$U_+ d = v \triangleq L^{-1}Py$$

Jelikož musí být zároveň zachována struktura horní trojúhelníkové matice, můžeme použít postup popsaný v oddílu 9.4. Výsledkem je aktualizace

$$U_+ = U + \sum_{i=1}^{n}\frac{e_i(v - Ud)e_i d^i}{(d^i)^T d^i} \qquad (\overline{\text{AD}})$$

kde d^i, $1 \leq i \leq n$, jsou vektory takové, že $d^i_j = d_j$, $U_{ij} \neq 0$ a $d^i_j = 0$, $U_{ij} = 0$ (členy s $d^i = 0$ odpadnou). Metoda, která používá aktualizaci ($\overline{\text{AD}}$) 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 ($\overline{\text{AD}}$). Po m aktualizacích ($\overline{\text{AD}}$) 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

$$e_i^T D_+ e_i = \frac{e_i^T Py}{e_i^T LUd} \qquad (\overline{\mathrm{AR}})$$

Také metodu škálování řádků je třeba po m iteračních krocích přerušovat s tím, že se provede nesymetrický trojúhelníkový rozklad $PJ = LU$.

9.6 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 9.3). 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 ($\overline{\mathrm{AS}}$).

9.7 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

- metody s krátkými rekurentními vztahy
- 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 48 *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 100 *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

$$(\alpha) \qquad \tilde{p}_j^T f_i = p_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1$$

$$(\beta) \qquad \tilde{f}_j^T f_i = f_j^T \tilde{f}_i = 0 \quad \forall 1 \le j < i \le n+1$$

$$(\gamma) \qquad \tilde{p}_j^T q_i = p_j^T \tilde{q}_i = 0 \quad \forall 1 \le j < i \le n$$

Z (β) 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 (γ) 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 $i = 1$ $(\alpha) - (\gamma)$ platí, neboť není co dokazovat.

(a) Nechť $i \le n$. Podle indukčních předpokladů (α) a (γ) 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 (α) a (γ) 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

$$\begin{aligned}
\tilde{p}_j^T q_{i+1} &= \tilde{p}_j^T A p_{i+1} = -\tilde{p}_j^T A f_{i+1} + \beta_i \tilde{p}_j^T A p_i = \\
&= -\left(\tilde{f}_{j+1} - \tilde{f}_j\right)^T f_{i+1}/\alpha_j + \beta_i \tilde{p}_j^T q_i = 0 \\
p_j^T \tilde{q}_{i+1} &= p_j^T A^T \tilde{p}_{i+1} = -p_j^T A^T \tilde{f}_{i+1} + \beta_i p_j^T A^T \tilde{p}_i = \\
&= -\left(f_{j+1} - f_j\right)^T \tilde{f}_{i+1}/\alpha_j + \beta_i p_j^T \tilde{q}_i = 0
\end{aligned}$$

$\forall 1 \leq j < i$. Použijeme-li navíc (b), dostaneme

$$\begin{aligned}
\tilde{p}_i^T q_{i+1} &= -\frac{1}{\alpha_i}\left(\tilde{f}_{i+1} - \tilde{f}_i\right)^T f_{i+1} + \beta_i \tilde{p}_i^T q_i = -\frac{\tilde{p}_i^T q_i}{\tilde{f}_i^T f_i} \tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} \tilde{p}_i^T q_i = 0 \\
p_i^T \tilde{q}_{i+1} &= -\frac{1}{\alpha_i}\left(f_{i+1} - f_i\right)^T \tilde{f}_{i+1} + \beta_i p_i^T \tilde{q}_i = -\frac{p_i^T \tilde{q}_i}{\tilde{f}_i^T f_i} \tilde{f}_{i+1}^T f_{i+1} + \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} p_i^T \tilde{q}_i = 0
\end{aligned}$$

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 151 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 32 *Nechť jsou splněny předpoklady věty 100. Pak vektory f_j, $1 \leq j \leq i \leq n$, (a také vektory p_j, $1 \leq j \leq i \leq n$) tvoří bázi v Krylovově podprostoru*

$$\mathcal{K}_i = \mathrm{span}\{f,\ Af, \ldots, A^{i-1}f\}$$

a vektory $\tilde{f}_j$, $1 \leq j \leq i \leq n$ (a také vektory $\tilde{p}_j$, $1 \leq j \leq i \leq n$) tvoří bázi v Krylovově podprostoru

$$\tilde{\mathcal{K}}_i = \mathrm{span}\{\tilde{f},\ (A^T)\tilde{f}, \ldots, (A^T)^{i-1}\tilde{f}\}$$

Důkaz (indukcí) pro $i = 1$ je tvrzení zřejmé. Předpokládejme, že tvrzení platí pro nějaký index $i < n$. Jelikož $f_i \in \mathcal{K}_i$ a $p_i \in \mathcal{K}_i$, dostaneme $f_{i+1} = f_i + \alpha_i A p_i \in \mathcal{K}_{i+1}$ a $p_{i+1} = -f_{i+1} + \beta_i p_i \in \mathcal{K}_{i+1}$, a jelikož vektory f_j, $1 \leq j \leq i+1$ (a také vektory p_j, $1 \leq j \leq i+1$) jsou lineárně nezávislé (důkaz věty 100), 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 100), tvoří tam bázi.

Poznámka 152 Nechť jsou splněny předpoklady věty 100. Pak platí

$$\begin{aligned} f_i = \varphi_i(A)f \quad &, \quad \tilde{f}_i = \varphi_i(A^T)\tilde{f} \\ p_i = -\psi_i(A)f \quad &, \quad \tilde{p}_i = -\psi_i(A^T)\tilde{f} \end{aligned}$$

$\forall 1 \leq i \leq n+1$, 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

$$\alpha_i = \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T A p_i} = \frac{\tilde{f}^T \varphi_i^2(A) f}{\tilde{f}^T A \psi_i^2(A) f} \quad \alpha_i = \frac{\tilde{f}_{i+1}^T f_{i+1}}{\tilde{f}_i^T f_i} = \frac{\tilde{f}^T \varphi_{i+1}^2(A) f}{\tilde{f}^T \varphi_i^2(A) f}$$

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 $\overline{s}_i \in R^n$, $1 \leq i \leq n+1$ tak, aby platilo $\overline{f}_i = A\overline{s}_i + f = \varphi_i^2(A)f$, $1 \leq i \leq n+1$.

Lemma 33 *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 49 *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

$$\begin{gathered}v_i = Ap_i, \quad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i \\ q_i = u_i - \alpha_i v_i \\ \overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i) \\ \overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i\end{gathered}$$

$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$
$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$

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 153 Jsou-li splněny předpoklady věty 100 platí

$$\|\overline{f}_i\| = \|\varphi_i^2(A)f\| \le \|\varphi_i(A)\|\|\varphi_i(A)f\| = \|\varphi_i(A)\|\|f_i\|$$

$1 \le i \le n+1$, takže metoda CGS najde řešení soustavy rovnic $As+f=0$ po nejvýše n krocích ($\|f_{n+1}\| = 0$ podle věty 100).

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 34 *Nechť* $\overline{f}_i$, $i \in N$, *je posloupnost reziduí určená metodou CGS. Nechť* $f_1 = \overline{f}_1$ *a*

$$\begin{aligned} \lambda_i &= -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2} \\ f_{i+1} &= \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) \end{aligned}$$

$1 \le i \le n$. *Pak platí*

$$\lambda_i = \arg\min_{\lambda\in R} \|\overline{f}_{i+1} + \lambda(f_i - \overline{f}_{i+1})\|$$

$1 \le i \le n$, *takže* $\|f_{i+1}\| \le \|f_i\|$ *(normy reziduí monotonně klesají) a* $\|f_{i+1}\| \le \|\overline{f}_{i+1}\|$ *(řešení je nalezeno po nejvýše n krocích).*

Důkaz Zřejmě pro $1 \le i \le n$ platí

$$\|f_{i+1}\|^2 = \|\overline{f}_{i+1}\|^2 + 2\lambda_i\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1}) + \lambda_i^2\|f_i - \overline{f}_{i+1}\|^2$$

tato kvadratická funkce nabývá minima pro $\lambda_i = -\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})/\|f_i - \overline{f}_{i+1}\|^2$.

Rekurentní vztahy pro f_i (lemma 34) spolu s odpovídajícími rekurentními vztahy pro s_i jsou základem jednoduše zhlazené metody CGS.

Definice 50 *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

$$v_i = Ap_i, \quad \alpha_i = \tilde{f}^T f_i/\tilde{f}^T v_i$$
$$q_i = u_i - \alpha_i v_i$$
$$\overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i)$$
$$\overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T f_{i+1}/\tilde{f}^T\overline{f}_i$$
$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$

$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$
$$\lambda_i = -\frac{\overline{f}_{i+1}^T(f_i - \overline{f}_{i+1})}{\|f_i - \overline{f}_{i+1}\|^2}$$
$$s_{i+1} = \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1})$$
$$f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1})$$

pro $1 \le i \le n$, *nazveme jednoduše zhlazenou metodou CGS (SSCGS) určenou maticí* $A \in R^{n\times n}$ *a vektory* $f \in R^n$, $\tilde{f} \in R^n$.

Ačkoliv normy reziduí jednoduše zhlazené metody CGS mají monotonní průběh, pro konstrukci metod s lokálně omezeným krokem je vhodnější dvojnásobně zhlazená metoda CGS.

Definice 51 *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

$$v_i = Ap_i, \quad \alpha_i = \tilde{f}^T\overline{f}_i/\overline{f}^T v_i$$
$$q_i = u_i - \alpha_i v_i$$
$$\overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i)$$
$$\overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \quad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i$$
$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$
$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$
$$[\lambda_i, \mu_i]^T = \arg\min_{[\lambda,\mu]^T\in R^2} \|\overline{f}_{i+1} + \lambda(f_i - \overline{f}_{i+1}) + \mu v_i\|$$
$$s_{i+1} = \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}) + \mu_i p_i$$
$$f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) + \mu_i v_i$$

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 154 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 34). 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

$$v_i = AC^{-1}p_i, \qquad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i$$
$$q_i = u_i - \alpha_i v_i$$
$$\overline{s}_{i+1} = \overline{s}_i + \alpha_i C^{-1}(u_i + q_i)$$
$$\overline{f}_{i+1} = \overline{f}_i + \alpha_i AC^{-1}(u_i + q_i), \qquad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i$$
$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$
$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$
$$[\lambda_i, \mu_i]^T = -(V_i^T V_i)^{-1}V_i^T\overline{f}_{i+1}$$
$$s_{i+1} = \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}) + \mu_i C^{-1}p_i$$
$$f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) + \mu_i v_i$$

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, $1 \le i \le n$, tak, že q_j $1 \le j \le i$, tvoří bázi v Krylovově podprostoru $\mathcal{K}_i$. Vektor $s_{i+1} \in R^n$ se volí tak, aby platilo

$$s_{i+1} = \arg\min_{s\in\mathcal{K}_i} \|As + f\| \tag{M}$$

Metoda GMRES je tedy založena na minimalizačním principu, což znamená, že normy reziduí monotonně klesají.

Ortonormální vektory q_i, $1 \le i \le n$ se generují pomocí Gramova-Schmidtova ortogonalizačního procesu. Klasický Gramův-Schmidtův ortogonalizační proces používá rekurentní vztahy

$$\beta_1 q_1 = f$$

a

$$q_{i+1}^1 = Aq_i$$
$$\left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^1 \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji}q_j \end{aligned}\right\} 1 \le j \le i$$
$$\beta_{i+1}q_{i+1} = q_{i+1}^{i+1}$$

$1 \le i \le n-1$, kde koeficienty β_i, $1 \le i \le n$ se vybírají tak, aby vektory q_i, $1 \le i \le n$, měly jednotkovou normu. Stabilnější je modifikovaný Gramův-Schmidtův ortogonalizační proces

$$\beta_1 q_1 = f$$

a

$$q_{i+1}^1 = Aq_i$$

$$\left.\begin{array}{l}\alpha_{ji} = q_j^T q_{i+1}^j \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j\end{array}\right\} 1 \le j \le i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1}$$

$1 \le i \le n-1$. Gramův-Schmidtův ortogonalizační proces generující ortonormální báze Krylovových podprostorů $\mathcal{K}_i$, $1 \le i \le n$, se také nazývá Arnoldiovým procesem určeným maticí $A \in R^{n\times n}$ a vektorem $f \in R^n$.

Označíme-li $Q_i = [q_1, q_2, \ldots, q_i]$ a

$$H_i = \begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1i} \\ \beta_2 & \alpha_{22} & \ldots & \alpha_{2i} \\ 0 & \beta_3 & \ldots & \alpha_{3i} \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & \beta_{i+1} \end{bmatrix}$$

($H_i \in R^{(i+1)\times i}$ je horní Hessenbergova matice), můžeme Arnoldiův proces zapsat v maticovém tvaru

$$AQ_i = Q_{i+1} H_i$$

Položíme-li $s_{i+1} = Q_i z_i$, kde $z_i \in R^n$, platí

$$\|As_{i+1} + f\| = \|AQ_i z_i + f\| = \|Q_{i+1} H_i z_i + Q_{i+1}(\beta_1 e_1)\| = \|H_i z_i + \beta_1 e_1\|$$

takže minimalizační podmínku můžeme zapsat ve tvaru

$$z_i = \arg\min_{z\in R^i} \|H_i z + \beta_1 e_1\| \qquad (\overline{\mathrm{M}})$$

Věta 101 *Nechť $\mathcal{K}_j \ne \mathcal{K}_{j+1}$, $1 \le j \le i$, $\mathcal{K}_i = \mathcal{K}_{i+1}$ a nechť platí (M). Pak $As_{i+1} + f = 0$.*

Důkaz Uvažujme Arnoldiův proces určený regulární maticí $A \in R^{n\times n}$ a vektorem f. Jestliže $\mathcal{K}_j \ne \mathcal{K}_{j+1}$, $1 \le j \le i$ a $\mathcal{K}_i = \mathcal{K}_{i+1}$, pak vektory q_i, $1 \le j \le i$, jsou lineárně nezávislé a $\beta_{i+1} = 0$. Platí tedy

$$AQ_i = Q_i \overline{H}_i$$

kde $\overline{H}_i \in R^{i\times i}$ je horní Hessenbergova matice, která vznikne z matice $H_i \in R^{(i+1)\times i}$ vyškrtnutím posledního řádku. Jelikož matice AQ_i má lineárně nezávislé sloupce a A je regulární), je matice $\overline{H}_i$ regulární a existuje řešení soustavy rovnic $\overline{H}_i z_i + \beta_1 e_1 = 0$. Položíme-li $s_{i+1} = Q_i z_i$ platí

$$\|As_{i+1} + f\| = \|\overline{H}_i z_i + \beta_1 e_1\| = 0$$

Důsledek Metoda GMRES nalezne řešení soustavy rovnic $As + f = 0$ po nejvýše n krocích. Jestliže totiž $\mathcal{K}_j \ne \mathcal{K}_{j+1}$, $1 \le j < n$, 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 $P_i = \overline{P}_i \overline{P}_{i-1} \ldots \overline{P}_1$ je součin Givensových matic elementárních rotací a

$$R_i = \begin{bmatrix} \rho_{11} & \rho_{12} & \ldots & \rho_{1i} \\ 0 & \rho_{22} & \ldots & \rho_{2i} \\ 0 & 0 & \ldots & \rho_{ii} \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \ldots \\ \eta_i \end{bmatrix}$$

Je to postup, který byl již použit v metodě LSQR (oddíl 7.8), proto ho nebudeme znovu odvozovat. Uvedeme pouze výsledné rekurentní vztahy metody GMRES.

Definice 52 *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

$$\begin{gathered}
q_{i+1}^1 = Aq_i \\
\overline{\alpha}_{1i} = q_1^T q_{i+1}^1, \qquad q_{i+1}^2 = q_{i+1}^1 - \overline{\alpha}_{1i} q_1 \\
\left.\begin{gathered}
\alpha_{ji} = q_j^T q_{i+1}^j, \qquad q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j \\
\rho_{j-1i} = \lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji} \\
\overline{\alpha}_{ji} = -\lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji}
\end{gathered}\right\} 1 < j \le i \\
\beta_{i+1} q_{i+1} = q_{i+1}^{i+1} \\
\rho_{ii} = \sqrt{\overline{\alpha}_{ii}^2 + \beta_{i+1}^2} \\
\lambda_i = \frac{\overline{\alpha}_{ii}}{\rho_{ii}}, \qquad \tau_i = \frac{\beta_{i+1}}{\rho_{ii}} \\
\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i
\end{gathered}$$

$1 \le i \le n$, 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 $|\overline{\eta}_i|$, $1 \le i \le n+1$, jsou tedy normy reziduí $f_i = As_i + f$, $1 \le i \le n+1$. Jakmile metoda GMRES získá dostatečně, malé rezidium $|(\overline{\eta}_{i+1}| \le \overline{\omega}\|f\|)$ 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

$$A\mathcal{K}_i = \mathrm{span}\{Af, \ldots, A^i f\}$$

odpadne použití ortogonálního rozkladu. Vektory q_j, $1 \le j \le i$ 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, $1 \le j \le i-1$, 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 53 *Nechť $A \in R^{n\times n}$ je regulární matice a $f \in R^n$. Pak iterační proces*

$$f_1 = f, \qquad \beta_1 q_1 = Af$$

a

$$\begin{gathered}
\gamma_i = q_i^T f_i \\
f_{i+1} = f_i - \gamma_i q_i \\
q_{i+1}^1 = Aq_i \\
\left.\begin{array}{l} \alpha_{ji} = q_j^T q_{i+1}^j \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j \end{array}\right\} 1 \le j \le i \\
\beta_{i+1} q_{i+1} = q_{i+1}^{i+1}
\end{gathered}$$

$1 \le i \le n-1$, *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 155 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 $C \approx A$.

9.8 Metody s lokálně omezeným krokem

Poznámka 156 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 46). 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 35 *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 36 *Nechť jsou splněny předpoklady lemmatu 35 a nechť $s \in \mathcal{R}^n$ je vektor určený metodou GMRES nebo dvojnásobně zhlazenou metodou CGS tak jako v poznámce 156. 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 35 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 35 spolu s odhadem $\underline{A}\|s\| \leq \|As\| \leq 2\|f\|$ dává

$$\|f\| - \|As + f\| \geq \tau_i\left(\|f\| - \|r_{i+1}\|\right) + \left(1 - \tau_i\left(\|f\| - \|r_i\|\right) \geq \frac{1}{2}\underline{\eta}^2\|f\| \geq \frac{1}{4}\underline{\eta}^2\|As\|\right.$$

(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 102 *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 156. Pak jsou splněny podmínky (T1a)-(T1c) a směrové vektory $s_i \in R^n$, $i \in N$, můžeme použít ke konstrukci nepřesné metody s lokálně omezeným krokem. Je-li tato metoda aplikována na funkci $f : R^n \to R^n$ vyhovující předpokladům (J3)-(J5) a splňují-li matice A_i, $i \in N$ podmínky (A5)-(A4), platí $x_i \to x^\star$ a $f(x^\star) = 0$.*

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 36 a věty 90.

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 $1 \le m \le 3$). Jestliže pro nějaký index $1 \le i \le m$ platí $\|s_{i+1}\| \le \Delta$ a $\|f_{i+1}\| \le \overline{\omega}\|f\|$, $0 \le \overline{\omega} < 1$, pokládáme $s = s_{i+1}$. Jestliže pro nějaký index $1 \le i \le m$ platí $\|s_j\| < \Delta$, $\|f_j\| > \overline{\omega}\|f\|$, $1 \le j \le i$, a $\|s_{i+1}\| \ge \Delta$, 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 36 nebo v důkazu lemmatu **??**), že pokud platí $\Delta \ge \underline{\gamma}\|f\|$ nebo $|f^T A f| \ge \underline{\varepsilon}\|f\|\|Af\|$, je splněna podmínka ($\overline{\text{T1c}}$).

Následující tabulka ukazuje srovnání několika realizací diferenční verze Newtonovy metody pro řídké úlohy (DSCGS značí dvojnásobně zhlazenou metodu CGS, GMRES(30) nebo GMRES(10) značí metodu GMRES restartovanou vždy po 30 nebo 10 krocích Arnoldiova procesu, S značí metodu spádových směrů, T značí metodu s lokálně omezeným krokem a LU značí předpodmiňování pomocí neúplného LU rozkladu) pro řešení 18 rozsáhlých řídkých systémů nelineárních rovnic se 100 neznámými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i celkový čas výpočtu).

Metoda	NIT-NFV	čas
Newtonova + DSCGS (S)	330-2010	8.07
Newtonova + DSCGS (S + LU)	235-1184	3.79
Newtonova + GMRES(30) (S)	346-2081	14.78
Newtonova + GMRES(10) (S + LU)	235-1184	3.85
Newtonova přímá (S + kompletní LU)	238-1200	5.21
Newtonova + DSCGS (T)	431-1851	8.96
Newtonova + DSCGS (T + LU)	234-1050	3.79
Newtonova + GMRES(30) (T)	337-1570	13.07
Newtonova + GMRES(10) (T + LU)	236-1061	3.95
Newtonova přímá (T + kompletní LU)	221- 975	5.00

V další tabulce je uvedeno srovnání několika metod (diferenční verze Newtonovy metody pro řídké úlohy, Schubertova kvazinewtonovská metoda, pětikroková Broydenova metoda s omezenou pamětí, diferenční verze nepřesné Newtonovy metody, Liova metoda se Schubertovou aktualizací) realizovaných s lokálně omezeným krokem s metodou DSCGS bez předpodmínění pro řešení 18 rozsáhlých řídkých systémů nelineárních rovnic se 100 neznámými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i celkový čas výpočtu).

Metoda	NIT-NFV	čas
Newtonova	389-1784	7.58
Schubertova	739-1288	10.98
5 - Broydenova	647-1322	12.03
Nepřesná Newtonova (diferenční verze)	517-6668	15.65
Liova s aktualizací	681-1592	11.09

10 Optimalizace dynamických systémů

Uvažujeme úlohu s účelovou funkcí

$$F(x) = \int_{t_0}^{t_1} f_A(y(x,t),t)\,dt + f_T(y(x,t_1)) \tag{O}$$

kde

$$\frac{dy(x,t)}{dt} = f_S(x,y(x,t),t), \qquad y(x,t_0) = f_I(x)\ . \tag{D}$$

Přitom $x \in R^n$, $y : R^n \times [t_0, t_1] \to R^{n_S}$, $F : R^n \to R$, $f_A : R^{n_S} \times [t_0, t_1] \to R$, $f_T : R^{n_S} \to R$, $f_S : R^n \times R^{n_S} \times [t_0, t_1] \to R^{n_S}$, $f_I : R^n \to R^{n_S}$. Odstranění integrálu:

$$F(x) = F_A(x, t_1) + f_T(y(x, t_1)) \qquad (\overline{\mathrm{O}})$$

kde

$$\frac{dy(x,t)}{dt} = f_S(x, y, t), \qquad y(x, t_0) = f_I(x)$$

$$\frac{dF_A(x,t)}{dt} = f_A(y, t), \qquad F_A(x, t_0) = 0 \qquad (\overline{\mathrm{D}})$$

Celkem se řeší $n_S + 1$ diferenciálních rovnic v přímém směru. Stačí spočítat hodnoty na konci intervalu. Úloha $(\overline{\mathrm{O}}) + (\overline{\mathrm{D}})$ se řeší pomocí gradientních optimalizačních metod (CG, VM, N) proto je třeba počítat derivace účelové funkce. Předpoklady:

(A1) Existuje spojité řešení systému (D) 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.

10.1 Přímý výpočet gradientu

Označme $u(x, t) = dy(x, t)/dx$, takže $u : R^n \times [t_0, t_1] \to R^{n_S \times n}$. Derivováním $(\overline{\mathrm{O}})$ a $(\overline{\mathrm{D}})$ dostaneme

$$g^T(x) = g_A^T(x, t_1) + \frac{\partial f_T(y(x, t_1))}{\partial y} u(x, t_1) \qquad (\overline{\mathrm{O1}})$$

kde

$$\frac{du(x,t)}{dt} = \frac{\partial f_S(x, y, t)}{\partial y} u(x, t) + \frac{\partial f_S(x, y, t)}{\partial x}, \qquad u(x, t_0) = \frac{df_I(x)}{dx}$$

$$\frac{dg_A^T(x,t)}{dt} = \frac{\partial f_A(y, t)}{\partial y} u(x, t), \qquad g_A^T(x, t_0) = 0 \qquad (\overline{\mathrm{D1}})$$

Přitom $g^T(x) = dF(x)/dx$, $g_A^T(x, t) = dF_A(x, t)/dx$. Celkem se řeší $(n_S + 1)(n + 1)$ diferenciálních rovnic v přímém směru.

10.2 Zpětný výpočet gradientů

Nechť $p(t)$ je libovolná funkce taková, že $p : [t_0, t_1] \to R^{n_S}$ a nechť $y(x, t)$ je řešení systému (D), takže $f_S(x, y, t) - dy(x, t)/dt = 0$ pro $t \in [t_0, t_1]$. Použijeme-li (O), můžeme psát

$$F(x) = \int_{t_0}^{t_1} \{f_A(y, t) + p^T(t)(f_S(x, y, t) - \frac{dy(x,t)}{dt})\}\, dt + f_T(y(x, t_1))$$

a použitím pravidla integrování per partes dostaneme

$$\begin{aligned} F(x) \quad = \quad & \int_{t_0}^{t_1} \{f_A(y, t) + p^T(t) f_S(x, y, t) + \frac{dp^T(t)}{dt} y(x, t)\}\, dt \\ & + p^T(t_0) y(x, t_0) - p^T(t_1) y(x, t_1) + f_T(y(x, t_1))\ . \end{aligned}$$

Nyní můžeme $F(x)$ derivovat podle x, takže

$$
\begin{aligned}
g^T(x) &= \int_{t_0}^{t_1} \left\{ \left[\frac{\partial f_A(y,t)}{\partial y} + p^T(t) \frac{\partial f_S(x,y,t)}{\partial y} + \frac{dp^T(t)}{dt} \right] \frac{dy(x,t)}{dx} \right. \\
&\quad \left. + p^T(t) \frac{\partial f_S(x,y,t)}{\partial x} \right\} dt \\
&\quad + p^T(t_0) \frac{df_I(x)}{dx} + \left[\frac{\partial f_T(y(x,t_1))}{\partial y} - p^T(t_1) \right] \frac{dy(x,t_1)}{dx} \; .
\end{aligned}
$$

Zvolíme-li funkci $p(t)$ tak, aby vypadly všechny členy s $dy(x,t)/dt$, čili tak, že

$$
-\frac{dp(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial y})^T p(x,t) + (\frac{\partial f_A(y,t)}{\partial y})^T, \qquad p(x,t_1) = (\frac{\partial f_T(y(x,t_1))}{\partial y})^T
$$

pak platí

$$
g^T(x) = \int_{t_0}^{t_1} p^T(x,t) \frac{\partial f_S(x,y,t)}{\partial x} \, dt + p^T(x,t_0) \frac{df_I(x)}{dx} \; .
$$

Dohromady to lze zapsat takto

$$
g(x) = \tilde{g}_A(x,t_0) + (\frac{df_I(x)}{dx})^T p(x,t_0) \tag{$\overline{\text{O2}}$}
$$

kde

$$
-\frac{dp(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial y})^T p(x,t) + (\frac{\partial f_A(y,t)}{\partial y})^T, \qquad p(x,t_1) = (\frac{\partial f_T(y(x,t_1))}{\partial y})^T
$$

a

$$
-\frac{d\tilde{g}_A(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial x})^T p(t), \qquad \tilde{g}_A(x,t_1) = 0 \tag{$\overline{\text{D2}}$}
$$

Celkem se řeší $n_S + 1$ diferenciálních rovnic v přímém směru a $2n_S + n$ diferenciálních rovnic ve zpětném směru.

10.3 Přímý výpočet Hessovy matice

Označme $v(x,t) = du(x,t)/dx = d^2y(x,t)/dx^2$, takže $v : R^n \times [t_0, t_1] \to R^{n_S \times n \times n}$. Derivováním ($\overline{\text{O1}}$) a ($\overline{\text{D1}}$) dostaneme

$$
G(x) = G_A(x,t_1) + u^T(x,t_1) \frac{\partial^2 f_T(y(x,t_1))}{\partial y^2} u(x,t_1) + \frac{\partial f_T(y(x,t_1))}{\partial y} v(x,t_1) \tag{$\overline{\text{O3}}$}
$$

kde

$$
\begin{aligned}
\frac{dv(x,t)}{dt} &= \frac{\partial f_S(x,y,t)}{\partial y} v(x,t) \\
&\quad + \left[\frac{\partial^2 f_S(x,y,t)}{\partial y^2} u(x,t) + \frac{\partial^2 f_S(x,y,t)}{\partial y \partial x} \right] u(x,t) \\
&\quad + \frac{\partial^2 f_S(x,y,t)}{\partial x \partial y} u(x,t) + \frac{\partial^2 f_S(x,y,t)}{\partial x^2},
\end{aligned}
$$

$$
v(x,t_0) = \frac{d^2 f_I(x)}{dx^2}
$$

$$\frac{dG_A(x,t)}{dt} = u^T(x,t)\frac{\partial^2 f_A(y,t)}{\partial y^2}u(x,t) + \frac{\partial f_A(y,t)}{\partial y}v(x,t), \qquad G_A(x,t_0) = 0 \tag{$\overline{\text{D3}}$}$$

Přitom $G(x) = d^2F(x)/dx^2$ a $G_A(x) = d^2 f_A(x,t)/dx^2$. Celkem se řeší $(n_S+1)(n^2+n+1)$ diferenciálních rovnic v přímém směru.

10.4 Přímá aproximace Hessovy matice (součet čtverců)

$$f_A(y,t) = \frac{1}{2}(y(x,t) - z(t))^T W(t)(y(x,t) - z(t))$$

$$\frac{\partial f_A(y,t)}{\partial y} = W(t)(y(x,t) - z(t)), \qquad \frac{\partial^2 f_A(y,t)}{\partial y^2} = W(t)$$

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 ($\overline{\text{O3}}$) a ($\overline{\text{D3}}$). Dostaneme tak

$$G(x) \approx B(x) = B_A(x,t_1) + u^T(x,t_1)W_1u(x,t_1) \tag{$\overline{\text{O4}}$}$$

kde

$$\frac{dB_A(x,t)}{dt} = u^T(x,t)W(t)u(x,t), \qquad B_A(x,t_0) = 0 \tag{$\overline{\text{D4}}$}$$

Celkem se řeší $(n_S+1)(n+1)+n^2$ diferenciálních rovnic v přímém směru.

11 Základy nehladké analýzy

11.1 Konvexní množiny

Definice 54 *Řekněme, že množina $C \in R^n$ je konvexní, jestliže z $x \in C$, $y \in C$ plyne*

$$\lambda x + (1-\lambda)y \in C, \tag{18}$$

pokud $0 \leq \lambda \leq 1$.

Poznámka 157 Vztah (18) můžeme zapsat ve tvaru

$$y + \lambda(x-y) \in C.$$

Definice 55 *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 103 *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 (18), 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 $x_1 \in C$ a $\lambda_1 = 1$). 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

$$x'_{m+1} = \sum_{i=1}^{m} \frac{\lambda_i}{\sum_{j=1}^{m} \lambda_j} x_i \stackrel{\Delta}{=} \sum_{i=1}^{m} \lambda'_i x_i \in C,$$

neboť $x_i \in C$, $\lambda'_i \geq 0$, $1 \leq i \leq m$, $\lambda'_1 + \ldots + \lambda'_m = 1$. □

Věta 104 *Průnik konvexních množin je konvexní množinou.*

Důkaz Nechť $C = \bigcap_i C_i$, kde $C_i \subset R^n$ jsou konvexní množiny. Nechť $x \in C$, $y \in C$. Pak platí $x \in C_i$ a $y \in C_i$ $\forall i$ a tedy $\lambda x + (1-\lambda)y \in C_i$ $\forall i$ pokud $0 \leq \lambda \leq 1$. Odtud plyne, že $\lambda x + (1-\lambda)y \in C$. □

Věta 105 *Lineární kombinace konvexních množin je konvexní množinou.*

Důkaz Nechť $C = \sum_i \lambda_i C_i$, kde $C_i \subset R^n$ jsou konvexní množiny. Nechť $x \in C$, $y \in C$. Pak pro $0 \leq \lambda \leq 1$ platí

$$\lambda x + (1-\lambda)y = \lambda \sum_i \lambda_i x_i + (1-\lambda) \sum_i \lambda_i y_i = \sum_i \lambda_i (\lambda x_i + (1-\lambda) y_i) \stackrel{\Delta}{=} \sum_i \lambda_i z_i.$$

Jelikož $x_i \in C_i$, $y_i \in C_i$, platí $z_i = \lambda x_i + (1-\lambda) y_i \in C_i$ takže $\lambda x + (1-\lambda)y \in C$. □

Definice 56 *Konvexním obalem množiny $C \subset R^n$ nazveme průnik*

$$\text{conv } C = \bigcap_{\alpha} C_\alpha$$

všech konvexních množin $C_\alpha \in R^n$ obsahujících C.

Poznámka 158 Zřejmě platí $C \subset \text{conv } C$.

Věta 106 *Konvexní obal množiny $C \subset R^n$ je množina všech konvexních kombinací bodů z C, tedy všech bodů tvaru*

$$y = \sum_{i=1}^{m} \lambda_i x_i, \tag{19}$$

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

$$y \in \text{conv } C = \bigcap_{\alpha} C_\alpha,$$

což dává $\tilde{C} \subset \text{conv } C$ □

Věta 107 *(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 (19) 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

$$\sum_{i=1}^{m} \alpha_i \hat{x}_i = 0. \tag{20}$$

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. Spojíme-li (19) a (20) 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 108 *Je-li množina C kompaktní, je i množina* conv C *kompaktní.*

Důkaz Nechť $\{x_i\} \subset \text{conv } C$ je posloupnost taková, že $x_i \to x \in R^n$. Máme dokázat, že $x \in \text{conv } C$. Jelikož $x_i \in \text{conv } C$, existují podle Věty 107 vektory $y_i^k \in C$ a čísla $\lambda_i^k \geq 0$, $1 \leq k \leq n+1$, $\lambda_i^1 + \ldots + \lambda_i^{n+1} = 1$ takové, že $x_i = \lambda_i^1 y_i^1 + \ldots + \lambda_i^{n+1} y_i^{n+1}$. Protože množina C je kompaktní a číslo $n+1$ je konečné, lze vybrat podposloupnost $\{\tilde{x}_i\} \subset \{x_i\}$ takovou, že odpovídající podposloupnosti $\{\tilde{y}_i^k\} \subset \{y_i^k\}$, $\{\tilde{\lambda}_i^k\} \subset \{\lambda_i^k\}$, $1 \leq k \leq n+1$, jsou konvergentní, čili $\tilde{y}_i^k \to \tilde{y}^k \in C$, $\tilde{\lambda}_i^k \to \tilde{\lambda}^k \geq 0$, $1 \leq k \leq n+1$, $\tilde{\lambda}^1 + \ldots + \tilde{\lambda}^{n+1} = 1$. Jelikož vybraná posloupnost má stejnou limitu jako původní konvergentní posloupnost, lze psát

$$\begin{aligned} x &= \lim_{i\to\infty} x_i = \lim_{i\to\infty} \tilde{x}_i = \lim_{i\to\infty}\left(\sum_{k=1}^{n+1} \tilde{\lambda}_i^k \tilde{y}_i^k\right) \\ &= \sum_{k=1}^{n+1}\left(\lim_{i\to\infty}\tilde{\lambda}_i^k\right)\left(\lim_{i\to\infty}\tilde{y}_i^k\right) = \sum_{k=1}^{n+1}\tilde{\lambda}^k\tilde{y}^k \in \text{conv } C. \end{aligned}$$

□

Definice 57 *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 159 Je-li množina $C \subset R^n$ uzavřená, platí

$$d_C(x) = \min_{y\in C} \|y - x\|.$$

Plyne to z toho, že pro libovolný bod $z \in C$ platí $d_C(x) = d_{C\cap B}(x)$, kde $B = \overline{B(x, \|z-x\|)}$, a konvexní množina $C \cap B$ je kompaktní, takže na ní nabývá vzdálenostní funkce svého minima. V dalším výkladu se omezíme na uzavřené množiny i když většina tvrzení má obecnější charakter.

Věta 109 *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 $x_1 \in R^n$, $x_2 \in R^n$. Podle Poznámky 159 existuje 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 $L = 1$ (Definice 68). Nechť nyní C je konvexní, $x_1 \in R^n$, $x_2 \in R^n$. Podle Poznámky 159 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 (Definice 65). Předpokládejme nyní, že C je konvexní 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

$$(y_2 - x)^T(y_1 - x) < d_C^2(x). \tag{21}$$

Položme nyní $y = \frac{1}{2}(y_2 + y_1)$. Jelikož C je konvexní, platí $y \in C$. Dále podle (21) 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 159 $d_C(x) \leq \|y - x\|$. □

Definice 58 *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 $y = P_C(x)$.*

Lemma 37 . *Nechť $C \subset R^n$ je uzavřená konvexní množina, $x \in R^n$ a $y = P_C(x)$. Pak platí*

$$(x - y)(z - y) \leq 0 \quad \forall z \in C$$

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 110 *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 37 platí

$$\begin{aligned}(x_1 - y_1)(z_1 - y_1) &\leq 0 \quad \forall z_1 \in C,\\ (x_2 - y_2)(z_2 - y_2) &\leq 0 \quad \forall z_2 \in C.\end{aligned}$$

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 59 *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.

Věta 111 *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 112 *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 $x \notin H(a, \alpha)$*

Důkaz Máme dokázat, že existuje vektor $a \in R^n$ a číslo $\alpha \in R$ tak, že

$$a^T x > \alpha \geq a^T y \quad \forall y \in C.$$

Podle Věty 107 existuje právě jeden vektor $\overline{y} \in C$ takový, že $\|\overline{y} - x\| = d_C(x)$. Položme $a = x - \overline{y}$ a $\alpha = a^T \overline{y}$. Pak platí

$$a^T x = (x - \overline{y})^T x = (x - \overline{y})^T (x - \overline{y}) + (x - \overline{y})^T \overline{y} = \|x - \overline{y}\|^2 + a^T \overline{y} > \alpha,$$

neboť $x \neq C$, takže $\|x - \overline{y}\| \neq 0$. Nerovnost $\alpha \geq a^T y \ \forall y \in C$ dokážeme sporem. Předpokládejme, že existuje bod $y \in C$ takový, že $\alpha = a^T \overline{y} < a^T y$, a označme $y(\lambda) = \overline{y} + \lambda b$, kde $b = y - \overline{y}$. Zřejmě $y(\lambda) \in C$, pokud $0 \leq \lambda \leq 1$ (Poznámka 157). Dále platí

$$\|y(\lambda) - x\|^2 = \|\lambda b - a\|^2 = \|a\|^2 - 2\lambda a^T b + \lambda^2 \|b\|^2$$

a

$$\left.\frac{d\|y(\lambda) - x\|^2}{d\lambda}\right|_{\lambda=0} = -2a^T b = -2(a^T y - a^T \overline{y}) < 0.$$

Tedy $\|y(0) - x\|^2 = \|a\|^2$ a existuje číslo $0 < \overline{\lambda} \leq 1$ takové, že $\|y(\lambda) - x\|^2 < \|a\|^2 = d_C^2(x) \ \forall 0 < \lambda \leq \overline{\lambda}$, což je ve sporu s definicí $d_C(x)$. □

Důsledek 12 *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 159, dvojice $x_1 \in C_1$, $x_2 \in C_2$ nemusí být určena jednoznačně). Protože $x_2 \notin C_1$ plyne z věty 112 (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 $a = a_1$, $\alpha = \alpha_1$). To však plyne z nekompatibility nerovností

$$a_1^T y = (x_2 - x_1)^T y \leq (x_2 - x_1)^T x_1 = \alpha_1,$$
$$a_2^T y = (x_1 - x_2)^T y \leq (x_1 - x_2)^T x_2 = \alpha_2,$$

jejichž sečtením dostaneme

$$0 \leq (x_2 - x_1)^T (x_1 - x_2) = -\|x_2 - x_1\|^2 < 0$$

(neboť $x_2 \neq x_1$). □

Věta 113 *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 111 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 22 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 $x \in C$ a $x \notin H$. □

Definice 60 *Nechť $C \subset R^n$. Pak funkci*

$$\delta_C(x) = \sup_{y \in C} y^T x$$

nazveme opěrnou funkcí množiny C.

Poznámka 160 Nechť množina $C \subset R^n$ je kompaktní. Pak platí

$$\delta_C(x) = \max_{y \in C} y^T x.$$

V dalším výkladu se omezíme na kompaktní množiny i když většina tvrzení má obecnější charakter.

Věta 114 *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 160 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 (Definice 68). □

Věta 115 *Nechť množina $C \subset R^n$ je kompaktní. Pak*

$$\delta_C(x) = \delta_{\mathrm{conv}\, C}(x) \quad \forall x \in R^n.$$

Důkaz Protože $C \subset \mathrm{conv}C$, platí podle Poznámky 160 $\delta_C(x) \leq \delta_{\mathrm{conv}\, C}(x) \ \forall x \in R^n$. Nechť $x \in R^n$. Podle Věty 107 lze každý vektor $y \in$ conv C vyjádřit jako konvexní kombinaci nejvýše $n+1$ vektorů $y_i \in C$, $1 \leq i \leq n+1$. Můžeme tedy psát

$$\begin{aligned} \delta_{\mathrm{conv}\, C}(x) &= \max_{y \in \mathrm{conv}\, C} y^T x = \max\left\{ \sum_{i=1}^{n+1} \lambda_i y_i^T x : y_i \in C, \lambda_i \geq 0, \sum_{i=1}^{n+1} \lambda_i = 1 \right\} \\ &\leq \max_{y \in C} y^T x = \delta_C(x). \end{aligned}$$ □

Věta 116 *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 160 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 112 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 13 *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 117 *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 118 *Množina $C \subset R^n$ leží v poloprostoru $H(a,\alpha)$ právě tehdy jestliže $\alpha \geq \delta_C(a)$*

Důkaz Tvrzení plyne z definice 60 a z toho, že $C \subset H(a,\alpha)$ právě tehdy, jestliže $\delta_C(a) = \sup_{y\in C} a^T y \leq \alpha$. □

Definice 61 *Řekneme, že množina $K \subset R^n$ je kuželem, jestliže z $x \in K$ a $\lambda \geq 0$ plyne $\lambda x \in K$.*

Věta 119 *Průnik kuželů je kuželem.*

Důkaz Nechť $K = \bigcap_i K_i$, kde $K_i \subset R^n$ jsou kužely. Nechť $x \in K$ a $\lambda \geq 0$. Pak platí $x \in K_i$ a tedy $\lambda x \in K_i\ \forall i$. Odtud plyne, že $\lambda x \in K$. □

Věta 120 *Lineární kombinace kuželů je kuželem.*

Důkaz Nechť $K = \sum_i \lambda_i K_i$, kde $K_i \subset R^n$ jsou kužely. Nechť $x \in K$ a $\lambda \geq 0$. Pak platí

$$\lambda x = \lambda \sum_i \lambda_i x_i \triangleq \sum_i \lambda_i z_i.$$

Jelikož $x_i \in K_i$ a $\lambda \geq 0$, platí $z_i = \lambda x_i \in K_i$, takže $\lambda x \in K$. □

Věta 121 *Nechť $C \subset R^n$. Označme*

$$\bigcup_{\lambda\geq 0} \lambda C = \{x \in R^n :\ x = \lambda y,\ \ y \in C,\ \ \lambda \geq 0\}$$

Pak platí

$$\bigcup_{\lambda\geq 0} \lambda C = \bigcap_\alpha K_\alpha,$$

kde $\bigcap_\alpha K_\alpha$ je průnik všech kuželů $K_\alpha \subset R^n$ obsahujících množinu C.

Důkaz Nechť $\tilde{K} = \bigcup_{\lambda\leq 0} \lambda C$. Jelikož $\tilde{K}$ je kužel obsahující množinu C, platí $\bigcap_\alpha K_\alpha \subset \tilde{K}$. Nechť naopak $y \in \tilde{K}$, takže $y = \lambda x$, kde $\lambda \geq 0$ a $x \in C$. Jelikož $x \in C \subset K_\alpha$ a $\lambda \geq 0$, platí $y \in K_\alpha$ pro libovolný kužel K_α a tedy $y \in \bigcap_\alpha K_\alpha$, což dává $\tilde{K} \subset \bigcap_\alpha K_\alpha$. □

Věta 122 *Množina $K \subset R^n$ je konvexním kuželem právě tehdy, obsahuje-li všechny kladné lineární kombinace svých bodů.*

Důkaz Obsahuje-li množina K všechny kladné lineární kombinace svých bodů, obsahuje též konvexní kombinace tvaru (18) a kladné násobky svých bodů, takže je konvexním kuželem. Nechť $x_i \in K$, $\lambda_i \geq 0$, $1 \leq i \leq m$. 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

$$x = \sum_{i=1}^{m} \lambda_i x_i = \lambda x' \in K.$$

□

Definice 62 *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.

Věta 123 *Nechť $C \in R^n$. 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 $\lambda x \in C^*$. □

Definice 63 *Nechť $C \subset R^n$ je uzavřená konvexní 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 124 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak $T_C(x)$ je uzavřeným konvexním kuželem.*

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 $x + t_i^k y_k^i \in C \; \forall i, k \in N$. Pro každé $k \in N$ tedy existuje index $\overline{i}_k \in N$ takový, že

$$\|y_i^k - y^k\| < \varepsilon/2 \text{ a } t_i^k < 1/k, \quad \forall i \geq \overline{i}_k.$$

Zkonstruujeme-li posloupnost indexů $\{i_k\} \subset N$ rekurentním předpisem $i_1 = \overline{i}_1$ a $i_{k+1} = \max(i_k + 1, \overline{i}_{k+1})$, platí

$$\|y_{i_k}^k - y^k\| < \varepsilon/2 \text{ a } t_{i_k}^k < 1/k$$

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 $y^1 \in T_C(x)$ a $y^2 \in T_C(x)$. Podle Definice 63 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 $x + t_i^1 y_i^1 \in C$, $x + t_i^2 y_i^2 \in C$. Jelikož C je konvexní, podle poznámky 157 platí $x + t_i y_i^1 \in C$, $x + t_i y_i^2 \in C$, 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

$$x + y_i t_i = \lambda(x + y_i^1) + (1-\lambda)(x + y_i^2) \in C,$$

takže $y \in T_C(x)$ a množina $T_C(x)$ je konvexní.
(c) Nechť $y \in T_C(x)$ a $\lambda \geq 0$. Podle Definice 63 existují posloupnosti $y_i \to y$, $t_i \downarrow 0$ takové, že $x + t_i y_i \in C$. Pak ale $\lambda y_i \to \lambda y$, $t_i/\lambda \downarrow 0$ 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. □

Věta 125 *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 (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 $\lambda t_i = t_i' \leq 1$. 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 157, takže $y \in T_C(x)$. Platí tedy $\bigcup_{\lambda \geq 0} \lambda(C - x) \subset T_C(x)$ a jelikož $T_C(x)$ je uzavřená množina, též

$$\overline{\bigcup_{\lambda \geq 0} \lambda(C - x)} \subset T_C(x)$$

.
(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 $y_i \in \bigcup_{\lambda \geq 0} \lambda(C - x)$. Jelikož $y_i \to y$, platí $y \in \overline{\bigcup_{\lambda \geq 0} \lambda(C - x)}$, takže

$$T_C(x) \subset \overline{\bigcup_{\lambda \geq 0} \lambda(C - x)}.$$

□

Definice 64 *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 161 Podle Věty 123 je množina $N_C(x)$ uzavřeným konvexním kuželem.

Věta 126 *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 62 a 64, druhá z Věty 125, třetí z invariance neostré nerovnosti vůči limitnímu přechodu a poslední z invariance neostré nerovnosti vůči násobení skalárem λ. □

11.2 Konvexní funkce

Definice 65 *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*

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2), \tag{22}$$

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 (22) pokud $x_1 \in C$, $x_2 \in C$ a $0 \leq \lambda \leq 1$.

Poznámka 162 Nerovnost (22) 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 163 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 103).

Věta 127 *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 vyjádřit jako konvexní kombinaci vrcholů $y^{(k)}$, $1 \le k \le 2^n$, 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 $\lambda_k \ge 0$, $1 \le k \le 2^n$, takže

$$f(y) = f\left(\sum_{k=1}^{2^n} \lambda_k y^{(k)}\right) \le \sum_{k=1}^{2^n} \lambda_k f(y^{(k)}) \le M \sum_{k=1}^{2^n} \lambda_k = M.$$

Funkce f je tedy omezená shora na $B(x,\varepsilon)$. 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

$$f(x) = f\left(\frac{y+y'}{2}\right) \le \frac{1}{2}(f(y) + f(y')),$$

takže

$$f(y) \ge 2f(x) - f(y') \ge 2f(x) - M$$

a funkce f je omezená zdola na $B(x,\varepsilon)$. Položme $\delta = \varepsilon/2$ a $m = 2f(x) - M$. Nechť $z \in B(x,\delta)$, $z' \in B(x,\delta)$ a $z \ne z'$. 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

$$\begin{aligned} f(z') - f(z) &\le \frac{\|z' - z\|}{\delta + \|z' - z\|} f(z'') + \frac{\delta}{\delta + \|z' - z\|} f(z) - f(z) \\ &= \frac{\|z' - z\|}{\delta + \|z' - z\|} (f(z'') - f(z)) \le \frac{1}{\delta} \|z' - z\| (M - m). \end{aligned}$$

Jelikož nezáleží na pořadí bodů z a z', dostaneme

$$|f(z') - f(z)| \le \frac{M - m}{\delta} \|z' - z\|,$$

takže f je lipschitzovská s konstantou $L = (M - m)/\delta$ na $B(x,\delta)$. □

Lemma 38 *Nechť funkce $\varphi : R \to R$ je konvexní na intervalu $[a,b]$ a nechť $a \le t_1 < t_2 < t_3 \le b$. Pak platí*

$$\frac{\varphi(t_2) - \varphi(t_1)}{t_2 - t_1} \le \frac{\varphi(t_3) - \varphi(t_1)}{t_3 - t_1} \le \frac{\varphi(t_3) - \varphi(t_2)}{t_3 - t_2}.$$

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 f (Poznámka 162) 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

$$t_2 = t_3 + \frac{t_2 - t_3}{t_1 - t_3}(t_1 - t_3).$$

Z konvexity funkce f pak plyne

$$\varphi(t_2) \leq \varphi(t_3) + \frac{t_2 - t_3}{t_1 - t_3}(\varphi(t_1) - \varphi(t_3)).$$

□

Definice 66 *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*

$$f'(x, h) = \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t}. \tag{23}$$

Věta 128 *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

$$\limsup_{i \to \infty} f'(x_i, h_i) \leq f'(x, h),$$

kdykoliv $x_i \to x$ a $h_i \to h$.

Důkaz (a) Nechť funkce f je konvexní v $B(x, \varepsilon)$. Podle Lemmatu 38 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 $(f(x + th) - f(x))/t \geq (f(x) - f(x - t'h))/t'$ pro libovolné $0 < t'h < \varepsilon$, přičemž výraz na levé straně poslední nerovnosti je konečný, neboť funkce f je spojitá). Existuje tedy limita (23). 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

$$\begin{aligned} f'(x, h_1 + h_2) &= \lim_{t \downarrow 0} \frac{f(x + t(h_1 + h_2)) - f(x)}{t} \\ &= \lim_{t \downarrow 0} \frac{f\left(\frac{1}{2}(x + 2th_1) + \frac{1}{2}(x + 2th_2)\right) - f(x)}{t} \\ &\leq \lim_{t \downarrow 0} \frac{f(x + 2th_1) - f(x)}{2t} + \lim_{t \downarrow 0} \frac{f(x + 2th_2) - f(x)}{2t} \\ &= f'(x, h_1) + f'(x, h_2), \end{aligned}$$

takže $f'(x, \cdot)$ je subaditivní. Dále platí

$$f(x + th_2) - f(x + th_1) \leq Lt\|h_2 - h_1\|$$

pro $t > 0$. Můžeme tedy psát

$$\lim_{t \downarrow 0} \frac{f(x + th_2) - f(x)}{t} \leq \lim_{t \downarrow 0} \frac{f(x + th_1) - f(x)}{t} + L\|h_2 - h_1\|,$$

takže

$$f'(x, h_2) - f'(x, h_1) \leq L\|h_2 - h_1\|.$$

Protože nezáleží na pořadí vektorů h_1 a h_2, platí

$$|f'(x, h_2) - f'(x, h_1)| \leq L\|h_2 - h_1\|,$$

což dokazuje lipschitzovskost $f(x, \cdot)$.
(c) Nechť $x_i \to x$ a $h_i \to h$. Položme $t_i = \sqrt{\|x_i - x\|} + 1/i$ 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

$$f'(x_i, h_i) \leq \frac{f(x_i + t_i h_i) - f(x_i)}{t_i} = \frac{f(x + t_i h) - f(x)}{t_i} + \frac{f(x_i + t_i h_i) - f(x + t_i h)}{t_i} + \frac{f(x) - f(x_i)}{t_i}.$$

Ale

$$\frac{|f(x_i + t_i h_i) - f(x + t_i h)|}{t_i} \leq \frac{L(\|x_i - x\| + t_i\|h_i - h\|)}{t_i} \leq L(\sqrt{\|x_i - x\|} + \|h_i - h\|) \to 0$$

a

$$\frac{|f(x_i) - f(x)|}{t_i} \leq \frac{L\|x_i - x\|}{t_i} \leq L\sqrt{\|x_i - x\|} \to 0.$$

Můžeme tedy psát

$$\limsup_{i \to \infty} f'(x_i, h_i) \leq \limsup_{i \to \infty} \frac{f(x + t_i h) - f(x)}{t_i} = \lim_{t \downarrow 0} \frac{f(x + th) - f(x)}{t} = f'(x, h)$$

□

Poznámka 164 Podle Definice 66 platí $f'(x,0)=0$, takže podle Věty 128 (b) dostaneme

$$|f'(x,h)| = |f'(x,h) - f'(x,0)| \leq L\|h\|.$$

Definice 67 *Nechť funkce $f : R^n \to R$ je konvexní v okolí bodu $x \in R^n$. Pak množinu*

$$\partial f(x) = \left\{ g \in R^n : f'(x,h) \geq g^T h \ \ \forall h \in R^n \right\}$$

nazveme subdiferenciálem funkce f v bodě x. Elementy $g \in \partial f(x)$ budeme nazývat subgradienty funkce f v bodě x.

Věta 129 *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 $\|g\| \leq L \ \forall g \in \partial f(x)$.

(b) Pro libovolný vektor $h \in R^n$ platí

$$f'(x,h) = \max\left\{ g^T h : g \in \partial f(x) \right\}.$$

(c) Jestliže $x_i \to x$, $g_i \in \partial f(x_i)$ a $g_i \to g$, pak $g \in \partial f(x)$ (polospojitost shora).

(d) Existuje číslo $\varepsilon > 0$ takové, že pro libovolný vektor $g \in \partial f(x)$ platí

$$f(x+h) - f(x) \geq g^T h \quad \forall h \in B(0,\varepsilon).$$

Důkaz (a) Podle Věty 128 (b) je funkce $f'(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (Věta **??**) 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 67 a Poznámky 164 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 67 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

$$f'(x,\overline{h}) > \max\left\{ g^T \overline{h} : g \in \partial f(x) \right\}. \tag{24}$$

Uvažujme lineární funkci $l(\lambda\overline{h}) \triangleq \lambda f'(x,\overline{h})$ 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 (24).
(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 128 (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 67 a Věty 128 (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 165 Porovnáme-li Větu 129 (b) s Poznámkou 160, vidíme, že směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$f'(x,h) = \delta_{\partial f(x)}(h).$$

Věta 130 *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 67 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 131 *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 128 (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 67 tedy platí $0 \in \partial f(x)$. Jestliže $0 \in \partial f(x)$, existuje podle Věty 129 (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 132 *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

$$\min_{z\in C} \|x+ty_i-z\| = \min_{z\in C} \| \left(1 - \frac{t}{t_i}\right) x + \frac{t}{t_i}(x+t_i y_i) - z\| = 0,$$

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 $d'_C(x,y) \leq 0$, čili $y \in K$. Odtud plyne $T_C(x) \subset K$.
(b) Nechť $y \in K$ a $t_i \downarrow 0$. Z definice množiny K plyne, že

$$d'_C(x,y) = \lim_{i\to\infty} \frac{d_C(x+t_i y)}{t_i} = 0.$$

Nechť body $z_i \in C$, $i \in N$, jsou zvoleny tak, že

$$\|x + t_i y - z_i\| \leq d_C(x+t_i y) + \frac{t_i}{i}$$

(což je možné vzhledem k definici vzdálenosti $d_C(x+t_i y)$). Položme $y_i = (z_i - x)/t_i$, $i \in N$. Pak platí

$$x + t_i y_i = x + \frac{z_i - x}{t_i} = z_i \in C$$

a

$$\|y - y_i\| = \|y - \frac{z_i - x}{t_i}\| = \frac{1}{t_i}\|x + t_i y - z_i\| \leq \frac{d_C(x+t_i y)}{t_i} + \frac{1}{i},$$

takže

$$\lim_{i\to\infty} \|y - y_i\| = d'_C(x,y) + \lim_{i\to\infty} \frac{1}{i} = 0.$$

Tím jsme dokázali, že $y \in T_C(x)$. Odtud plyne, že $K \subset T_C(x)$. □

Věta 133 *Nechť $C \subset R^n$ je uzavřená konvexní množina a $x \in C$. Pak platí*

$$N_C(x) = \overline{\bigcup_{\lambda\geq 0} \lambda \partial d_C(x)}$$

Důkaz (a) Předpokládejme, že $z \in \partial d_C(x)$. Pak podle Definice 67 platí

$$d'_C(x,y) \geq z^T y \quad \forall y \in R^n.$$

Jestliže $y \in T_C(x)$, platí podle věty 132 $d'_C(x,y) = 0$, takže

$$z^T y \leq 0 \quad \forall y \in T_C(x),$$

což podle definic 62 a 64 dává $z \in N_C(x)$. Jelikož $N_C(x)$ je uzavřený konvexní kužel, platí

$$\overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)} \subset N_C(x)$$

(b) Nechť $z \in N_C(x)$. Pak podle definic 62, 64 a věty 132 platí

$$z^T y \leq 0 = d'_C(x,y) = \lambda(y) d'_C(x,y) \quad \forall y \in T_C(x),$$

kde $\lambda(y) = 1 \ \forall y \in T_C(x)$. Zbývá dokázat podobnou nerovnost i pro $y \notin T_C(x)$ (kde obecně $\lambda(y) \neq 1$). Nechť $y \notin T_C(x)$. Jelikož $d'_C(x,y) > 0$ pro $y \notin T_C(x)$ ($d'_C(x,y) \geq 0$ a $d'_C(x,y) \neq 0$ pro $y \notin T_C(x)$ podle věty 132), platí

$$\lambda(y) \triangleq \frac{\|z\|\|y\|}{d'_C(x,y)} \geq 0.$$

Použitím Schwarzovy nerovnosti dostaneme

$$z^T y \leq \|z\|\|y\| = \lambda(y) d'_C(x,y).$$

Dokázali jsme tedy, že pro libovolný vektor $y \in R^n$ existuje $\lambda(y) \geq 0$ tak, že $z^T y \leq \lambda(y) d'_C(x,y)$. Odtud plyne, že $z \in \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)}$, takže

$$N_C(x) \subset \overline{\bigcup_{\lambda \geq 0} \lambda \partial d_C(x)}$$

. □

11.3 Lipschitzovské funkce

Definice 68 *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*

$$|f(x_2) - f(x_1)| \leq L\|x_2 - x_1\|, \tag{25}$$

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 oblasti Ω, je-li lipschitzovská v okolí každého bodu $x \in \Omega$.

Definice 69 *Zobecněnou (Clarkovu) směrovou derivaci funkce $f : R^n \to R$ v bodě $x \in R^n$ ve směru $h \in R^n$ definujeme předpisem*

$$f^0(x,h) = \limsup_{\substack{y \to x \\ t \downarrow 0}} \frac{f(y+th) - f(y)}{t}. \tag{26}$$

Poznámka 166 Je-li $f^0(x,h)$ zobecněnou směrovou derivací funkce f ve smyslu Definice 69, 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 134 *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

$$\limsup_{i\to\infty} f^0(x_i, h_i) \le f^0(x, h),$$

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

$$\begin{aligned}
f^0(x, h_1 + h_2) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y + t(h_1 + h_2)) - f(y)}{t} \\
&= \limsup_{\substack{y\to x\\ t\downarrow 0}} \left(\frac{f(y + t(h_1 + h_2)) - f(y + th_1)}{t} + \frac{f(y + th_1) - f(y)}{t} \right) \\
&\le \limsup_{\substack{y'\to x\\ t\downarrow 0}} \frac{f(y' + th_2) - f(y')}{t} + \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y + th_1) - f(y)}{t} \\
&= f^0(x, h_2) + f^0(x, h_1),
\end{aligned}$$

kde $y' = y + th_1 \to x$, takže $f^0(x, h)$ je subaditivní. Jelikož f je lipschitzovská s konstantou L v okolí bodu x, platí v tomto okolí

$$f(y + th_2) \le f(y + th_1) + Lt\|h_2 - h_1\|$$

(viz (25)), takže

$$\begin{aligned}
f^0(x, h_2) &= \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y + th_2) - f(y)}{t} \\
&\le \limsup_{\substack{y\to x\\ t\downarrow 0}} \frac{f(y + th_1) - f(y)}{t} + L\|h_2 - h_1\| \\
&= f^0(x, h_1) + L\|h_2 - h_1\|,
\end{aligned}$$

neboli

$$f^0(x, h_2) - f^0(x, h_1) \le L\|h_2 - h_1\|.$$

Protože nezáleží na pořadí vektorů h_1 a h_2, platí

$$|f^0(x, h_2) - f^0(x, h_1)| \le L\|h_2 - h_1\|.$$

Funkce $f^0(x,\cdot)$ je tedy lipschitzovská s konstantou L.
(b) Nechť $x_i \to x$ a $h_i \to h$. Z definice horní limity (limes superior) existují posloupnosti $y_i \to x$ a $t_i \downarrow 0$ takové, že

$$\begin{aligned} f^0(x_i,h_i) &\leq \frac{f(y_i+t_ih_i)-f(y_i)}{t_i}+\frac{1}{i} \\ &= \frac{f(y_i+t_ih)-f(y_i)}{t_i}+\frac{f(y_i+t_ih_i)-f(y_i+t_ih)}{t_i}+\frac{1}{i}. \end{aligned}$$

Z lipschitzovské spojitosti funkce f plyne

$$\left\| \frac{f(y_i+t_ih_i)-f(y_i+t_ih)}{t_i} \right\| \leq L\|h_i-h\|$$

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 167 Podle Definice 69 platí $f^0(x,0) = 0$, takže podle Věty 134 (a) dostaneme

$$|f^0(x,h)| = |f^0(x,h) - f^0(x,0)| \leq L\|h\|.$$

Definice 70 *Nechť funkce $f : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$. Pak množinu*

$$\partial f(x) = \{g \in R^n : f^0(x,h) \geq g^T h \quad \forall h \in R^n\}$$

nazveme subdiferenciálem funkce f v bodě x. Elementy $g \in \partial f(x)$ budeme nazývat subgradienty funkce f v bodě x.

Věta 135 *Nechť funkce $f : R^n \to R$ je lipschitzovská s konstantou L v okolí bodu $x \in R^n$. Pak:*

(a) Subdiferenciál $\partial f(x)$ je neprázdná konvexní kompaktní množina taková, že $\|g\| \leq L$ $\forall g \in \partial f(x)$.

(b) Pro libovolný vektor $h \in R^n$ platí

$$f^0(x,h) = \max\left\{g^T h : g \in \partial f(x)\right\}.$$

(c) Jestliže $x_i \to x$, $g_i \in \partial f(x_i)$ a $g_i \to g$, pak $g \in \partial f(x)$ (polospojitost shora).

(d) Platí $\partial(-f)(x) = -\partial f(x)$.

Důkaz (a) Podle Věty 134 (a) je funkce $f^0(x,\cdot)$ positivně homogenní a subaditivní. Podle Hahn-Banachovy věty (Věta **??**) 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 70 a Poznámky 167 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

$$f^0(x,\overline{h}) > \max\left\{g^T \overline{h} : g \in \partial f(x)\right\}. \tag{27}$$

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 $f^0(x,h) \ge \overline{g}^T h \ \forall h \in R^n$ 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 (27).
(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

$$g^T h = \lim_{i\to\infty} g_i^T h \le \limsup_{i\to\infty} f^0(x_i,h).$$

Podle Věty 134 (b) je funkce $f^0(\cdot,\cdot)$ shora polospojitá, takže $g^T h \le f^0(x,h)$.
(d) Vztah $g \in \partial(-f)(x)$ platí podle Definice 70 právě tehdy, jestliže $(-f)^0(x,h) \ge g^T h \ \forall h \in R^n$, což je podle Věty 134 (c) ekvivalentní $f^0(x,-h) \ge g^T h \ \forall h \in R^n$, což podle Definice 70 znamená $-g \in \partial f(x)$. Tedy $\partial(-f)(x) = -\partial f(x)$. □

Poznámka 168 Porovnáme-li Větu 135 (b) s Poznámkou 160 vidíme, že zobecněná směrová derivace je opěrnou funkcí subdiferenciálu, neboli

$$f^0(x,h) = \delta_{\partial f(x)}(h).$$

Věta 136 *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*

$$\partial f(x) = \{\nabla f(x)\}. \tag{28}$$

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 169 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 (28) lze dokázat pouze v případě spojité diferencovatelnosti.

Věta 137 *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 70. Abychom dokázali (a), stačí dokázat, že $f^0(x,h) \leq f'(x,h)$, neboť obrácenou nerovnost dostaneme ihned z Definice 69 (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 166). Položme $\overline{t}_i = \max\left(t_i, \sqrt{\|x_i - x\|}\right)$, takže $\|x_i - x\| \leq \overline{t}_i^2$, $t_i \leq \overline{t}_i$ a $\overline{t}_i \to 0$. Podle Věty 127 je funkce f lipschitzovská (s nějakou konstantou L) v okolí bodu x (bez újmy na obecnosti budeme předpokládat, že body x_i, $x_i + \overline{t}_i h$ a $x + \overline{t}_i h$ leží v tomto okolí). Použijeme-li Lemma 38 (levou nerovnost) dostaneme

$$\begin{aligned}
\frac{f(x_i+t_ih)-f(x_i)}{t_i} &\leq \frac{f(x_i+\overline{t}_ih)-f(x_i)}{\overline{t}_i} \\
&\leq \frac{f(x+\overline{t}_ih)-f(x)}{\overline{t}_i}+\frac{f(x_i+\overline{t}_ih)-f(x+\overline{t}_ih)}{\overline{t}_i}-\frac{f(x_i)-f(x)}{\overline{t}_i} \\
&\leq \frac{f(x+\overline{t}_ih)-f(x)}{\overline{t}_i}+\frac{2L\|x_i-x\|}{\overline{t}_i} \\
&\leq \frac{f(x+\overline{t}_ih)-f(x)}{\overline{t}_i}+2L\overline{t}_i
\end{aligned}$$

pro dostatečně velké indexy i. Provedeme-li limitní přechod na obou stranách této nerovnosti, dostaneme

$$f^0(x,h) \leq f'(x,h)+\lim_{\overline{t}_i\to 0} 2L\overline{t}_i = f'(x,h)$$

□

Poznámka 170 Věta 137 říká, že v případě konvexních funkcí je zobecněná směrová derivace totožná s obyčejnou směrovou derivací a subdiferenciál podle Definice 70 splývá se subdiferenciálem podle Definice 67.

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 71 *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 138 *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 136 (neboť $f^0(x,h) = \max_{g\in\partial f(x)} g^Th = (\nabla f(x))^Th = f'(x,h)$ $\forall h \in R^n$). Konvexní funkce je regulární podle Věty 137.
(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 128 (b) a Věty 134 (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 66 plyne, že $(f_1+f_2)'$ existuje a platí $(f_1+f_2)' = f_1'+f_2'$. Podle Definice 69 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 69 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 139 *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 70 platí $\partial(\lambda_1 f_1)(x) = \lambda_1 \partial f_1(x)$. V opačném případě s použitím Věty 135 (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 138 (a)). Použijeme-li Poznámku 168 a Větu 117, dostaneme

$$\delta_{\partial(f_1+f_2)(x)}(h) \leq \delta_{\partial f_1(x)}(h) + \delta_{\partial f_2(x)}(h) = \delta_{\partial f_1(x) + \partial f_2(x)}(h) \tag{29}$$

$\forall h \in R^n$, takže podle Věty 116 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 138 (a) platí $(f_1 + f_2)^0 = (f_1 + f_2)' = f_1' + f_2' = f_1^0 + f_2^0$, takže v (29) a tedy i v (b) nastane rovnost. Je-li funkce f_1 spojitě diferencovatelná v bodě x, pak podle Definice 69 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 171 *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 140 *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 70 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 135 (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 141 *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

$$\partial \varphi(\lambda) \subset (\partial f(x + \lambda(y - x)))^T (y - x). \tag{30}$$

Podle Věty 135 (a) jsou množiny na obou stranách této inkluze intervaly. Podle Věty 116 stačí dokázat, že

$$\delta_{\partial\varphi(\lambda)}(\beta) \le \delta_{(\partial f(x+\lambda(y-x)))^T(y-x)}(\beta) \tag{31}$$

pro $\beta = 1$ a $\beta = -1$. Podle Definice 69 a Věty 135 (b) platí

$$\begin{aligned}
\varphi^0(\lambda,\beta) &= \limsup_{\substack{\lambda'\to\lambda \\ t\downarrow 0}} \frac{\varphi(\lambda'+t\beta)-\varphi(\lambda')}{t} \\
&= \limsup_{\substack{\lambda'\to\lambda \\ t\downarrow 0}} \frac{f(x+(\lambda'+t\beta)(y-x))-f(x+\lambda'(y-x))}{t} \\
&\le \limsup_{\substack{y'\to x+\lambda(y-x) \\ t\downarrow 0}} \frac{f(y'+t\beta(y-x))-f(y')}{t} \\
&= f^0(x+\lambda(y-x),\beta(y-x)) \\
&= \max\left\{\beta g^T(y-x) : g\in\partial f(x+\lambda(y-x))\right\}
\end{aligned}$$

pro $\beta = 1$ a $\beta = -1$, což podle Poznámky 160 a Poznámky 168 dává (31) a tedy i (30). 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 140 dává $0 \in \partial\psi(\lambda^*)$. Použijeme-li Větu 139 a vztah (30), 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 oblasti Ω, je podle Rademacherovy věty (Věta **??**) 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 142 *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

$$f^0(x,h) \le \limsup_{\substack{y\to x \\ y\notin\Omega_f}} \nabla^T f(y)h. \tag{32}$$

Zvolme $h \in R^n$, $\varepsilon > 0$ libovolně a označme α pravou stranu v (32). 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

$$f(y+th)-f(y)\le t(\alpha+\varepsilon). \tag{33}$$

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 (33) platit pro všechny body $y\in B(x,\delta/2)$ a pro všechna $t\in(0,\delta/(2\|h\|))$, což podle Definice 69 dává

$$f^0(x,h)\le\alpha+\varepsilon.$$

Jelikož $\varepsilon>0$ je libovolné, dostáváme (32).
(b) Protože Ω_f má Lebesgueovu míru nula, existuje alespoň jedna posloupnost $y_i\to x$, $y_i\notin\Omega_f$. Podle Poznámky 169 platí $\nabla f(y_i)\in\partial f(y_i)$, takže podle Věty 135 (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 135 (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 168 a nerovnost (32), dostaneme

$$\begin{aligned}\delta_{\partial f(x)}(h) &= f^0(x,h)\le\limsup_{\substack{y\to x\\ y\notin\Omega_f}}\nabla^T f(y)h=\sup_{g\in\partial_B f(x)}g^Th\\ &\le\sup_{g\in\text{conv }\partial_B f(x)}g^Th=\delta_{\text{conv }\partial_B f(x)}(h)\end{aligned}$$

pro libovolný vektor $h\in R^n$, takže podle Věty 116 platí $\partial f(x)\subset$ conv $\partial_B f(x)$. □

11.4 Lipschitzovská zobrazení

Přístup použitý ve Větě 142 můžeme využít k definici zobecněného Jakobiánu lokálně lipschitzovského zobrazení $f:R^n\to R^m$. Stejně jako v případě lokálně lipschitzovské funkce zavedeme množinu

$$\Omega_f=\{x\in\Omega:\mathcal{J}f(x)\text{ neexistuje}\},$$

kde

$$\mathcal{J}f(x)=\begin{bmatrix}\frac{\partial f_1(x)}{\partial x_1}, & \dots, & \frac{\partial f_1(x)}{\partial x_n}\\ \dots & \dots & \dots\\ \frac{\partial f_m(x)}{\partial x_1}, & \dots, & \frac{\partial f_m(x)}{\partial x_n}\end{bmatrix},$$

která má opět Lebesgueovu míru nula.

Definice 72 *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak množinu*

$$\partial f(x) = \operatorname{conv} \partial_B f(x),$$

kde

$$\partial_B f(x) = \left\{ \lim_{i\to\infty} \mathcal{J} f(x_i) : x_i \to x, x_i \notin \Omega_f \right\},$$

nazveme zobecněným Jakobiánem zobrazení f.

Poznámka 172 Poznamenejme, že se dopouštíme jisté nedůslednosti, neboť pro $m = 1$ se Definice 72 odlišuje od Definice 70 (nyní jde o řádkový vektor). Tato konvence se však běžně používá v literatuře, takže se jí také přidržíme.

Poznámka 173 Je-li zobrazení $f : R^n \to R^m$ diferencovatelné v bodě $x \in R^n$, pak přímo z Definice 72 plyne, že

$$\mathcal{J} f(x) \in \partial f(x)$$

(stačí zvolit posloupnost $x_i = x \to x \notin \Omega_f$).

Věta 143 *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské s konstantou L v okolí bodu $x \in R^n$. Pak*

(a) Platí $\partial f(x) \subset [\partial f_1(x), \dots, \partial f_m(x)]^T$, kde $\partial f_i(x)$, $1 \le i \le m$, jsou subdiferenciály funkcí $f_i : R^n \to R$ (i-tých složek zobrazení f) v bodě $x \in R^n$.

(b) Zobecněný Jakobián $\partial f(x)$ je neprázdná konvexní kompaktní množina taková, že $\|J\| \le L \ \forall J \in \partial f(x)$.

(c) Jestliže $x_i \to x$, $J_i \in \partial f(x_i)$ a $J_i \to J$, pak $J \in \partial f(x)$ (polospojitost shora).

Důkaz (a) plyne bezprostředně z Věty 142.
(b) Kompaktnost plyne bezprostředně z (a) a z Věty 135 (a). Konvexita plyne přímo z Definice 72. Neprázdnost plyne z existence alespoň jedné posloupnosti $x_i \to x, x_i \notin \Omega_f$, pro kterou $\{\mathcal{J} f(x_i)\}$ konverguje (argumentace je stejná jako v důkazu Věty 142). Nerovnost $\|J\| \le L$ plyne z Definice 72 a z toho, že $\|\mathcal{J} f(x_i)\| \le L$ pokud $\mathcal{J} f(x_i)$ 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 72 musí existovat bod $y_i \in B(x_i, 1/(2i)) \subset B(x, 1/i)$ takový, že $\|\mathcal{J} f(y_i) - \overline{J}_i\| < \varepsilon/2$, takže

$$\mathcal{J} f(y_i) \notin \partial f(x) + B(0, \varepsilon/2). \tag{34}$$

Podle (a) jsou matice $\overline{J}_i$ a tedy i $\mathcal{J} f(y_i)$ stejnoměrně omezené v okolí bodu x. Můžeme tedy předpokládat, že existuje limita

$$\lim_{i\to\infty} \mathcal{J} f(y_i) = \overline{J}$$

(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 $\mathcal{J} f(y_i)$ existuje) a $\mathcal{J} f(y_i) \to \overline{J}$. Podle Definice 72 tedy platí $\overline{J} \in \partial f(x)$, což je ve sporu s (34). □

Lemma 39 *Nechť zobrazení $f : R^n \to R^m$ je lokálně lipschitzovské v okolí bodu $x \in R^n$ a funkce $g : R^m \to R$ je spojitě diferencovatelná v okolí bodu $f(x)$. Pak funkce $\varphi = g \circ f : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ a platí*

$$\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x)).$$

Důkaz Lipschitzovskost funkce $g \circ f$ je zřejmá (stačí použít Větu 136 a Definici 68). Nechť $J \in \partial_B f(x)$. Pak existuje posloupnost $x_i \to x$, $x_i \notin \Omega_f$ taková, že $\mathcal{J}f(x_i) \to J$ a tudíž $\nabla\varphi(x_i) = (\mathcal{J}f(x_i))^T \nabla g(f(x_i)) \to J^T \nabla g(f(x))$. Platí tedy $J^T \nabla g(f(x)) \in \partial_B \varphi(x)$, což dává

$$(\partial_B f(x))^T \nabla g(f(x)) \subset \partial_B \varphi(x).$$

Nechť naopak $w \in \partial_B \varphi(x)$. Pak existuje posloupnost $x_i \to x$, $x_i \notin \Omega_\varphi \supset \Omega_f$ taková, že $\nabla\varphi(x_i) = (\mathcal{J}f(x_i))^T \nabla g(f(x_i)) \to w$. Jelikož Jacobiovy matice $\mathcal{J}f(x_i)$ jsou podle Věty 143 (b) omezené v okolí bodu x, existuje podposloupnost $\{x_i'\} \subset \{x_i\}$ taková, že $\mathcal{J}f(x_i') \to J \in \partial_B f(x)$, což spolu s $(\mathcal{J}f(x_i'))^T \nabla g(f(x_i')) \to w$ dává

$$\partial_B \varphi(x) \subset (\partial_B f(x))^T \nabla g(f(x)).$$

Spojením obou inkluzí dostaneme $\partial_B \varphi(x) = (\partial_B f(x))^T \nabla g(f(x))$, což po přechodu ke konvexním obalům dává $\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x))$. □

Abychom mohli zformulovat větu o střední hodnotě, zavedeme označení

$$\partial f([x,y]) = \operatorname{conv} \bigcup_{z \in [x,y]} \partial f(z). \tag{35}$$

Lemma 40 *Množina $\partial f([x,y])$ je konvexní a kompaktní.*

Důkaz Konvexita plyne bezprostředně z (35). Abychom dokázali kompaktnost, stačí podle Věty 108 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 143 (c) dostaneme $J \in \partial f(z) \subset \bigcup_{z\in[x,y]} \partial f(z)$. □

Věta 144 *Nechť zobrazení $f : R^n \to R^m$ je lokálně lipschitzovské na otevřené množině Ω obsahující úsečku $[x,y]$. Pak platí*

$$f(y) - f(x) \in \partial f([x,y])\,(y-x). \tag{36}$$

Důkaz Podle Lemmatu 39 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 141, podle které pro libovolný vektor $v \in R^m$ existuje bod $z \in (x,y)$ takový, že

$$v^T(f(y) - f(x)) \in \partial(v^T f)(z)(y-x) = v^T \partial f(z)(y-x). \tag{37}$$

Vztah (36) dokážeme sporem. Předpokládejme, že $f(y) - f(x) \notin \partial f([x,y])\,(y-x)$. Jelikož množina na pravé straně je podle Lemmatu 40 konvexní a kompaktní, musí podle Věty 112 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 (37), neboť podle (37) existuje prvek $J \in \partial f(z) \subset \partial f([x,y])$ takový, že $v^T(f(y)-f(x)) = v^T J(y-x)$. □

Věta 145 *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$ a funkce $g : R^m \to R$ je lipschitzovská v okolí bodu $f(x)$. Pak funkce $\varphi = g \circ f : R^n \to R$ je lipschitzovská v okolí bodu $x \in R^n$ a platí*

$$\partial\varphi(x) \subset \text{conv } (\partial f(x))^T \partial g(f(x)) \stackrel{\Delta}{=} \text{conv } \left\{J^T v : J \in \partial f(x), v \in \partial g(f(x))\right\}, \tag{38}$$

přičemž rovnost nastává zejména v těchto případech

(a) Funkce g je spojitě diferencovatelná v bodě $f(x)$. V tomto případě platí

$$\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x)). \tag{39}$$

(b) Funkce g je regulární v bodě $f(x)$ a zobrazení f je spojitě diferencovatelné v bodě x. V tomto případě je funkce φ regulární v bodě x a platí

$$\partial\varphi(x) = (\mathcal{J}f(x))^T \partial g(f(x)). \tag{40}$$

(c) Funkce g 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 $v \in \partial g(f(x))$ platí $v_i \geq 0$, $1 \leq i \leq m$. V tomto případě je funkce φ regulární v bodě x.

Důkaz Lipschitzovskost funkce $g \circ f$ je zřejmá (stačí dvakrát použít Definici 68). Označme S množinu na pravé straně (38). Abychom dokázali inkluzi $\partial\varphi(x) \subset S$, použijeme Větu 116 a Poznámku 168. Jelikož podle Věty 115 pro libovolný vektor $h \in R^n$ platí

$$\delta_S(h) = \max\left\{v^T Jh : J \in \partial f(x), v \in \partial g(f(x))\right\},$$

stačí podle Věty 116 a Poznámky 168 ukázat, že pro libovolný vektor $h \in R^n$ existuje matice $J \in \partial f(x)$ a vektor $v \in \partial g(f(x))$ tak, že

$$\delta_{\partial\varphi(x)}(h) = \varphi^0(x,h) \leq v^T Jh. \tag{41}$$

Podle Poznámky 166 můžeme vybrat posloupnosti $x_i \to x$ a $t_i \downarrow 0$ tak, že

$$\varphi^0(x,h) = \lim_{i\to\infty} \frac{\varphi(x_i + t_i h) - \varphi(x_i)}{t_i}.$$

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 141 (aplikované na funkci g) a existuje tedy bod $u_i \in [f(x_i), f(x_i + t_i h)]$ a subgradient $v_i \in \partial g(u_i)$ tak, že

$$\varphi(x_i + t_i h) - \varphi(x_i) = g(f(x_i + t_i h)) - g(f(x_i)) = v_i^T(f(x_i + t_i h) - f(x_i)).$$

Podle Věty 144 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 (35) a podle Věty 107 znamená, že

$$\frac{\varphi(x_i+t_ih)-\varphi(x_i)}{t_i}=v_i^T\frac{f(x_i+t_ih)-f(x_i)}{t_i}=v_i^T\sum_{k=1}^{m+1}\lambda_i^kJ_i^kh,$$

kde $J_i^k\in\partial f(y_i^k)$, $y_i^k\in[x_i,x_i+t_ih]$, $\lambda_i^k\geq 0$, $k\in[1,m+1]$, $\lambda_i^1+\ldots+\lambda_i^{m+1}=1$. Z tohoto důvodu musí alespoň pro jeden index $k\in[1,m+1]$ platit

$$\frac{\varphi(x_i+t_ih)-\varphi(x_i)}{t_i}\leq v_i^TJ_i^kh. \tag{42}$$

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 135 (c) a Věty 143 (c) platí $v\in\partial g(f(x))$ a $J\in\partial f(x)$, takže z (42) plyne (41). Nyní vyšetříme speciální případy:
(a) Tento případ je tvrzením Lemmatu 39.
(b) Je-li zobrazení f spojitě diferencovatelné, můžeme množinu S zapsat ve tvaru $S=(\mathcal{J}f(x))^T\partial g(f(x))$ (protože množina $\partial f(x)=\{\mathcal{J}f(x)\}$ je jednoprvková nemusíme používat její konvexní obal). Použijeme-li Definici 60, Poznámku 168 a regularitu funkce g (Definice 71), můžeme psát

$$\begin{aligned}\delta_S(h) &= \max_{v\in\partial g(f(x))}v^T\mathcal{J}f(x)h=\max_{v\in\partial g(f(x))}v^Tf'(x,h)\\ &= g^0(f(x),f'(x,h))=g'(f(x),f'(x,h))\\ &= \lim_{t\downarrow 0}\frac{g(f(x)+tf'(x,h))-g(f(x))}{t}=\lim_{t\downarrow 0}\left(\frac{g(f(x+th))-g(f(x))}{t}+T(t)\right),\end{aligned}$$

kde pro dostatečně malá t platí

$$\begin{aligned}\|T(t)\| &= \frac{\|g(f(x)+tf'(x,h))-g(f(x+th))\|}{t}\leq\frac{L\|f(x)+tf'(x,h)-f(x+th)\|}{t}\\ &= L\left\|f'(x,h)-\frac{f(x+th)-f(x)}{t}\right\|,\end{aligned}$$

neboť funkce g 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

$$\varphi'(x,h)=\lim_{t\downarrow 0}\frac{g(f(x+th))-g(f(x))}{t}$$

existuje a platí $\delta_S(h)=\varphi'(x,h)\leq\varphi^0(x,h)$, což podle Věty 116 dává $S\subset\partial\varphi(x)$, takže z (38) plyne $\partial\varphi(x)=S$. Z nerovnosti $\varphi^0(x,h)\leq\delta_S(h)=\varphi'(x,h)\leq\varphi^0(x,h)$ pak plyne regularita funkce φ v bodě x.
(c) Označme

$$S'=\operatorname{conv}\left\{\sum_{i=1}^m u_iv_i:u_i\in\partial f_i(x),v\in\partial g(f(x))\right\}.$$

Podle (38) platí $\partial\varphi(x)\subset S$ a podle Věty 143 (a) platí $S\subset S'$, takže $\partial\varphi(x)\subset S'$. Jsou-li funkce g a f_i, $1\leq i\leq m$, regulární a platí-li $v_i\geq 0$, $1\leq i\leq m$, můžeme psát

$$\delta_{S'}(h) = \max\left\{\sum_{i=1}^m v_iu_i^Th:u_i\in\partial f_i(x),v\in\partial g(f(x))\right\}$$

$$
\begin{aligned}
&\leq \max\left\{\sum_{i=1}^{m} v_i \max_{u_i\in\partial f_i(x)} u_i^T h : v\in\partial g(f(x))\right\}\\
&= \max\left\{\sum_{i=1}^{m} v_i f_i^0(x,h) : v\in\partial g(f(x))\right\}\\
&= \max\left\{\sum_{i=1}^{m} v_i f_i'(x,h) : v\in\partial g(f(x))\right\}\\
&= g^0(f(x), f'(x,h)) = g'(f(x), f'(x,h)).
\end{aligned}
$$

Konec důkazu je již stejný jako konec důkazu tvrzení (b). Dostaneme $\delta_{S'}(h) = \varphi'(x,h) \leq \varphi^0(x,h)$, což podle Věty 116 dává $S' \subset \partial\varphi(x)$, takže z $\partial\varphi(x) \subset S \subset S'$ plyne $\partial\varphi(x) = S = S'$. Z nerovnosti $\varphi^0(x,h) \leq \delta_S(h) \leq \delta_{S'}(h) = \varphi'(x,h) \leq \varphi^0(x,h)$ pak plyne regularita funkce φ v bodě x. □

Důsledek 14 *Jsou-li splněny předpoklady Věty 145, platí*

$$
\partial\varphi(x) \subset \operatorname{conv}\left\{\sum_{i=1}^{m} u_i v_i : u_i\in\partial f_i(x), v\in\partial g(f(x)),\right\} \tag{43}
$$

přičemž rovnost nastává zejména v těchto případech:

(a) Funkce g je spojitě diferencovatelná v bodě $f(x)$ a $m = 1$.

(b) Funkce g je regulární v bodě $f(x)$ a funkce f_i, $1 \leq i \leq m$, jsou spojitě diferencovatelné v bodě x.

(c) Funkce g je regulární v bodě $f(x)$ a funkce f_i, $1 \leq i \leq m$, jsou regulární v bodě x a pro libovolný prvek $v \in \partial g(f(x))$ platí $v_i \geq 0$, $1 \leq i \leq m$.

Důkaz Stačí použít Větu 145 a některé úvahy (například $S \subset S'$) z jejího důkazu. □

Důsledek 15 *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 $\varphi = f_1 f_2$ 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

$$
\partial\varphi(x) = (\partial f(x))^T P f(x) \subset \partial f_1(x) f_2(x) + f_1(x)\partial f_2(x)
$$

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 $\varphi = f_1 f_2$ regulární.

Důkaz Definujme funkci $g : R^2 \to R$ předpisem $g(u_1, u_2) = u_1 u_2$. Tato funkce je spojitě diferencovatelná a tedy (podle Věty 136) lipschitzovská v okolí libovolného bodu $u \in R^2$, přičemž platí

$$
\nabla g(u) = \begin{bmatrix} u_2 \\ u_1 \end{bmatrix} = Pu.
$$

Podle Věty 145 je funkce $g \circ f = f_1 f_2$ lipschitzovská v okolí bodu x a platí

$$
\partial(f_1 f_2) = \left\{J^T \nabla g(f(x)) : J \in \partial f(x)\right\} = (\partial f(x))^T P f.
$$

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 14 (c). □

Důsledek 16 *Nechť zobrazení $f : R^n \to R^m$ je lipschitzovské v okolí bodu $x \in R^n$. Pak funkce $\varphi = (1/2)f^T f$ je lipschitzovská v okolí bodu x a platí*

$$\partial\varphi(x) = \frac{1}{2}\partial(f^T f)(x) = (\partial(f(x))^T f(x) = \left\{J^T f(x) : J \in \partial(f(x))\right\}. \tag{44}$$

Důkaz Definujme funkci $g : R^m \to R$ předpisem

$$g(u) = \frac{1}{2}u^T u = \frac{1}{2}\sum_{i=1}^{m} u_i^2.$$

Tato funkce je spojitě diferencovatelná a tedy (podle Věty 136) lipschitzovská v okolí libovolného bodu $u \in R^m$, přičemž platí $\nabla g(u) = u$. Podle Věty 145 (a) je funkce $\varphi = g \circ f = (1/2)f^T f$ lipschitzovská v okolí bodu x a platí $\partial\varphi(x) = (\partial f(x))^T \nabla g(f(x)) = (\partial f(x))^T f(x)$. □

Věta 146 *Nechť funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou lipschitzovské v okolí bodu $x \in R^n$. Pak funkce*

$$\varphi(x) = \max_{1\le i\le m} f_i(x)$$

je lipschitzovská v okolí bodu x a platí

$$\partial\varphi(x) \subset \text{conv } \{\partial f_i(x) : i \in I(x)\}, \tag{45}$$

kde $I(x) = \{i \in \{1,\dots,m\} : f_i(x) = \varphi(x)\}$. Jsou-li funkce f_i, $1 \le i \le m$, regulární v bodě x, je funkce φ regulární v bodě x a v (45) platí rovnost.

Důkaz Definujme funkci $g : R^m \to R$ předpisem $g(u) = \max(u_1,\dots,u_m)$. Tato funkce je konvexní v R^m, neboť

$$\begin{aligned} g(\lambda u + (1-\lambda)v) &= \max_{1\le i\le m}(\lambda u_i + (1-\lambda)v_i) \le \lambda \max_{1\le i\le m}(u_i) + (1-\lambda)\max_{1\le i\le m}(v_i) \\ &= \lambda g(u) + (1-\lambda)g(v) \end{aligned}$$

pro $u \in R^m$, $v \in R^m$ a $1 \le \lambda \le 1$, takže je lokálně lipschitzovská podle Věty 127. Nechť $I(u) = \{i \in \{1,\dots,m\} : u_i = g(u)\}$. Pak platí

$$\begin{aligned} g'(u,d) &= \lim_{t\downarrow 0}\frac{g(u+td)-g(u)}{t} = \lim_{t\downarrow 0}\max_{1\le i\le m}\left(\frac{u_i+td_i-g(u)}{t}\right) = \\ &= \lim_{t\downarrow 0}\max_{i\in I(u)}\left(\frac{u_i+td_i-g(u)}{t}\right) = \max_{i\in I(u)}(d_i), \end{aligned}$$

takže $g^0(u,d) = g'(u,d) = \max_{i\in I(u)}(d_i)$ a podle Definice 70 platí

$$\partial g(u) = \left\{v \in R^n : \max_{i\in I(u)}(d_i) \ge v^T d \ \ \forall d \in R^n\right\}.$$

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 $v \notin \partial g(u)$. 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 $v \notin \partial g(u)$. 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 $v \notin \partial g(u)$. 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 $v \notin \partial g(u)$. Musí tedy platit

$$\partial g(u) = \left\{ v \in R^n : v_i \geq 0, \sum_{i \in I(u)} v_i = 1, \sum_{i \notin I(u)} v_i = 0 \right\}.$$

Podle Důsledku 14 pak platí

$$\begin{aligned} \partial f(x) &\subset \text{conv} \left\{ \sum_{i=1}^{m} v_i u_i : u_i \in \partial f_i(x), v \in \partial g(f(x)) \right\} \\ &= \text{conv} \left\{ \sum_{i \in I(u)} v_i \partial f_i(x) : v_i \geq 0, \sum_{i \in I(u)} v_i = 1 \right\} \\ &= \text{conv} \left\{ \partial f_i(x), i \in I(u) \right\}. \end{aligned}$$

Funkce g je konvexní, takže je podle Věty 138 regulární. Jsou-li funkce f_i, $1 \leq i \leq m$, regulární, je podle Věty 138 i funkce φ regulární a jelikož $v_i \geq 0$, $1 \leq i \leq m$, platí v (45) rovnost. □

11.5 Polohladká zobrazení

Definice 73 *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*

$$\lim_{\substack{J \in \partial f(x+th) \\ t \downarrow 0}} Jh \tag{46}$$

(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

$$\lim_{\substack{J \in \partial f(x+th') \\ h' \to h, t \downarrow 0}} Jh' \tag{47}$$

(nezávislá na volbě $J \in \partial f(x+th')$), řekneme, že zobrazení f je polohladké v bodě x.

Poznámka 174 Jelikož $\partial f(x)$ je množinové zobrazení, mohlo by se zdát, že existence limity (47) je výjimečná. V dalším textu však ukážeme (Poznámka 178), že polohladkost je vlastnost převážné většiny zajímavých lokálně lipschitzovských zobrazení.

Poznámka 175 Z Definice 73 plyne, že každé polohladké zobrazení je slabě polohladké. Slabá polohladkost se však nezachovává při skládání funkcí a také Věta 152 vyžaduje platnost vztahu (47).

Věta 147 *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 144, dostaneme

$$\frac{f(x+t_ih)-f(x)}{t_i} \in \partial f\left([x, x+t_ih]\right) h = \left(\operatorname{conv} \bigcup_{t\in[0,1]} \partial f(x+th)\right) h = \operatorname{conv}\left(\bigcup_{t\in[0,1]} \partial f(x+th)h\right) \subset R^m$$

Podle Věty 107 existuje nejvýše $m+1$ prvků $J_i^k \in \partial f(x + t_i^k h)$, $t_i^k \in [0, t_i]$, $1 \le k \le m+1$, tak, že

$$\frac{f(x+t_ih)-f(x)}{t_i} = \sum_{k=1}^{m+1} \lambda_i^k J_i^k h,$$

kde $0 \le \lambda_i^k \le 1$ a $\lambda_1^k + \ldots + \lambda_{m+1}^k = 1$. 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 (46) platí

$$\begin{aligned}\lim_{i\to\infty} \frac{f(x+t_ih)-f(x)}{t_i} &= \lim_{i\to\infty}\left(\sum_{k=1}^{m+1} \lambda_i^k J_i^k h\right) = \sum_{k=1}^{m+1}\left(\lim_{i\to\infty}\lambda_i^k\right)\left(\lim_{i\to\infty} J_i^k h\right)\\ &= \left(\sum_{k=1}^{m+1}\lambda^k\right) \lim_{\substack{J\in\partial f(x+th)\\ t\downarrow 0}} Jh = \lim_{\substack{J\in\partial f(x+th)\\ t\downarrow 0}} Jh,\end{aligned}$$

takže limita na levé straně nezávisí na výběru posloupnosti $t_i \downarrow 0$ a rovná se směrové derivaci $f'(x, h)$.
(b) Nechť vektor $h \in R^n$ je libovolný. Jelikož zobrazení f je lipschitzovské v okolí bodu $x \in R^n$ (s nějakou konstantou L) platí

$$\lim_{h'\to h,\, t\downarrow 0} \frac{\|f(x+th')-f(x+th)\|}{t} \le \lim_{h'\to h} L\|h'-h\| = 0.$$

Každé polohladké zobrazení je slabě polohladké. Můžeme tedy psát

$$\begin{aligned}\lim_{h'\to h,\, t\downarrow 0} \frac{f(x+th')-f(x)}{t} &= \lim_{t\downarrow 0}\frac{f(x+th)-f(x)}{t}\\ &\quad + \lim_{h'\to h,\, t\downarrow 0}\frac{f(x+th')-f(x+th)}{t} = f'(x,h).\end{aligned}$$

Zbytek tvrzení plyne z (a). □

Poznámka 176 Zobrazení $f'(x, .) : R^n \to R^m$ (směrová derivace) vystupující ve větě 147 je pozitivně homogení a lipschitzovské. Není však subaditivní jako v případě konvexních funkcí.

Poznámka 177 Podle věty 147 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 172.

Věta 148 *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 (47) 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 149 *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 150 *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 129 (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 67 platí

$$g^T h' \le f'(x+th', h').$$

Použijeme-li tyto nerovnosti spolu s Větou 147, Definicí 66 a Větou 128 (c), dostaneme

$$\begin{aligned} f'(x,h) &\le \lim_{\substack{h' \to h \\ t \downarrow 0}} \frac{f(x+th') - f(x)}{t} \le \liminf_{\substack{g \in \partial f(x+th') \\ h' \to h,\ t \downarrow 0}} g^T h' \le \limsup_{\substack{g \in \partial f(x+th') \\ h' \to h,\ t \downarrow 0}} g^T h' \\ &\le \limsup_{h' \to h,\, t \downarrow 0} f'(x+th', h') \le f'(x,h), \end{aligned}$$

což dokazuje existenci požadované limity. □

Věta 151 *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ a funkce $g : R^m \to R$ je polohladká v bodě $f(x)$. Pak složené zobrazení $\varphi = g \circ f$ je polohladké 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 145 platí $\partial\varphi(x_k) \subset S_k$, kde symbol $S_k \subset R^n$ označuje kompaktní množinu na pravé straně výrazu (38) (s x_k místo x). Nechť

$$\begin{aligned} w_k^- &= (J_k^-)^T v_k^- = \arg\min_{w \in S_k} w^T h, \quad v_k^- \in \partial g(f(x_k)), \quad J_k^- \in \partial f(x_k), \\ w_k^+ &= (J_k^+)^T v_k^+ = \arg\max_{w \in S_k} w^T h, \quad v_k^+ \in \partial g(f(x_k)), \quad J_k^+ \in \partial f(x_k). \end{aligned}$$

Pak pro libovolný vektor $w_k \in \partial\varphi(x_k) \subset S_k$ platí

$$(w_k^-)^T h \le w_k^T h \le (w_k^+)^T h. \tag{48}$$

Jelikož všechny veličiny v těchto vzorcích jsou podle Věty 135 (a) omezené, můžeme předpokládat (po případném přechodu k podposloupnostem), že

$$\begin{aligned} J_k^- &\to J^- \in \partial f(x), \qquad v_k^- \to v^- \in \partial g(f(x)), \\ J_k^+ &\to J^+ \in \partial f(x), \qquad v_k^+ \to v^+ \in \partial g(f(x)) \end{aligned}$$

(používáme Větu 135 (c)). Jelikož zobrazení f je polohladké, platí $J^- h = J^+ h = f'(x,h)$, takže s použitím (48) 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 g je polohladká a podle Poznámky 177 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 73, 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 $w_k \in \partial\varphi(x_k)$. □

Důsledek 17 *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ a funkce $g : R^m \to R$ je buď spojitě diferencovatelná nebo konvexní v okolí bodu $f(x)$. Pak funkce $\varphi = g \circ f$ je polohladká v bodě x.*

Důkaz Tvrzení plyne bezprostředně z Věty 149, Věty 150 a Věty 151. □

Důsledek 18 *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 $\varphi_1 = \sum_{i=1}^m \lambda_i f_i$ (lineární kombinace) a $\varphi_2 = \prod_{i=1}^m f_i$ (součin) jsou polohladké v bodě x.*

Důkaz Podle Věty 148 je zobrazení $f = [f_1, \ldots, f_m]^T$ polohladké v bodě x. Funkce $g_1(u) = \sum_{i=1}^m \lambda_i u_i$ a $g_2(u) = \prod_{i=1}^m u_i$ jsou spojitě diferencovatelné, takže podle Důsledku 17 jsou funkce $\varphi_1 = g_1 \circ f$ a $\varphi_2 = g_2 \circ f$ polohladké v bodě x. □

Důsledek 19 *Nechť funkce $f_i : R^n \to R$, $1 \le i \le m$ jsou polohladké v bodě $x \in R^n$. Pak funkce $\varphi = \|[f_1, \ldots, f_m]^T\|$, kde $\|\cdot\|$ je libovolná norma v R^m, je polohladká v bodě x. Speciálně funkce $\varphi_1 = \max_{1\le i\le m}(|f_i|)$ (maximum absolutních hodnot) a $\varphi_2 = \sum_{i=1}^m |f_i|$ (součet absolutních hodnot) jsou polohladké v bodě x. Dále funkce $\varphi_3 = \max_{1\le i\le m}(f_i)$ (bodové maximum) je polohladká v bodě x.*

Důkaz Podle Věty 148 je zobrazení $f = [f_1, \ldots, f_m]^T$ polohladké v bodě x. Funkce $g(u) = \|u\|$ je konvexní, neboť z vlastností vektorové normy plyne, že pro $0 \le \lambda \le 1$ platí

$$g(\lambda u + (1-\lambda)v) = \|\lambda u + (1-\lambda)v\| \le \lambda\|u\| + (1-\lambda)\|v\|.$$

Funkce $\varphi = g \circ f$ je tedy podle Důsledku 17 polohladká. Také funkce $g(u) = \max_{1\le i\le m}(u_i)$ je konvexní (důkaz Věty 146), takže funkce $\varphi_3 = g \circ f$ je podle Důsledku 17 polohladká. □

Důsledek 20 *(Obrácení Věty 148). Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$ přičemž $f = [f_1, \ldots, f_m]^T$. Pak funkce $f_i : R^n \to R$, $1 \le i \le m$, jsou polohladké v bodě x.*

Důkaz Zřejmě $f_i = g_i \circ f$, $1 \le i \le m$, kde funkce $g_i : R^m \to R$, definované předpisem $g_i(u) = e_i^T u = u_i$, jsou spojitě diferencovatelné. Polohladkost funkcí $f_i : R^n \to R$, $1 \le i \le m$, tedy plyne z Důsledku 17. □

Důsledek 21 *Lineární kombinace polohladkých zobrazení je polohladké zobrazení. Skalární součin polohladkých zobrazení je polohladká funkce.*

Důkaz Podle Důsledku 20 jsou složky polohladkých zobrazení polohladkými funkcemi. Podle Důsledku 18 je lineární kombinace polohladkých funkcí polohladkou funkcí, takže podle Věty 148 je lineární kombinace polohladkých zobrazení polohladkým zobrazením. Polohladkost skalárního součinu plyne z Důsledku 20, Důsledku 18 a Věty 148. □

Poznámka 178 Z předchozího textu vyplývá, že vycházíme-li ze spojitě diferencovatelných a konvexních zobrazení, dostáváme běžnými operacemi (součet, součin, maximum, skládání funkcí) pouze polohladká zobrazení. Proto má teorie polohladkých zobrazení velké uplatnění v praktických aplikacích. Navíc je polohladkost základním předpokladem pro konstrukci numerických metod pro řešení nehladkých rovnic.

V následujících úvahách budeme používat symbol $o(\|h\|)$ pokud $h \to 0$. Tento symbol znamená, že pro libovolnou posloupnost $h_i \to 0$, $h_i \neq 0$ platí $o(\|h_i\|)/\|h_i\| \to 0$.

Věta 152 *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*

$$Jh - f'(x,h) = o(\|h\|) \tag{49}$$

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

$$\lim_{i\to\infty} \frac{J_i h_i - f'(x,h_i)}{\|h_i\|} = 0. \tag{50}$$

pro libovolné posloupnosti $\{h_i\} \subset R^n$ a $\{J_i\} \subset R^{m\times n}$ takové, že $h_i \to 0$ a $J_i \in \partial f(x+h_i)$. Předpokládejme naopak, že existují posloupnosti $\{h_i\} \subset R^n$ a $\{J_i\} \subset R^{m\times n}$ takové, že $h_i \to 0$ a $J_i \in \partial f(x+h_i)$, a číslo $\varepsilon > 0$ takové, že že

$$\frac{\|J_i h_i - f'(x,h_i)\|}{\|h_i\|} = \|J_i h_i' - f'(x,h_i')\| \geq \varepsilon \quad \forall i \in N,$$

kde $h_i' = h_i/\|h_i\|$ a $t_i = \|h_i\|$ (takže $J_i \in \partial f(x+t_i h_i')$). Pak podle Věty 147 (b) 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 177 spojitá.
(b) Předpokládejme nyní, že 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

$$\|J_i h_i' - f'(x,h)\| \geq 2\varepsilon \quad \forall i \in N \tag{51}$$

(v opačném případě by existovala limita (47) rovnající se $f'(x,h)$, takže zobrazení f by bylo podle Definice 73 polohladké). Jelikož směrová derivace je podle Poznámky 176 lipschitzovská, platí pro dostatečně velké indexy $\|f'(x,h_i') - f'(x,h)\| \leq \varepsilon$, což spolu s (51) 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 (50) a tudíž ani (49). □

Poznámka 179 Vzhledem k platnosti Věty 152 se polohladké zobrazení často definuje jako lokálně lipschitzovské zobrazení, které vyhovuje podmínce (49).

Definice 74 *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*

$$f(x+h) - f(x) - f'(x,h) = o(\|h\|), \tag{52}$$

pokud $h \to 0$ (to znamená, že zobrazení $f'(x,.)$ má stejné aproximační vlastnosti jako Frechetova derivace).

Věta 153 *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

$$\frac{|f(x+h_i) - f(x) - f'(x,h_i)|}{\|h_i\|} = \left| \frac{f(x+t_i h_i') - f(x)}{t_i} - f'(x,h_i') \right| \geq \varepsilon \quad \forall i. \tag{53}$$

kde $h_i' = h_i/\|h_i\|$ a $t_i = \|h_i\|$. Jelikož vektory h_i' jsou omezené (neboť $\|h_i'\| = 1$), můžeme tuto podposloupnost bez újmy na obecnosti vybrat tak, že $h_i' \to h$. Pak podle Věty 147 (b) 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 (53), neboť funkce $f'(x,\cdot)$ je podle Poznámky 177 spojitá. □

Důsledek 22 *Nechť zobrazení $f : R^n \to R^m$ je polohladké v bodě $x \in R^n$. Pak platí*

$$f(x+h) - f(x) - Jh = o(\|h\|), \tag{54}$$

pokud $h \to 0$ a $J \in \partial f(x+h)$.

Důkaz Tvrzení plyne bezprostředně z Věty 152 a Věty 153. □

12 Metody pro řešení soustav nehladkých rovnic

12.1 Newtonova metoda

Nyní se budeme zabývat řešením soustavy rovnic

$$f(x) = 0, \tag{55}$$

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

$$x_{k+1} = x_k + d_k, \tag{56}$$

kde vektor d_k se vybírá tak, aby platilo

$$\omega_k = \frac{\|A_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega \tag{57}$$

a matice A_k se vybírá tak, aby platilo

$$\Delta_k = \|A_k - J_k\| \leq \Delta \tag{58}$$

pro nějaký prvek $J_k \in \partial_B f(x_k)$. Přitom $\omega \geq 0$, $\Delta \geq 0$ a normy v (57) a (58) jsou euklidovské.

Definice 75 *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 72).*

Poznámka 180 V iteračním procesu (56)-(58) 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 154 *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 $\mathcal{J}f(z)$ jsou regulární a platí

$$\|(\mathcal{J}f(z))^{-1}\| \leq c, \tag{59}$$

pokud $z \in B(x, \delta) \backslash \Omega_f$ (množina Ω_f je uvedena v Definici 72). Předpokládejme, že (59) neplatí. Pak musí existovat posloupnost $x_i \to x$, $x_i \in B(x, \delta) \backslash \Omega_f$ taková, že buď všechny Jacobiho matice $\mathcal{J}f(x_i)$ jsou singulární nebo $\|(\mathcal{J}f(x_i))^{-1}\| \to \infty$. Jelikož zobrazení f je lipschitzovské v okolí bodu x, jsou podle Věty 143 (b) Jacobiovy matice $\mathcal{J}f(x_i)$ omezené v okolí bodu x. Existuje tedy podposloupnost $\{x_i'\} \subset \{x_i\}$ taková, že $\mathcal{J}f(x_i') \to J$. Ze spojité závislosti vlastních čísel na koeficientech matice plyne, že J musí být singulární. Podle Definice 72 platí $J \in \partial_B f(x)$, což je v rozporu s Definicí 75. 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 (59) platí pokud $z \in B(y, \delta') \backslash \Omega_f$. Jelikož podle Definice 72 platí

$$J = \lim_{i \to \infty} \mathcal{J}f(y_i)$$

pro nějakou posloupnost $y_i \to y$, $y_i \in B(y, \delta') \backslash \Omega_f$, dostaneme z (59) a ze spojité závislosti vlastních čísel na koeficientech matice nerovnost $\|J^{-1}\| \leq c$. □

Věta 155 *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 (56)-(58) 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 154. Položme $\Delta = 1/(5c)$ a zvolme $\varepsilon \leq \delta$ tak, aby zobrazení f bylo lipschitzovské (s nějakou konstantou L) v $B(x^*, \varepsilon)$ a aby platilo

$$\|f(x) - f(x^*) - J(x - x^*)\| \leq \frac{\Delta}{2}\|x - x^*\| \quad \forall J \in \partial_B f(x), \tag{60}$$

pokud $x \in B(x^*, \varepsilon)$ (to je možné vzhledem k (54)). Dále položme $\omega = \Delta/(2L)$. Předpokládejme, že $x_k \in B(x^*, \varepsilon)$ (platí to pro $k = 1$). Pak podle Věty 154 platí $\|J_k^{-1}\| \leq c$. Zřejmě

$$A_k^{-1} + J_k^{-1}(A_k - J_k)A_k^{-1} = J_k^{-1}.$$

Jelikož rozdíl norem není větší než norma rozdílu, můžeme psát

$$\|A_k^{-1}\| - \|J_k^{-1}\|\|A_k - J_k\|\|A_k^{-1}\| \leq \|J_k^{-1}\|,$$

neboli

$$\|A_k^{-1}\| \le \frac{\|J_k^{-1}\|}{1-\|J_k^{-1}\|\|A_k-J_k\|} \le \frac{c}{1-c\Delta} = \frac{5}{4}c,$$

což podle (56)-(58) a (60) (s využitím vztahu $f(x^*)=0$) dává

$$\begin{aligned}
\|x_{k+1}-x^*\| &= \|x_k+d_k-x^*\| = \|x_k+A_k^{-1}(A_kd_k+f(x_k)-f(x_k))-x^*\| \\
&= \|A_k^{-1}(A_kd_k+f(x_k)-(f(x_k)-J_k(x_k-x^*))+(A_k-J_k)(x_k-x^*))\| \\
&\le \|A_k^{-1}\|\,(\|f(x_k)-f(x^*)-J_k(x_k-x^*)\| \\
&\quad +\|A_k-J_k\|\|x_k-x^*\|+\omega_k\|f(x_k)-f(x^*)\|) \\
&\le \frac{5}{4}c\,(\|f(x_k)-f(x^*)-J_k(x_k-x^*)\| \\
&\quad +\Delta_k\|x_k-x^*\|+\omega_kL\|x_k-x^*\|) \\
&\le \frac{5}{4}c\left(\frac{1}{2}\Delta+\Delta+\frac{1}{2}\Delta\right)\|x_k-x^*\| = \frac{1}{2}\|x_k-x^*\|. \qquad (61)
\end{aligned}$$

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 (54) a (61) platí

$$\begin{aligned}
\|x_{k+1}-x^*\| &\le \frac{5}{4}c\,(\|f(x_k)-f(x^*)-J_k(x_k-x^*)\| \\
&\quad +\Delta_k\|x_k-x^*\|+\omega_kL\|x_k-x^*\|) \\
&= \frac{5}{4}c\,(o\,(\|x_k-x^*\|)+o\,(\|x_k-x^*\|)+o\,(\|x_k-x^*\|)) \\
&= o\,(\|x_k-x^*\|) \qquad (62)
\end{aligned}$$

a posloupnost $\{x_k\}$ konverguje k bodu x^* Q-superlineárně. Jelikož $f(x^*)=0$, můžeme podle (62) psát

$$\lim_{k\to\infty}\frac{\|f(x_{k+1})\|}{\|x_k-x^*\|} \le L\lim_{k\to\infty}\frac{\|x_{k+1}-x^*\|}{\|x_k-x^*\|} = 0. \qquad (63)$$

S použitím (56)-(58) a (61) dostaneme

$$\begin{aligned}
\|x_k-x^*\| &\le \|x_{k+1}-x_k\|+\|x_{k+1}-x^*\| \\
&\le \|A_k^{-1}\|\|A_kd_k+f(x_k)\|+\|A_k^{-1}\|\|f(x_k)\|+\|x_{k+1}-x^*\| \\
&\le \frac{5}{4}c(1+\omega)\|f(x_k)\|+\frac{1}{2}\|x_k-x^*\|,
\end{aligned}$$

neboli

$$\|x_k-x^*\| \le \frac{5}{2}c(1+\omega)\|f(x_k)\|,$$

takže podle (63) platí

$$\lim_{k\to\infty}\frac{\|f(x_{k+1})\|}{\|f(x_k)\|} \le \frac{5}{2}c(1+\omega)\lim_{k\to\infty}\frac{\|f(x_{k+1})\|}{\|x_k-x^*\|} = 0$$

a $\{f(x_k)\}$ konverguje k nule Q-superlineárně. □

Věta 155 říká, že nepřesná Newtonova metoda (56)-(58) 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 (56) nahradit výběrem délky kroku. V následujícím algoritmu se pro výběr délky kroku používá funkce

$$\varphi(x) = \frac{1}{2} f^T(x) f(x)$$

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 $\varphi(x) \leq \varepsilon$, ukončíme výpočet. V opačném případě zvolíme $J_k \in \partial_B f(x_k)$ a určíme směrový vektor d_k tak, aby platilo

$$\omega_k = \frac{\|J_k d_k + f(x_k)\|}{\|f(x_k)\|} \leq \omega. \tag{64}$$

Krok 3 (Délka kroku). Nechť $t_k = \varrho^{i_k}$, kde i_k je nejmenší nezáporné celé číslo i vyhovující podmínce

$$\varphi(x_k + \varrho^i d_k) - \varphi(x_k) \leq -2\sigma\varrho^i \varphi(x_k). \tag{65}$$

Krok 4 (Aktualizace). Položíme $x_{k+1} := x_k + t_k d_k$ a $k := k+1$. Přejdeme na Krok 2.

Věta 156 *Nechť množina $X = \{x \in R^n : \varphi(x) \leq \varphi(x_1)\}$ je kompaktní, nechť zobrazení $f : R^n \to R$ je polohladké a silně BD-regulární na $X \subset R^n$ a funkce $\varphi(x)$ je spojitě diferencovatelná na $X \subset R^n$. Pak:*

(a) Každý hromadný bod posloupnosti $\{x_k\}$, generovaný Algoritmem 4.1, je řešením rovnice (55).

(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 (64) platí

$$\|d_k\| = \|J_k^{-1}(J_k d_k + f(x_k)) - J_k^{-1} f(x_k)\| \leq (1+\omega)\|J_k^{-1}\|\|f_k\| \leq c(1+\omega)\sqrt{2\varphi(x_1)}. \tag{66}$$

Ukážeme, že i Krok 3 algoritmu je dobře definován. Předpokládejme naopak, že pro libovolný exponent i platí

$$\varphi(x_k + \varrho^i d_k) - \varphi(x_k) > -2\sigma\varrho^i \varphi(x_k),$$

neboli v limitě

$$\varphi'(x_k, d_k) \geq -2\sigma\varphi(x_k).$$

Jelikož φ je spojitě diferencovatelná, podle Důsledku 16 a podle (64) platí

$$\begin{aligned}\varphi'(x_k,d_k) &= (\nabla\varphi(x_k))^T d_k = f^T(x_k)J_k d_k \\ &= f^T(x_k)f(x_k) + f^T(x_k)J_k d_k - f^T(x_k)f(x_k) \\ &\leq \|f(x_k)\|\|f(x_k)+J_k d_k\| - \|f(x_k)\|^2 \\ &\leq (\omega-1)\|f(x_k)\|^2 = -2(1-\omega)\varphi(x_k). \end{aligned} \tag{67}$$

Jelikož platí $\varphi(x_k) \neq 0$ (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 (65) platí

$$\begin{aligned}\varphi(x_1) &\geq \varphi(x_1) - \lim_{k\to\infty}\varphi(x_k) = \sum_{k=1}^{\infty}(\varphi(x_k)-\varphi(x_{k+1})) \\ &\geq \sum_{k=1}^{\infty} 2\sigma t_k\varphi(x_k) \geq 2\tau\sigma\sum_{k\in K}\varphi(x_k),\end{aligned}$$

takže nutně $\varphi(x_k) \overset{K}{\to} 0$, což spolu s $x_k \overset{K}{\to} x^*$ dává $\varphi(x^*) = 0$ (neboť funkce φ 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 (65) neplatí pro $i = i_k - 1$, můžeme s použitím věty o střední hodnotě psát

$$(\nabla\varphi(x'_k))^T d_k = \frac{\varphi\left(x_k + \frac{t_k}{\varrho}d_k\right) - \varphi(x_k)}{\frac{t_k}{\varrho}} > -2\sigma\varphi(x_k),$$

kde $x'_k \in (x_k, x_k + (t_k/\varrho)d_k)$. Jelikož posloupnost $\{\|d_k\|\}_{K_1}$ je podle (66) 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

$$(\nabla\varphi(x^*))^T d^* \geq -2\sigma\varphi(x^*).$$

Z druhé strany podle (67) platí $(\nabla\varphi(x_k))^T d_k \leq -2(1-\omega)\varphi(x_k)$, což v limitě dává

$$(\nabla\varphi(x^*))^T d^* \leq -2(1-\omega)\varphi(x^*).$$

Jelikož podle předpokladu platí $\sigma < 1-\omega$, dostaneme porovnáním obou nerovností $\varphi(x^*) = 0$.

Dokázali jsme tedy, že pokud x^* je hromadným bodem posloupnosti generované algoritmem, platí $\varphi(x^*) = 0$ a tedy i $f(x^*) = 0$.
(b) Nechť K je indexová množina použitá v části (a) důkazu. Naším 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 155. 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 155 ($x_k \in B(x^*,\varepsilon)$ a $\omega_k \leq 1/(10cL)$), takže pro bod $x_k + d_k$ platí (61) (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

$$\begin{aligned} \frac{\varphi(x_k + d_k) - \varphi(x_k)}{\varphi(x_k)} &= \frac{(\|f(x_k + d_k)\| - \|f(x_k)\|)\,(\|f(x_k + d_k)\| + \|f(x_k)\|)}{\|f(x_k)\|^2} \\ &\leq \frac{\|f(x_k + d_k)\| - \|f(x_k)\|}{\|f(x_k)\|} \leq -2\sigma, \end{aligned}$$

takže podmínka (65) je splněna s $i_k = 0$. Platí tedy $x_{k+1} = x_k + d_k$ a vzhledem k (61) můžeme množinu K formálně doplnit o index $k+1$. Pokračujeme-li takto pro další hodnoty indexu, vidíme (tak jako v důkazu Věty 155), že $x_k \to x^*$ superlineárně. □

Poznámka 181 Požadavek spojité diferencovatelnosti funkce $\varphi = (1/2)f^T f$ 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 182 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 $J_k = \mathcal{J}f(x_k)$. Pokud $x_k \in \Omega_f$, bývá výpočet $J_k \in \partial_B f(x_k)$ obtížnější. Z Definice 72 plyne, že

$$\partial_B f(x_k) \subset [\partial_B f_1(x_k), \ldots, \partial_B f_n(x_k)]^T \stackrel{\Delta}{=} \partial_b f(x_k),$$

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 20 polohladké, podle Důsledku 22 platí

$$\begin{gathered} f_1(x + h) - f_1(x) - e_1^T Jh = o(\|h\|), \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ f_n(x + h) - f_1(x) - e_n^T Jh = o(\|h\|) \end{gathered}$$

a n je konečné, zůstává klíčový vztah (54) v platnosti i pro $J \in \partial_b f(x)$ a v důkazech Věty 155 a Věty 156 se v podstatě nic nezmění.

12.2 Aplikace nehladkých rovnic

Definice 76 *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*

$$x_i^* \geq 0, \qquad p_i(x^*) \geq 0, \qquad x_i^* p_i(x^*) = 0 \tag{68}$$

pro libovolný index $1 \leq i \leq n$.

Úlohu nelineární komplementarity lze snadno převést na řešení ekvivalentní soustavy polohladkých rovnic $f(x) = 0$, kde

$$f(x) = \begin{bmatrix} \psi(x_1, p_1(x)) \\ \ldots\ldots\ldots\ldots \\ \psi(x_n, p_n(x)) \end{bmatrix} \tag{69}$$

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 19 (neboť $\min(u_1,u_2) = -\max(-u_1,-u_2)$). Nevýhodou Pangovy funkce je to, že není zaručena spojitá diferencovatelnost zobrazení $\varphi = (1/2)f^Tf$, které se používá při výběru délky kroku. Výhodnější vlastnosti má Fischerova-Burmeisterova funkce

$$\psi(u) = \sqrt{u_1^2+u_2^2} - (u_1+u_2). \tag{70}$$

Lemma 41 *Funkce $\psi : R^2 \to R$ definovaná vztahem (70) 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 $u_1u_2 = 0$. 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

$$\nabla\psi(u) = \begin{bmatrix} \frac{u_1}{\sqrt{u_1^2+u_2^2}} - 1 \\ \frac{u_2}{\sqrt{u_1^2+u_2^2}} - 1 \end{bmatrix}. \tag{71}$$

Polohladkost funkce ψ v bodě 0 plyne z Věty 150, 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

$$\psi(u) = \sqrt{u_1^2+u_2^2} - (u_1+u_2) < \sqrt{u_1^2+2u_1u_2+u_2^2} - (u_1+u_2) = 0.$$

Pokud $u_1 = 0$ a $u_2 > 0$, platí

$$\psi(u) = |u_2| - u_2 = 0$$

(stejný výsledek dostaneme pro $u_1 > 0$ a $u_2 = 0$). Rovnost $\psi(0) = 0$ je zřejmá. Druhou mocninu funkce ψ můžeme vyjádřit ve tvaru

$$\psi^2(u) = u_1^2 + u_2^2 + (u_1+u_2)^2 - 2(u_1+u_2)\sqrt{u_1^2+u_2^2}.$$

Tato funkce je spojitě diferencovatelná v $R^2\backslash\{0\}$ a je spojitě diferencovatelná v bodě 0 právě tehdy, je-li funkce $\overline{\psi}(u) = (u_1+u_2)\sqrt{u_1^2+u_2^2}$ spojitě diferencovatelná v bodě 0. Ale

$$\lim_{\|u\|\to 0} \frac{\overline{\psi}(u) - \overline{\psi}(0)}{\|u\|} = \lim_{\|u\|\to 0} (u_1+u_2)\frac{\sqrt{u_1^2+u_2^2}}{\sqrt{u_1^2+u_2^2}} = 0,$$

takže $\overline{\psi}$ je diferencovatelná v bodě 0 a platí $\nabla\overline{\psi}(0) = 0$. Spojitost parciální derivace $\partial\overline{\psi}/\partial u_1$ v bodě 0 plyne z nerovnosti

$$\left|\frac{\partial\overline{\psi}(u)}{\partial u_1}\right| = \left|\frac{u_1}{\sqrt{u_1^2+u_2^2}}(u_1+u_2)+\sqrt{u_1^2+u_2^2}\right|$$
$$\leq \frac{|u_1|}{\sqrt{u_1^2+u_2^2}}|u_1+u_2|+\sqrt{u_1^2+u_2^2} \leq |u_1+u_2|+\sqrt{u_1^2+u_2^2}$$

a z toho, že pravá strana této nerovnosti konverguje k nule pokud $u \to 0$ (stejný výsledek dostaneme pro parciální derivaci $\partial\overline{\psi}/\partial u_2$). □

Věta 157 *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 (69), kde $\psi : R^2 \to R$ je funkce definovaná předpisem (70). 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

$$\partial_B f_i(x) = \nabla f_i(x) = \left(\frac{x_i}{\sqrt{x_i^2+p_i^2(x)}}-1\right)e_i + \left(\frac{p_i(x)}{\sqrt{x_i^2+p_i^2(x)}}-1\right)\nabla p_i(x), \tag{72}$$

pokud $x_i^2 + p_i^2(x) \neq 0$ a

$$\partial_B f_i(x) = \bigcup_{\varphi\in[0,2\pi]} [(\cos\varphi - 1)e_i + (\sin\varphi - 1)\nabla p_i(x)], \tag{73}$$

pokud $x_i^2 + p_i^2(x) = 0$.

(c) Funkce $\varphi = (1/2)f^T f$ je spojitě diferencovatelná v bodě x.

Důkaz (a) Polohladkost zobrazení f plyne z Věty 148 a Věty 151, neboť $f_i(x) = \psi(x_i, p_i(x))$, funkce ψ je polohladká podle Lemmatu 41 a zobrazení p je spojitě diferencovatelné.
(b) Podle Lemmatu 41 je funkce $\psi(x_i, p_i)$ spojitě diferencovatelná, pokud $x_i^2 + p_i^2 \neq 0$. Vztah (72) plyne z (71) 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 41, 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 (73).
(c) Platí

$$\varphi(x) = \frac{1}{2}f^T(x)f(x) = \frac{1}{2}\sum_{i=1}^{n}\psi^2(x_i, p_i(x)).$$

Zobrazení p je spojitě diferencovatelné. Podle Lemmatu 41 je druhá mocnina funkce ψ spojitě diferencovatelná, takže i funkce φ je spojitě diferencovatelná. □

Věta 157 naznačuje jednu z možností jak řešit úlohy nelineární komplementarity. Úloha nelineární komplementarity se převede na ekvivalentní soustavu nehladkých rovnic (69), které se řeší pomocí Algoritmu 4.1. Podle Poznámky 182 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 (72)-(73). Funkce $\varphi = (1/2)f^T f$ 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

$$\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) = 0, \tag{74}$$

$$\left.\begin{array}{ll} -c_i(x) \geq 0, & \lambda_i \geq 0, \\ \lambda_i c_i(x) = 0, & 1 \leq i \leq m \end{array}\right\} \tag{75}$$

(Věta **??**). Podmínky (75) jsou v podstatě podmínkami nelineární komplementarity (68). Můžeme tedy sestavit soustavu $n + m$ nehladkých rovnic

$$F(x, \lambda) \triangleq \left[\begin{array}{l} \nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla c_i(x) \\ \psi(\lambda_1, -c_1(x)) \\ \ldots\ldots\ldots\ldots\ldots\ldots \\ \psi(\lambda_m, -c_m(x)) \end{array}\right] = 0, \tag{76}$$

kde ψ je Fischerova-Burmeisterova funkce (70). Funkce $F : R^{n+m} \to R^{n+m}$ je polohladká a funkce $\varphi = (1/2)F^T F$ je spojitě diferencovatelná, takže soustavu rovnic (76) lze řešit pomocí Algoritmu 4.1.

13 Metody pro nehladkou optimalizaci

13.1 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

$$f_j = f(y_j), \quad g_j \in \partial f(y_j), \tag{77}$$

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

$$f_j^k = f_j + g_j^T(x_k - y_j), \tag{78}$$

$$\alpha_j^k = f(x_k) - f_j^k \tag{79}$$

pro $j \in \mathcal{J}_k$. Tato po částech lineární funkce je v konvexním případě majorizována funkcí f.

Věta 158 *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 129 (d) $f(x) \geq f_j + g_j^T(x - y_j)$ $\forall j \in \mathcal{J}_k$, takže podle (78) dostaneme $f(x_k) \geq f_j^k$, což podle (79) 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 158 neplatí. Abychom v tomto případě zaručili vhodnost po částech lineárního modelu $f_L^k(x)$, je třeba čísla α_j^k, $j \in \mathcal{J}_k$, definovat jiným způsobem. Jednou z možností je pro $j \in \mathcal{J}_k$ položit

$$\alpha_j^k = \max\left(|f(x_k) - f_j^k|, \gamma\|x_k - y_j\|^\nu\right),$$

kde $\gamma \geq 0$ a $\nu \geq 1$. Jelikož by však bylo nutné ukládat body y_j, $j \in \mathcal{J}_k$, využívá se toho, že pro $j \in \mathcal{J}_k$ platí

$$\|x_k - y_j\| \leq \|x_j - y_j\| + \sum_{i=j}^{k-1} \|x_{i+1} - x_i\| \triangleq s_j^k \tag{80}$$

a čísla α_j^k se určují podle vzorce

$$\alpha_j^k = \max\left(|f(x_k) - f_j^k|, \gamma(s_j^k)^\nu\right), \quad j \in \mathcal{J}_k. \tag{81}$$

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 věrohodnosti 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 (88), (94), (95) a transformační vztahy (99)). 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

$$\alpha_a^k = \max\left(|f(x_k) - f_a^k|, \gamma(s_a^k)^\nu\right). \tag{82}$$

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

$$\frac{1}{2}d^T G_k d + v \tag{83}$$

na množině určené omezeními

$$\begin{aligned} -\alpha_j^k + d^T g_j &\leq v, \quad j \in \mathcal{J}_k, &(84)\\ -\alpha_a^k + d^T g_a^k &\leq v, &(85)\end{aligned}$$

(minimalizuje se přes všechny dvojice $(d, v) \in R^{n+1}$ vyhovující nerovnostem (84), (85)).

Věta 159 *Řešení úlohy (83)-(85) lze vyjádřit ve tvaru*

$$\begin{aligned} d_k &= -G_k^{-1}\tilde{g}_a^k, &(86)\\ v_k &= -d_k^T G_k d_k - \tilde{\alpha}_a^k, &(87)\end{aligned}$$

kde

$$\begin{aligned} \tilde{g}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k g_j + \lambda_a^k g_a^k, &(88)\\ \tilde{\alpha}_a^k &= \sum_{j\in\mathcal{J}_k} \lambda_j^k \alpha_j^k + \lambda_a^k \alpha_a^k &(89)\end{aligned}$$

a kde Lagrangeovy multiplikátory λ_j^k, $j \in \mathcal{J}_k$, λ_a^k, jsou řešením duální úlohy kvadratického programování: Minimalizovat funkci

$$\frac{1}{2}\left(\sum_{j\in\mathcal{J}_k} \lambda_j g_j + \lambda_a g_a^k\right)^T G_k^{-1} \left(\sum_{j\in\mathcal{J}_k} \lambda_j g_j + \lambda_a g_a^k\right) + \sum_{j\in\mathcal{J}_k} \lambda_j \alpha_j^k + \lambda_a \alpha_a^k \tag{90}$$

na množině určené omezeními

$$\left.\begin{array}{l} \lambda_j \geq 0, \quad j \in \mathcal{J}_k, \quad \lambda_a \geq 0, \\ \sum_{j\in\mathcal{J}_k} \lambda_j + \lambda_a = 1. \end{array}\right\} \tag{91}$$

Minimální hodnota funkce (90), odpovídající řešení úlohy (90)-(91), je

$$w_k = \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1} \tilde{g}_a^k + \tilde{\alpha}_a^k = -v_k - \frac{1}{2}(\tilde{g}_a^k)^T G_k^{-1} \tilde{g}_a^k. \tag{92}$$

Důkaz Jelikož matice G_k je positivně definitní, je funkce (83) konvexní. Omezení (84)-(85) jsou lineární a tudíž také konvexní, takže pár $(d_k, v_k) \in R^{n+1}$ je podle Věty **??** řešením úlohy (83)-(85) právě tehdy, existují-li Lagrangeovy multiplikátory $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$, takové, že

$$\begin{bmatrix} G_k d_k \\ 1 \end{bmatrix} + \sum_{j\in\mathcal{J}_k} \lambda_j^k \begin{bmatrix} g_j \\ -1 \end{bmatrix} + \lambda_a^k \begin{bmatrix} g_a^k \\ -1 \end{bmatrix} = 0, \tag{93}$$

přičemž

$$\begin{aligned} \lambda_j^k > 0 &\Rightarrow -\alpha_j^k + d_k^T g_j = v_k,\\ \lambda_a^k > 0 &\Rightarrow -\alpha_a^k + d_k^T g_a^k = v_k \end{aligned}$$

(podmínky komplementarity). Z poslední rovnice soustavy (93) dostaneme

$$\sum_{j\in\mathcal{J}_k} \lambda_j^k + \lambda_a^k = 1.$$

Platí tedy (86) (88) a (91). Použijeme-li označení (88)-(89) a podmínky komplementarity, můžeme psát

$$-\tilde{\alpha}_a^k + d_k^T \tilde{g}_a^k = v_k,$$

což spolu s (86) dává (87). 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 (90)-(91). 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 (91)) a $\mu_j^k \geq 0$, $j \in \mathcal{J}_k$, $\mu_a^k \geq 0$ (odpovídající nerovnostem v (91)) 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 (86) a (88)). Poslední rovnosti však nejsou nic jiného než nerovnosti (84), (85), 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 (83)-(85). □

Poznámka 183 Poznamenejme, že omezení (85) není třeba používat pokud $\mathcal{J}_k = \{1, \ldots, k\}$, neboť je v tomto případě lineární kombinací omezení (84). Pak ale $\lambda_a^k = 0$ v (88)-(89).

Poznámka 184 Kromě agregovaných gradientů (88) se pomocí Lagrangeových multiplikátorů $\lambda_j^k \geq 0$, $j \in \mathcal{J}_k$, $\lambda_a^k \geq 0$ definují agregované hodnoty

$$\tilde{f}_a^k = \sum_{j\in\mathcal{J}_k} \lambda_j^k f_j^k + \lambda_a^k f_a^k, \tag{94}$$

$$\tilde{s}_a^k = \sum_{j\in\mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k. \tag{95}$$

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 77 a Definici 78. V obou definicích používáme označení

$$\beta_{k+1} = \max\left(|f(x_k) - f_{k+1} - (x_k - y_{k+1})^T g_{k+1}|, \gamma|x_k - y_{k+1}|^\nu\right) \tag{96}$$

a konstanty $0 < \sigma_L < \sigma_T < \sigma_R < 1$, $0 < \sigma_A < \sigma_R - \sigma_T$, $0 < \tau < 1$ a $D > 0$.

Definice 77 *(Spádový krok) Spádovým krokem nazveme krok, ve kterém platí $t_R^k = t_L^k > 0$,*

$$f(x_{k+1}) \leq f(x_k) - \sigma_L t_L^k w_k \tag{97}$$

a buď $t_L^k \geq \tau$ nebo $\beta_{k+1} > \sigma_A w_k$.

Definice 78 *(Nulový krok) Nulovým krokem nazveme krok, ve kterém platí* $t_R^k > t_L^k = 0$,

$$d_k^T g_{k+1} \geq \beta_{k+1} - \sigma_R w_k \tag{98}$$

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

$$\left.\begin{array}{ll} f_j^{k+1} = f_j^k + (x_{k+1} - x_k)^T g_j, & j \in J_k \\ f_a^{k+1} = f_a^k + (x_{k+1} - x_k)^T \tilde{g}_a^k & \\ f_{k+1}^{k+1} = f_{k+1} + (x_{k+1} - y_{k+1}) g_{k+1} & \\ g_a^{k+1} = \tilde{g}_a^k & \\ s_j^{k+1} = s_j^k + \|x_{k+1} - x_k\|, & j \in J_k \\ s_a^{k+1} = \tilde{s}_a^k + \|x_{k+1} - x_k\| & \\ s_{k+1}^{k+1} = \|x_{k+1} - y_{k+1}\| & \end{array}\right\} \tag{99}$$

Zbývá uvést podmínky, které by měly splňovat matice G_k. Abychom zaručili globální konvergenci svazkové metody, použijeme tento předpoklad.

Předpoklad 13.1 *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 5.1

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 (83)-(85) (omezení (85) 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 159). 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 77) nebo nulový krok (Definice 78). 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 (99) a určíme matici G_{k+1} tak, aby vyhovovala Předpokladu 13.1. 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 185 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 5.2

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 $t^1 = 1$, $t^1_A = 0$, $t^1_U = 1$ 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 $t^i_A = t^i$. V opačném případě položíme $t^i_U = t^i$.

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 $t_R = t_L = t^i$, $t_A = t^i_A$, $\beta^+ = \beta^i$ a ukončíme výpočet.

Krok 4 (Nulový krok). Jestliže $d^T g^i \geq \beta^i - \sigma_R w$ a $(t^i - t^i_A)\|d\| \leq D$, položíme $t_R = t^i$, $t_L = 0$, $t_A = t^i_A$, $\beta^+ = \beta^i$ a ukončíme výpočet.

Krok 5 (Aktualizace). Zvolíme $t^{i+1} \in [t^i_A + \kappa(t^i_U - t^i_A), t^i_U - \kappa(t^i_U - t^i_A)]$, položíme $i := i + 1$ a přejdeme na Krok 2.

Věta 160 *Nechť funkce $f : R^n \to R$ je lokálně lipschitzovská a nechť pro libovolnou posloupnost $t^i \downarrow 0$ platí*

$$\limsup_{\substack{g^i \in \partial f(x + t^i d) \\ i \to \infty}} d^T g^i \geq \liminf_{i \to \infty} \frac{f(x + t^i d) - f(x)}{t^i}. \tag{100}$$

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

$$f(x+t^*d) \le f - \sigma_T t^* w, \tag{101}$$

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 $t_U^{\bar{i}} = t_U^i \downarrow t^*\ \forall i > \bar{i}$, neboli $t^* = t_U^{\bar{i}} \notin S$, 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 (101) 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 (100), dostaneme

$$-\sigma_T w \le \liminf_{i \overset{I}{\to} \infty} \frac{f(x+t^*d+(t^i-t^*)d) - f(x+t^*d)}{t^i - t^*} \le \limsup_{i \overset{I}{\to} \infty} d^T g^i. \tag{102}$$

Vyšetříme nyní dva případy.
(a) Nechť $t^* > 0$. Podle (101) pro dostatečně velké indexy platí $f(x+t^i d) \le f - \sigma_L t^i w$, 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 $\beta^i \le \sigma_A w$ (Krok 3 algoritmu) a $\beta^i - d^T g^i > \sigma_R w$ (Krok 4 algoritmu), což dohromady dává

$$d^T g^i < \beta^i - \sigma_R w \le -(\sigma_R - \sigma_A) w < -\sigma_T w$$

(neboť $w > 0$) a což je pro $i \in I$ (I je nekonečná) ve sporu s (102).
(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 135 (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

$$\limsup_{i \overset{I}{\to} \infty} d^T g^i \le -\sigma_R w < -\sigma_T w,$$

což je opět ve sporu s (102). □

Poznámka 186 Podle Věty 147 splňuje podmínku (100) každá slabě polohladká funkce, neboť výraz na pravé straně (100) je v tomto případě směrovou derivací (která existuje) a výraz na levé straně je roven limitě (46).

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 13.2 *Funkce $f : R^n \to R$ je lokálně lipschitzovská na množině $X + \overline{B(0,D)}$, kde množina $X = \{x \in R^n : f(x) \le f(x_1)\}$ je kompaktní, a je splněna podmínka (100) (například, když f je slabě polohladká).*

Poznámka 187 Protože ve spádových krocích hodnota funkce f neroste, platí $x_k \in X$ a protože X je kompaktní, je posloupnost $\{x_k\}$ omezená. Jelikož podle Věty 160 platí $\|y_k - z_k\| \le D$, kde $z_k \in X$, můžeme psát $y_k \in X + \overline{B(0,D)}$. Množina $X + \overline{B(0,D)}$ je kompaktní, takže posloupnost $\{y_k\}$ je omezená. Z lokální lipschitzovskosti funkce f na $X + \overline{B(0,D)}$ plyne omezenost posloupnosti $\{g_k\}$. Podle (103) je i posloupnost $\{\tilde{g}_a^k\}$ omezená. Z (86) a Předpokladu 13.1 pak plyne omezenost posloupnosti $\{d_k\}$.

Lemma 42 *Existují čísla $\tilde{\lambda}_i^k \ge 0$, $1 \le i \le k$, $\tilde{\lambda}_1^k + \ldots + \tilde{\lambda}_k^k = 1$ taková, že hodnoty $\tilde{f}_a^k$, $\tilde{g}_a^k$, $\tilde{s}_a^k$ získané v Kroku 2 Algoritmu 5.1 vyhovují vztahům*

$$\left(\tilde{f}_a^k, \tilde{g}_a^k, \tilde{s}_a^k\right) = \sum_{i=1}^{k} \tilde{\lambda}_i^k (f_i^k, g_i, s_i^k) \tag{103}$$

(závorky v (103) 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 (103) (platí to zřejmě pokud $\mathcal{J}_k = \{1, \ldots, k\}$, kdy $\lambda_a^k = 0$, takže vztahy (88), (94), (95) implikují (103) 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 (83)-(85) (nebo úlohy (90)-(91)), 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 (91) 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 (99), (88), (94), (95) 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 43 *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 $x_k \xrightarrow{K} x^*$ a $w_k \xrightarrow{K} 0$, pak bod x^* je stacionárním bodem funkce f (platí $0 \in \partial f(x^*)$).*

Důkaz Podle Lemmatu 42 platí (103). Podle Věty 107 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

$$(\tilde{g}_a^k, \tilde{s}_a^k) = \sum_{i=1}^{n+2} \lambda^{k,i}\left(g^{k,i}, s^{k,i}\right), \tag{104}$$

kde $\lambda^{k,i} \geq 0$, $1 \leq i \leq n+2$, $\lambda^{k,1} + \ldots + \lambda^{k,n+2} = 1$. Podle Poznámky 187 jsou vektory $y^{k,i}$, $g^{k,i}$, $1 \leq i \leq n+2$, omezené, takže existuje podmnožina $\overline{K} \subset K$ taková, že $y^{k,i} \xrightarrow{\overline{K}} y_i^*$, $g^{k,i} \xrightarrow{\overline{K}} g_i^*$, $\lambda^{k,i} \xrightarrow{\overline{K}} \lambda_i^*$, $1 \leq i \leq n+2$. Podle Věty 135 (c) platí $g_i^* \in \partial f(y_i^*)$, $1 \leq i \leq n+2$. Z (104) pak plyne $(\tilde{g}_a^k, \tilde{s}_a^k) \to (\tilde{g}_a^*, \tilde{s}_a^*)$, kde

$$(\tilde{g}_a^*, \tilde{s}_a^*) = \sum_{i=1}^{n+2} \lambda_i^*(g_i^*, s_i^*) \tag{105}$$

a $\lambda_i^* \geq 0$, $1 \leq i \leq n+2$, $\lambda_1^* + \ldots + \lambda_{n+2}^* = 1$. Navíc (80) implikuje $s^{k,i} \geq \|x_k - y^{k,i}\|$, což spolu s $x_k \xrightarrow{\overline{K}} x^*$, $y^{k,i} \xrightarrow{\overline{K}} y_i^*$ a $s^{k,i} \xrightarrow{\overline{K}} s_i^*$ dává

$$s_i^* \geq \|x^* - y_i^*\| \tag{106}$$

pro $1 \leq i \leq n+2$. Jelikož $w_k \overset{K}{\to} 0$, matice G_k jsou stejnoměrně pozitivně definitní a $\tilde{\alpha}_a^k \geq 0$, musí podle (92) platit $\tilde{g}_a^k \overset{K}{\to} 0$, $\tilde{\alpha}_a^k \overset{K}{\to} 0$. Podle (81), (82) a (89) dostaneme

$$\begin{aligned}
\tilde{\alpha}_a^k &= \sum_{j \in \mathcal{J}_k} \lambda_j^k \max\left(|f(x_k) - f_j^k|, \gamma(s_j^k)^\nu\right) + \lambda_a^k \max\left(|f(x_k) - f_a^k|, \gamma(s_a^k)^\nu\right) \\
&\geq \max\left(\sum_{j \in \mathcal{J}_k} \lambda_j^k |f(x_k) - f_j^k| + \lambda_a^k |f(x_k) - f_a^k|, \gamma\left(\sum_{j \in \mathcal{J}_k} \lambda_j^k s_j^k + \lambda_a^k s_a^k\right)^\nu\right) \\
&\geq \max\left(|f(x_k) - \tilde{f}_a^k|\right), \gamma\left(\tilde{s}_a^k\right)^\nu), \qquad (107)
\end{aligned}$$

neboť funkce $\max(\cdot,\cdot)$ a $|\cdot|^\nu$, $\nu \geq 1$, jsou konvexní. Platí tedy $\tilde{g}_a^k \overset{K}{\to} 0$, $\tilde{s}_a^k \overset{K}{\to} 0$, což s použitím (105) a (106) 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 188 *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 43. Platí $\tilde{g}_a^k = 0$ a $\tilde{s}_a^k = 0$, což jako v důkazu Lemmatu 43 dává $0 \in \partial f(x_k)$.*

Lemma 44 *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 (99) a (82) 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 (107) dává $\alpha_a^{k+1} \leq \tilde{\alpha}_a^k$. Nechť $0 \leq \lambda \leq 1$. Označme

$$\begin{aligned}
g_{k+1}(\lambda) &= \lambda g_{k+1} + (1-\lambda) g_a^{k+1} = \lambda g_{k+1} + (1-\lambda)\tilde{g}_a^k \overset{\Delta}{=} \tilde{g}_k(\lambda), \\
\alpha_{k+1}(\lambda) &= \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\alpha_a^{k+1} \leq \lambda \alpha_{k+1}^{k+1} + (1-\lambda)\tilde{\alpha}_a^k \overset{\Delta}{=} \tilde{\alpha}_k(\lambda).
\end{aligned}$$

Vzhledem k tomu, že w_{k+1} je podle Věty 159 minimem funkce (90) (s indexem $k+1$ místo k), musí pro $k > l$ platit

$$w_{k+1} \leq \frac{1}{2} g_{k+1}^T(\lambda) G_{k+1}^{-1} g_{k+1}(\lambda) + \alpha_{k+1}(\lambda) \leq \frac{1}{2} \tilde{g}_k^T(\lambda) G_k^{-1} \tilde{g}_k(\lambda) + \tilde{\alpha}_k(\lambda) \overset{\Delta}{=} w_k(\lambda),$$

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 13.1). Dále poznamenejme, že pro $k > l$ z (86) a (98) 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 (96) 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 187) a ze stejnoměrné pozitivní definitnosti matic G_k (Předpoklad 13.1). 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

$$
w_{k+1} \leq w_k - \frac{(1-\sigma_R)^2 w_k^2}{4M}. \tag{108}
$$

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 (108) nerostoucí pro $k > l$). Pak bychom z (108) 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 43. □

Věta 161 *Nechť funkce $f : R^n \to R$ splňuje Předpoklad 13.2. 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 44 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 (97), 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 160 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 43 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 13.1 a z omezenosti směrových vektorů (Poznámka 187) 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 (96) 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 43 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 13.1). 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 13.1. 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 (83)-(85).