NUMERICKÉ OPTIMALIZAČNÍ METODY PRO ÚLOHY BEZ OMEZUJÍCÍCH PODMÍNEK

(učební text)

L. Lukšan

Výzkumná zpráva č. V-640, květen 1995
Tato práce byla podpořena grantem č. 201/93/0429 poskytnutým grantovou agenturou ČR

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 pak popsány speciální metody pro rozsáhlé a strukturované optimalizační úlohy. Věty a lemata jsou v této práci vždy dokazovány. Tvrzení z příbuzných oborů, která lze nalézt v běžných učebních textech, jsou uváděny bez důkazu. Většinu chybějících důkazů lze nalézt v knize: L.Lukšan, Metody s proměnnou metrikou, Academia, Praha 1991.

1.1. Základní pojmy

Budeme používat označování:

$$x \in R^n \text{ pro } n - \text{dimensionální vektor}$$

$$F(x) \text{ pro funkci } F : R^n \to R$$

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

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

Zde $F(x)$ je účelová funkce, $g(x)$ je její gradient a $G(x)$ je její Hessova matice (matice druhých parciálních derivací). 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 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 $\forall \overline{F}$ (prázdná množina se předpokládá kompaktní).

Definice 3 Řekneme, že funkce $F : R^n \to R$ má omezené druhé derivace, jestliže platí

$$|d^T G(x)d| \leq \overline{G} \parallel d \parallel^2 \tag{F3}$$

$\forall x \in R^n$, $\forall d \in R^n$. Je to ekvivalentní podmínce $\parallel G(x) \parallel \leq \overline{G}$ $\forall x \in R^n$.

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

$$\parallel g(x+d) - g(x) \parallel \leq \overline{G} \parallel d \parallel$$

$\forall x \in R^n$, $\forall d \in R^n$

Definice 4 Řekneme, že funkce $F : R^n \to R$ je stejnoměrně (nebo silně) konvexní, jestliže platí

$$d^T G(x)d \geq \underline{G} \parallel d \parallel^2 \tag{F4}$$

$\forall x \in R^n$, $\forall d \in R^n$ (zde $\underline{G} > 0$).

Definice 5 Řekneme, že funkce $F : R^n \to R$ má lipschitzovské druhé derivace, jestliže

$$\parallel G(x+d) - G(x) \parallel \leq \overline{L} \parallel d \parallel \tag{F5}$$

$\forall x \in R^n$, $\forall d \in R^n$.

Při konvergenčních důkazech budeme často používat věty o střední hodnotě:

(a) $F(x+d) = F(x) + d^T g(x) + \frac{1}{2} d^T G(\tilde{x})d$, kde $\tilde{x} = x + \tilde{\lambda} d$ a $0 \leq \tilde{\lambda} \leq 1$.
Z (F3) plyne

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

Z (F4) plyne

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

(b) $g(x+d) = g(x) + \int_0^1 G(x + \lambda d) d d\lambda$
z (F3) plyne

$$\parallel g(x+d) - g(x) \parallel \leq \overline{G} \parallel d \parallel$$
$$d^T(g(x+d) - g(x)) \leq \overline{G} \parallel d \parallel^2$$

z (F4) plyne

$$\parallel g(x+d) - g(x) \parallel \geq \underline{G} \parallel d \parallel$$
$$d^T(g(x+d) - g(x)) \geq \underline{G} \parallel d \parallel^2$$

Důkaz posledního tvrzení:

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

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

1.2. Podmínky optimality

Definice 6 Řekneme, že bod $x^* \in R^n$ je lokálním minimem funkce $F : R^n \to R$, jestliže existuje číslo $\varepsilon > 0$ takové, že

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

kde $\mathcal{B}(x^*,\varepsilon) = \{x \in R^n : \| x - x^* \| < \varepsilon\}$. Jestliže navíc $F(x^*) < F(x)$ pokud $x^* \neq x$, řekneme, že bod $x^* \in R^n$ je ostrým lokálním minimem funkce $F : R^n \to R$.

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

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

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

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

(matice $G(x^*)$ je positivně semidefinitní)

Tvrzení 2 Nechť $F : R^n \to R \in C^2$ na $\mathcal{B}(x^*,\varepsilon)$ a nechť platí

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

a

$$G(x^*) > 0$$

(matice $G(x^*)$ je pozitivně definitní). Pak bod $x^* \in R^n$ je ostrým lokálním minimem funkce $F : R^n \to R$.

1.3. Základní pojmy z teorie konvergence

Nyní se budeme zabývat vlastnostmi konvergentních posloupností.
Definice 7 Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů. Jestliže pro libovolné $\varepsilon > 0$ existuje index $k \in N$ tak, že $x_i \in \mathcal{B}(x^*,\varepsilon)$ $\forall k \geq i$, 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 $F_i = F(x_i)$, $g_i = g(x_i)$, $G_i = G(x_i)$.

Věta 3 Nechť $x_i \in R^n$, $d_i \in R^n$, $i \in N$, jsou dvě posloupnosti. Nechť $e_i = x_i - x^*$, $i \in N$, kde $x^* \in R^n$ je lokální minimum funkce $F \in C^2$ která vyhovuje podmínce (F5). 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

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

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

Zde $\| O(\xi_i) \| \leq C \| \xi_i \|$ pokud $\| \xi_i \| \to 0$.

Důkaz Z (F5) a z věty (a) o střední hodnotě plyne

$$\begin{aligned}
F(x_i + d_i) - F(x_i) &= d_i^T g_i + \frac{1}{2} d_i^T (G(x_i + \tilde{\lambda} d_i)) d_i = \\
&= d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{1}{2} d_i^T (G(x_i + \tilde{\lambda} d_i) - G_i) d_i \leq \\
&\leq d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{1}{2} \| G(x_i + \tilde{\lambda} d_i) - G_i \| \| d_i \|^2 \leq \\
&\leq d_i^T g_i + \frac{1}{2} d_i^T G_i d_i + \frac{\tilde{\lambda}}{2} \overline{L} \| d_i \|^3
\end{aligned}$$

kde $0 \leq \tilde{\lambda} \leq 1$. Pobobně z (F5) a z věty (b) o střední hodnotě plyne

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

a

$$\begin{aligned} \| \int_0^1 (G(x_i+\lambda d_i)-G_i)d_i d\lambda \| &\le \int_0^1 \| G(x_i+\lambda d_i)-G_i \| \| d_i \| d\lambda \le \\ &\le \overline{L} \int_0^1 \| d_i \|^2 d\lambda = \overline{L} \| d_i \|^2 \end{aligned}$$

Tím jsme dokázáli 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 $g(x^*) = 0$.

Poznámka 2 Bez podmínky (F5) lze odvodit odhady

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

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

a

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

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

Zde $\| o(\xi_i) \| / \| \xi_i \| \to 0$ pokud $\| \xi_i \| \to 0$.

Definice 8 Nechť $x_i \to x^*$. Jestliže existuje index $k \in N$ a hodnoty $0 < M_k < \infty$ a $0 < q < 1$, tak že

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

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

Věta 4 Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* R-lineárně právě tehdy jestliže

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

($\limsup$ existuje a je menší než 1).

Důkaz Z definice 8 plyne

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

$\forall i \ge k$, kde $q < 1$, takže

$$\| x_i - x^* \|^{\frac{1}{i}} \le (M_k \| x_k - x^* \|)^{\frac{1}{i}} q^{1-\frac{k}{i}} \le \max(1, M_k \| x_k - x^* \|)$$

Tato posloupnost je omezená, takže existuje $\limsup$ a platí

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

Z druhé strany nechť existuje výraz na levé straně poslední nerovnosti a je roven $\hat{q} < 1$. Pak pro libovolné číslo $\hat{q} < q < 1$ existuje index $k \in N$ tak, že platí

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

neboli

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

$\forall i \geq k$. Zvolme

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

Pak platí

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

Definice 9 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 10 Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* 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 11 Posloupnost $x_i \in R^n$, $i \in N$, konverguje k bodu x^* Q-superlineárně, jestliže

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

Věta 5 Nechť $x_i \to x^*$ Q-lineárně (Q-superlineárně). Pak $x_i \to x^*$ R-lineárně (R-superlineárně)

Důkaz R-lineární konvergence plyne v Q-lineární konvergence bezprostředně (stačí volit $M_k = 1$). Nechť $0 < q < 1$ je libovolné (malé) číslo. Z Q-superlineární konvergence plyne existence indexu $k \in N$ takového, že

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

takže podle věty 4 platí

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

Protože q je libovolné (malé) musí platit

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

Poznámka 3 Q-lineární (Q-superlineární) konvergence implikuje monotonnost posloupnosti $\| x_i - x^* \|$, $i \in N$ (počínaje vhodným indexem $k \in N$).

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

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

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

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

Poznámka 4 Někdy se používá slabší předpoklad $\| x_{jl+m} - x^* \| \leq M \| x_{jl} - x^* \|^2$ $\forall j \in N$ (vybraná posloupnost $i = jl$).

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 14 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

$$s_i = -g(x_i)$$

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

Výhody:

- je globálně konvergentní
- používá pouze vektory z R^n
  $O(n)$ - paměťových míst
  $O(n)$ - operací na iteraci

Nevýhody:

- přesný výběr délky kroku
- 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 jestliže $\kappa(G(x^*)) = 10^6$ potřebujeme ke snížení chyby $\| x - x^* \|$ o 8 řádů zhruba 9200000 iterací!

Newtonova metoda je definována vztahy

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

$$\alpha_i = 1$$

Výhody:

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

Nevýhody:

- není globálně konvergentní. Pokud x_1 je daleko od x^*, nemusí konvergovat.
- používá matici a je třeba řešit soustavu lineárních rovnic.
  $O(n^2)$ - 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.

a) Metody spádových směrů vyvinuté z metody největšího spádu:

- nepřesný výběr délky kroku
- urychlení konvergence (princip sdružených směrů),

b) Metody s lokálně omezeným krokem vyvinuté z Newtonovy metody:

- zajištění globální konvergence
- snížení počtu operací (nepřesné řešení lineárních rovnic)

2. Metody spádových směrů

2.1. Základní vlastnosti metod spádových směrů

Klíčový význam pro konstrukci metod spádových směrů má pojem spádových a stejnoměrně spádových směrových vektorů.

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

$$s_i^T g_i < 0 \qquad (\overline{S1})$$

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

$$-s_i^T g_i \ge \varepsilon_0 \parallel s_i \parallel \parallel g_i \parallel \qquad (S1)$$

$\forall i \in N$.

Směrové vektory $s_i \in R^n$, $i \in N$, se často určují nepřesným řešením soustav lineárních rovnic $B_i s_i + g_i = 0$, $i \in N$.

Věta 6 Nechť B_i, $i \in N$, je posloupnost symetrických pozitivně definitních matic takových, že $\underline{B} \le \underline{\lambda}(B_i) \le \overline{\lambda}(B_i) \le \overline{B}$ $\forall i \in N$ ($\underline{\lambda}(B_i)$ a $\overline{\lambda}(B_i)$ je nejmenší a největší vlastní číslo matice B_i). Nechť $\parallel B_i s_i + g_i \parallel \le \overline{\omega} \parallel g_i \parallel$ $\forall i \in N$, kde $\overline{\omega} < \underline{B}/\overline{B}$. Pak je splněna podmínka (S1) s

$$\varepsilon_0 = \frac{\underline{B}/\overline{B} - \overline{\omega}}{1 + \overline{\omega}}$$

Důkaz Označme $r_i = B_i s_i + g_i$, $i \in N$, takže $\parallel r_i \parallel \le \overline{\omega} \parallel g_i \parallel$, $i \in N$. Zřejmě platí $s_i = B_i^{-1}(r_i - g_i)$, $i \in N$, takže

$$-s_i^T g_i = g_i^T B_i^{-1} g_i - r_i^T B_i^{-1} g_i \ge \parallel g_i \parallel^2 / \overline{\lambda}(B_i) - \parallel r_i \parallel \parallel g_i \parallel / \underline{\lambda}(B_i) \ge (1/\overline{B} - \overline{\omega}/\underline{B}) \parallel g_i \parallel^2$$

Dále platí

$$\| s_i \| = \| B_i^{-1}(r_i - g_i) \| \le (\| r_i \| + \| g_i \|)/\underline{\lambda}(B_i) \le (1+\overline{\omega}) \| g_i \| / \underline{B}$$

což po dosazení dává

$$-\frac{s_i^T g_i}{\| s_i \| \| g_i \|} \ge \frac{1/\overline{B} - \overline{\omega}/\underline{B}}{(1+\overline{\omega})/\underline{B}} = \frac{\underline{B}/\overline{B} - \overline{\omega}}{1+\overline{\omega}} \triangleq \varepsilon_0$$

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

Definice 16 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 $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$ (nezávislá na indexu $i \in N$) tak, že

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

a buď

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

nebo

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

nebo

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

nebo $\alpha_i > 0$ je první člen posloupnosti α_i^j, $j \in N$, takové, že $\alpha_i^1 = 1$ a

$$\underline{\beta} \alpha_i^j \le \alpha_i^{j+1} < \alpha_i^j \qquad \forall j \in N \tag{S3d}$$

pro který platí (S2) ($0 < \underline{\beta} < 1$ je nějaká konstanta).

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

Definice 17 Řekneme, že základní optimalizační metoda $x_{i+1} = x_i + \alpha_i s_i$ $i \in N$ je metodou spádových směrů jestliže směrové vektory $s_i \in R^n$, $i \in N$, splňují podmínku (S1) (jsou stejnoměrně spádové) a délky kroku $\alpha_i > 0$, $i \in N$ splňují podmínku (S2) a některou z podmínek (S3a)-(S3d).

Poznámka 5 Pro metody sdružených gradientů odvozené z metody největšího spádu se používá podmínka (S3a). Pro metody s proměnnou metrikou odvozené z Newtonovy metody (kde $\alpha_i \to 1$ pro $i \to \infty$) se používá podmínka (S3b). Pro bezderivační metody se používá podmínka (S3c). Pro nehladké úlohy se používá podmínka (S3d).

Poznámka 6 Podmínka (S1) se často nahražuje nějakou slabší podmínkou (směrové vektory nemusí být stejnoměrně spádové). Důkaz globální konvergence je pak obtížnější, jak je patrné z důkazu věty 14.

Poznámka 7 Metoda největšího spádu je metodou spádových směrů, neboť $s_i = -g_i$, takže $s_i^T g_i = -\| g_i \|^2 = -\| s_i \| \| g_i \|$ a (S1) platí pro $\varepsilon_0 = 1$.

Lemma 7 (Konzistence) Nechť funkce $F : R^n \to R$ splňuje podmínky (F1) a (F3) a směrový vektor s_i splňuje podmínku (S1). Pak pro libovolnou z podmínek (S3a)-(S3d) existuje délka kroku $\alpha_i > 0$, která vyhovuje této podmínce i podmínce (S2).

Důkaz (pro (S3b)). Označme

$$\mathcal{M}_i = \{\alpha \ge 0 : F(x_i + \alpha s_i) - F_i \le \varepsilon_1 \alpha s_i^T g_i\}$$

zřejmě $\mathcal{M}_i \neq \emptyset$ neboť $0 \in \mathcal{M}_i$. Nechť $\alpha_i = \sup \mathcal{M}_i$. Podle (F1) platí $F(x_i + \alpha s_i) \geq \underline{F}$ takže pro $\alpha_i \in \mathcal{M}_i$ dostaneme $\alpha_i \leq (\underline{F} - F_i)/\varepsilon_1 s_i^T g_i < \infty$. Ukážeme nyní, že platí $s_i^T g(x_i + \alpha_i s_i) \geq \varepsilon_1 s_i^T g_i$. Předpokládejme naopak, že

$$s_i^T g(x_i + \alpha_i s_i) = \varepsilon s_i^T g_i$$

kde $\varepsilon > \varepsilon_1$. Potom podle (F3) platí

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

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

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

což je spor neboť $\alpha_i = \sup \mathcal{M}_i$. Z $s_i^T g(x_i + \alpha_i s_i) \geq \varepsilon_1 s_i^T g_i$ a $s_i^T g_i < 0$ plyne $\alpha_i > 0$..

Věta 8 (Globální konvergence) Nechť funkce $F : R^n \to R$ splňuje podmínky (F1) a (F3). Pak metoda spádových směrů je globálně konvergentní.

Důkaz (pro (S3b)) Dokážeme nejprve, že existuje konstanta $\underline{c} > 0$ tak, že pro libovolný index $i \in N$ platí

$$\alpha_i \geq \underline{c} \parallel g_i \parallel / \parallel s_i \parallel \tag{a}$$

a

$$F_{i+1} - F_i \leq -\varepsilon_0 \varepsilon_1 \underline{c} \parallel g_i \parallel \tag{b}$$

Z podmínky (S3b) dostaneme

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

neboli

$$\alpha_i \geq \frac{(\varepsilon_2 - 1) s_i^T g_i}{\overline{G} \parallel s_i \parallel^2} \geq \frac{\varepsilon_0 (1 - \varepsilon_2) \parallel g_i \parallel}{\overline{G} \parallel s_i \parallel}$$

takže platí (a) pro $\underline{c} = \varepsilon_0 (1 - \varepsilon_2)/\overline{G} > 0$. Dále z podmínky (S2) a z (a) dostaneme

$$F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \leq -\varepsilon_0 \varepsilon_1 \alpha_i \parallel s_i \parallel \parallel g_i \parallel \leq -\varepsilon_0 \varepsilon_1 \underline{c} \parallel g_i \parallel^2$$

takže platí (b). Nyní můžeme psát

$$F_{i+1} = F_1 + \sum_{j=1}^{i} (F_{j+1} - F_j) \leq F_1 - \varepsilon_0 \varepsilon_1 \underline{c} \sum_{j=1}^{i} \parallel g_j \parallel^2$$

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

$$\underline{F} \leq \lim_{i \to \infty} F_i \leq F_1 - \varepsilon_0 \varepsilon_1 \underline{c} \sum_{i=1}^{\infty} \parallel g_i \parallel^2$$

takže

$$\sum_{i=1}^{\infty} \parallel g_i \parallel^2 \leq \frac{F_1 - \underline{F}}{\varepsilon_0 \varepsilon_1 \underline{c}}$$

takže nutně $\| g_i \| \to 0$.

Důsledek Metoda největšího spádu je globálně konvergentní.

Poznámka 7 Tvrzení $\| g_i \| \to 0$ je silnější než podmínka v definici 14. K tomu aby metoda spádových směrů byla globálně konvergentní ve smyslu definice 14 stačí, když (S1) platí pro nějakou nekonečnou podmnožinu množiny N.

Poznámka 8 Je-li základní optimalizační metoda globálně konvergentní (ve smyslu definice 14), 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. Vyhovuje-li nějaký hromadný bod $x^* \in R^n$ posloupnosti $x_i \in R^n$, $i \in N$, nutným podmínkám pro lokální minimum (Tvrzení 2), pak již platí $x_i \to x^*$. V dalším výkladu budu předpokládat, že globální konvergence automaticky znamená $x_i \to x^*$.

Věta 9 (Lineární konvergence) Nechť $x_i \in R^n$, $i \in N$ je posloupnost generovaná metodou spádových směrů taková, že $x_i \to x^*$. Nechť funkce $F : R^n \to R$ vyhovuje podmínkám (F3) a (F4). Pak pro libovolný index $k \in N$ platí

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

$\forall i \ge k$, kde $q = \sqrt{1 - 2\varepsilon_0\varepsilon_1\underline{c}\underline{G}^2/\overline{G}}$.

Důkaz (pro (S3b)) Podle (b) platí

$$F_{i+1} - F^* \le F_i - F^* - \varepsilon_0\varepsilon_1\underline{c} \| g_i \|^2$$

z vět o střední hodnotě dostaneme

$$F_i - F^* \le \frac{1}{2}\overline{G} \| x_i - x^* \|^2 \tag{c}$$

$$F_i - F^* \ge \frac{1}{2}\underline{G} \| x_i - x^* \|^2 \tag{d}$$

$$\| g_i \| \le \overline{G} \| x_i - x^* \| \tag{e}$$

$$\| g_i \| \ge \underline{G} \| x_i - x^* \| \tag{f}$$

Můžeme tedy psát

$$\frac{F_{i+1} - F^*}{F_i - F^*} \le 1 - \varepsilon_0\varepsilon_1\underline{c}\frac{\| g_i \|^2}{F_i - F^*} \le 1 - \varepsilon_0\varepsilon_1\underline{c}\frac{2\underline{G}^2 \| x_i - x^* \|^2}{\overline{G} \| x_i - x^* \|^2} = 1 - 2\varepsilon_0\varepsilon_1\underline{c}\underline{G}^2/\overline{G} = q^2 \tag{g}$$

Pro podmínku (S3b) platí $\underline{c} = \varepsilon_0(1 - \varepsilon_2)/\overline{G}$ (důkaz věty 8), takže

$$q^2 = 1 - 2\varepsilon_0^2\varepsilon_1(1 - \varepsilon_2)\underline{G}^2/\overline{G}^2 > 1 - 2\varepsilon_0^2\varepsilon_1(1 - \varepsilon_1)\underline{G}^2/\overline{G}^2 > 1 - \varepsilon_0^2\underline{G}^2/\overline{G}^2 \ge 0$$

neboť $0 < \varepsilon_0 \le 1$, $0 < \varepsilon_1 < 1/2$, $\varepsilon_1 < \varepsilon_2 < 1$. Několikanásobným použitím nerovnosti (g) dostaneme

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

což s použitím (c) a (d) dává

$$\frac{\| x_i - x^* \|}{\| x_k - x^* \|} \le \sqrt{\frac{\overline{G}}{\underline{G}}}\sqrt{\frac{F_i - F^*}{F_k - F^*}} \le \sqrt{\frac{\overline{G}}{\underline{G}}}q^{i-k}$$

Poznámka 9 Často se (F4) předpokládá až od nějakého indexu k (například, když $x_i \to x^*$ a x^* splňuje postačující podmínky pro lokální minimum, pak předpokládáme (F4) až v okolí bodu x^*).

Poznámka 10 Podle věty 9 je $\| e_{i+1} \| = O(\| e_i \|)$. Tento vztah platí dokonce i tehdy jesliže $q = 1$, takže kromě (F3) a (F4) stačí předpokládat, že $F_{i+1} \leq F_i$.

Definice 18 Řekneme, že výběr délky kroku je asymptoticky přesný, jestliže

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

Lemma 10 Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou spadových směrů s asymptoticky přesným výběrem délky kroku, taková, že $x_i \to x^*$. Nechť funkce $F : R^n \to R$ vyhovuje podmínkám (F3)-(F5). Pak platí

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

Důkaz Budeme psát $u_i \sim v_i$ jestliže $u_i = O(v_i)$ a $v_i = O(u_i)$. Podle věty 4 platí

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

což s použitím (F3) a (F4) dává $g_i \sim e_i$. Jelikož z (F3) a (F4) plyne $s_i^T G^* s_i \sim \| s_i \|^2$, z (S1) plyne $s_i^T g_i \sim \| s_i \| \| g_i \|$ a předpokládáme, že $\| e_i \| \to 0$, můžeme psát

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

takže $\| d_i \| = \| \alpha_i s_i \| \sim \| g_i \| \sim \| e_i \|$ a podle věty 3 platí

$$\begin{aligned}
F_{i+1} - F_i &= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G_i s_i + O(\| d_i \|^3) = \\
&= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G^* s_i + \frac{1}{2} d_i^T (G_i - G^*) d_i + O(\| d_i \|^3) = \\
&= \alpha_i s_i^T g_i + \frac{1}{2}\alpha_i^2 s_i^T G^* s_i + O(\| e_i \|^3)
\end{aligned}$$

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

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

neboť $(s_i^T g_i)^2 / s_i^T G^* s_i \sim \| g_i \|^2 \sim \| e_i \|^2$.

Lemma 11 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

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

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

Důkaz (a) Nechť vektory $u \in R^n$ a $v \in R^n$ vyhovují podmínce $(u^Tv)^2/(u^Tuv^Tv) \leq \varepsilon^2$. 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 $u^Tw = 0$. 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 $u^Tv = \alpha$ a $v^Tv = \alpha^2 + \beta^2$. Z nerovnosti $(u^Tv)^2/(u^Tuv^Tv) \leq \varepsilon^2$ a z $\| v \|= 1$ pak plyne

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

a

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

Položmě $W = [u, w]$. 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

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

Můžeme tedy psát

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

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

$$\begin{aligned}
\frac{(u^T Bv)^2}{u^T Buv^T Bv} &= 1-\frac{\det(V^T BV)}{u^T Buv^T Bv} \le 1-\frac{4\kappa(V^T BV)}{(1+\kappa(V^T BV))^2} = \\
&= \left(\frac{\kappa(V^T BV)-1}{\kappa(V^T BV)+1}\right)^2 \le \left(\frac{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon}-1}{\kappa(B)\frac{1+\varepsilon}{1-\varepsilon}+1}\right)^2 = \\
&= \left(\frac{\kappa(B)-1+(\kappa(B)+1)\varepsilon}{\kappa(B)+1+(\kappa(B)-1)\varepsilon}\right)^2
\end{aligned}$$

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

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

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

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 ortogonální bázi v_i, $1 \le i \le n$, tak, aby platilo $v_1 = u/\parallel u \parallel$ a $v_2 = w/\parallel w \parallel$. Pak platí

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

a

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

takže

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

a použijeme-li (a), dostaneme

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

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

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

Věta 12 Nechť jsou splněny podmínky lemmatu 10. 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 4 platí

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

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

takže s použitím (F3) a (F4) a z toho, že $\| g_i \| \sim \| e_i \|$ dostaneme

$$e_i = (G^*)^{-1} g_i (1 + O(\| e_i \|))$$

a

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

Použijeme-li lemma 10 můžeme psát

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

(platí $(1 + O(\| e_i \|))/(1 + O(\| e_i \|)) = 1 + O(\| e_i \|)$). Podle (S1) platí $(s_i^T g_i)^2 \geq \varepsilon_0^2 \| s_i \|^2 \| g_i \|^2$ takže s použitím lemmatu 11 dostaneme

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

takže

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

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 9, 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 4 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 11 Pro metodu největšího spádu je $\varepsilon_0 = 1$, takže

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

Poznámka 12 Asymptoticky přesný výběr délky kroku dostaneme, vybíráme-li délku kroku pomocí kvadratické nebo kubické interpolace (viz poznámku 15).

Věta 13 (Superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná metodou spádových směrů taková, že $x_i \to x^*$. Nechť funkce $F : R^n \to R$ vyhovuje podmínkám (F3) a (F4). Nechť $\alpha_i = 1$, kdykoliv tato hodnota vyhovuje podmínkám (S2) a (S3). Nechť směrový vektor se určuje tak, že

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

kde

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

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 $x^* \in R^n$. Navíc $-s_i^T g_i \geq \varepsilon_0 \| s_i \| \| g_i \|\ \forall i \geq k$, pokud $\varepsilon_0 < 1/\kappa(G^*)$.

Důkaz (a) ukážeme, že existuje index $k_1 \in N$ tak, že

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

pokud $\underline{G} < \underline{\lambda}(G^*)$ a $\overline{G} > \overline{\lambda}(G^*)$. 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

$$\| s_i \| \geq \frac{1 - \| \omega_i \|}{\| G_i \| + \| \vartheta_i \|} \| g_i \|$$

a jelikož $\| \vartheta_i \| \to 0$ a $\| \omega_i \| \to 0$ (podle (α) a (β)) a $\| G_i \| \to \| G^* \| = \overline{\lambda}(G^*) < \overline{G}$, existuje index $k_0 \in N$ tak, že $\| s_i \| \geq \| g_i \| / \overline{G}\ \forall i \geq k_0$. Podobně platí

$$s_i = G_i^{-1}(\omega_i \| g_i \| - \vartheta_i \| s_i \| - g_i)$$

takže

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

a jelikož $\| \vartheta_i \| \to 0$ a $\| \omega_i \| \to 0$ (podle (α) a (β)) a $\| G_i^{-1} \| \to \| (G^*)^{-1} \| = 1/\underline{\lambda}(G^*) < 1/\underline{G}$, existuje index $k_1 \geq k_0$ tak, že $\| s_i \| \leq \| g_i \| / \underline{G}\ \forall i \geq k_1$.
(b) Ukážeme, že existuje index $k_2 \geq k_1$ takový, že $-s_i^T g_i \geq \varepsilon_0 \| s_i \| \| g_i \|\ \forall i \geq k_2$, pokud $\varepsilon_0 < 1/\kappa(G^*)$. Zvolme $\underline{G} < \underline{\lambda}(G^*)$ a $\overline{G} > \overline{\lambda}(G^*)$ tak, aby platilo $\varepsilon_0 = \underline{G}/\overline{G}$. Z definice ω_i a ϑ_i a z (a) plyne, že

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

a jelikož $\| \omega_i \| \to 0$ a $\| \vartheta_i \| \to 0$ (podle (α) a (β)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, existuje index $k_2 \geq k_1$ tak, že $-s_i^T g_i \geq (\underline{G}/\overline{G}) \| s_i \| \| g_i \| = \varepsilon_0 \| s_i \| \| g_i \|\ \forall i \geq k_2$.
(c) Ukážeme, že existuje index $k \geq k_2$ tak, že hodnota $\alpha_i = 1$ vyhovuje podmínkám (S2) a (S3b). Označme

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

Pak podle předchozích výsledků dostaneme pro $i \geq k_2$

$$|\eta_i| = \frac{|s_i^T g_i + s_i^T G_i s_i|}{|s_i^T g_i|} \leq \frac{\| s_i \| \| g_i + G_i s_i \|}{\varepsilon_0 \| s_i \| \| g_i \|} \leq \frac{\overline{G}}{\underline{G}} \left(\frac{\| g_i + B_i s_i \|}{\| g_i \|} + \frac{\| (B_i - G_i) s_i \|}{\| g_i \|} \right)$$

a podle (α), (β) a (a) platí $|\eta_i| \to 0$. Nyní použijeme věty o střední hodnotě.

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

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

Můžeme tedy psát

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

$$\lim_{i \to \infty} \frac{s_i^T g(x_i + s_i)}{s_i^T g_i} = \lim_{i \to \infty} (\eta_i + o(\| s_i \|^2) / \| s_i \|^2) = 0$$

a protože $0 < \varepsilon_1 < 1/2$ a $\varepsilon_1 < \varepsilon_2 < 1$ existuje index $k \geq k_2$ tak, že (S2) a (S3b) (s $\alpha_i = 1$) platí $\forall i \geq k$.
(d) Superlineární konvergence. Použijeme větu o střední hodnotě

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

Podle předchozích výsledků dostaneme $x_{i+1} = x_i + s_i$ $\forall i \geq k$ a $\| s_i \| \sim \| g_i \| \to 0$
Můžeme tedy psát

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

takže podle (α), (β) a (a) platí

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

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 $\gamma_1 \overline{\alpha} \leq \alpha \leq \gamma_2 \overline{\alpha}$ a přejdeme na krok 2.

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

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 13 Jsou-li splněny podmínky (F1) a (F3) najde algoritmus délku kroku vyhovující podmínkám (S2) a (S3b) po konečném počtu kroků. Je-li splněna podmínka (F4) nezávisí tento počet na indexu $i \in N$.

Poznámka 14 Extrapolace a interpolace. Označme $\varphi(\alpha) = F(x_i + \alpha s_i)$, $\varphi(\alpha) = s_i^T g(x_i + \alpha s_i)$ a

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

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

Kvadratická interpolace (dvě hodnoty):

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

Kvadratická interpolace (dvě derivace):

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

Kubická interpolace:

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

kde

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

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

Poznámka 15 Určuje-li se délka kroku pomocí (Ka) nebo (Kb) nebo (C) a platí-li (F3) a (F4), je výběr délky kroku asymptoticky přesný.

Poznámka 16 Počáteční výběr délky kroku. Pro superlineárně konvergentní metody volíme $\alpha = 1$. U metod sdružených gradientů volíme $\alpha = 2(F_{i-1} - F_i)/s_i^T g_i$ nebo (v prvním iteračním kroku) $\alpha = 2(\underline{F} - F_i)/s_i^T g_i$.

Shrnutí: Pro metody spádových směrů (s parametry $\varepsilon_0, \varepsilon_1, \varepsilon_2$ vystupujími v podmínkách (S1), (S2), (S3)) platí tyto implikace:

a) (F1) - (F3) $\Rightarrow$ globální konvergence
b) (F3) - (F4) $\Rightarrow$ lineární konvergence
c) (F3) - (F5) $\Rightarrow$ asymptotický odhad

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

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

$$\frac{\| B_i s_i + g_i \|}{\| g_i \|} \to 0$$

$$\frac{\| (G^* - B_i) s_i \|}{\| g_i \|} \to 0$$

2.2. Metody sdružených gradientů

Definice 19 Řekneme, že základní optimalizační metoda je metodou sdružených gradientů jestliže $s_1 = -g_1$ a

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

(Hestenes, Stiefel) nebo

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

(Polak, Ribiere) nebo

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

(Fletcher, Reeves) pro $i \in N$. Přitom $y_i = g_{i+1} - g_i$.

Poznámka 17 Metoda (CGa) je teoreticky nejpodloženější. Metoda (CGb) dává nejlepší praktické výsledky. Metoda (CGc) je nejjednodušší a je globálně konvergentní bez přerušování iteračního procesu.

Věta 14 (Globální konvergence). Nechť funkce $F : R^n \to R$ splňuje podmínky (F1)-(F3). Pak metoda sdružených gradientů Fletchera a Reevese (CGc) s výběrem délky kroku splňujícím silnou Wolfeho podmínku (S2) a (S3a), kde $\varepsilon_1 < \varepsilon_2 < 1/2$, je globálně konvergentní.

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

$$-1 - \frac{\varepsilon_2}{1-\varepsilon_2} \le \frac{s_i^T g_i}{\| g_i \|^2} \le -1 + \frac{\varepsilon_2}{1-\varepsilon_2}$$

$\forall i \in N$. Pro $i = 1$ to platí, neboť $s_1 = -g_1$ a tedy $s_1^T g_1 / \| g_1 \|^2 = -1$. Použijeme-li (CGc), dostaneme

$$\frac{s_{i+1}^T g_{i+1}}{\| g_{i+1} \|^2} = -1 + \frac{\| g_{i+1} \|^2}{\| g_i \|^2} \frac{s_i^T g_{i+1}}{\| g_{i+1} \|^2} = -1 + \frac{s_i^T g_{i+1}}{\| g_i \|^2}$$

Podle (S3a) platí

$$|s_i^T g_{i+1}| \le -\varepsilon_2 s_i^T g_i$$

takže

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

(b) Z (S3a) plyne (S3b). Podle (F3) a (S3b) platí

$$\varepsilon_2 s_i^T g_i \le s_i^T g_{i+1} \le s_i^T g_i + \alpha_i \overline{G} \| s_i \|^2$$

takže

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

Podle (S2) platí

$$F_i - F_{i+1} \ge -\varepsilon_1 \alpha_i s_i^T g_i \ge \frac{\varepsilon_1 (1-\varepsilon_2)(s_i^T g_i)^2}{\overline{G} \| s_i \|^2} = \frac{\varepsilon_1 (1-\varepsilon_2)}{\overline{G}} \frac{(s_i^T g_i)^2}{\| g_i \|^4} \frac{\| g_i \|^4}{\| s_i \|^2}$$

Z Al-Baaliho nerovnosti (pravá část) plyne

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

neboť $\varepsilon_1 < \varepsilon_2 < 1/2$. Platí tedy

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

Předpokládejme, že neplatí

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

Pak existuje $\underline{\varepsilon} > 0$ tak, že $\| g_i \| \geq \underline{\varepsilon}\ \forall i \in N$, takže

$$\infty > \frac{(F_1 - \underline{F})\overline{G}(1-\varepsilon_2)}{\underline{\varepsilon}^4 \varepsilon_1 (1-2\varepsilon_2)^2} \geq \sum_{i=1}^{\infty} \frac{1}{\| s_i \|^2}$$

(c) Z (S3a) plyne (S3b). Podle (S3b) a Al-Baaliho nerovnosti (levá část) platí

$$-s_i^T g_{i+1} \leq -\varepsilon_2 s_i^T g_i \leq \varepsilon_2 \left(1 + \frac{\varepsilon_2}{1-\varepsilon_2}\right) \| g_i \|^2 = \frac{\varepsilon_2}{1-\varepsilon_2} \| g_i \|^2$$

Použijeme-li (CGc), dostaneme

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

Toto je rekurzivní vztah pro $\| s_i \|$. Postupným dosazováním dostaneme

$$\begin{aligned}
\| s_{i+1} \|^2 &\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \| g_{i+1} \|^2 + \frac{\| g_{i+1} \|^4}{\| g_i \|^4} \| s_i \|^2 \leq \\
&\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \| g_{i+1} \|^2 + \frac{\| g_{i+1} \|^4}{\| g_i \|^4} \left(\frac{1+\varepsilon_2}{1-\varepsilon_2} \| g_i \|^2 + \frac{\| g_i \|^4}{\| g_{i-1} \|^4} \| s_{i-1} \|^2 \right) \leq \\
&\leq \frac{1+\varepsilon_2}{1-\varepsilon_2} \| g_{i+1} \|^4 \sum_{j=1}^{i+1} \frac{1}{\| g_j \|^2}
\end{aligned}$$

Předpokládejme, že neplatí

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

Pak existuje $0 < \underline{\varepsilon} < \overline{\varepsilon} < \infty$ tak, že $\underline{\varepsilon} \leq \| g_i \| \leq \overline{\varepsilon}\ \forall i \in N$ (existence $\overline{\varepsilon}$ plyne z (F2) a (F3)), takže

$$\| s_i \|^2 \leq \frac{(1+\varepsilon_2)\overline{\varepsilon}^4}{(1-\varepsilon_2)\underline{\varepsilon}^2} i$$

a můžeme psát

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

což je spor, neboť v (b) jsme dokázali že tento součet je konečný.

Poznámka 18 Označme β_i koeficient u s_i v (CGa)-(CGc). Dá se dokázat, že metoda sdružených gradientů je globálně konvegentní, pokud

$$0 \leq \beta_i \leq \frac{1}{2\varepsilon_3} \frac{\| g_{i+1} \|^2}{\| g_i \|^2} \qquad \forall i \in N$$

kde $0 < \varepsilon_2 < \varepsilon_3 < 1/2$. Toto však nelze obecně zaručit. Proto se používá přerušování iteračního procesu ($s_i = -g_i$ pokud podmínka není splněna).

Věta 15 (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 $k \leq n$ tak, že $g_{k+1} = 0$ a $x_{k+1} = x^*$.

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

$$(\alpha) \qquad s_j^T g_i = 0 \qquad \forall 1 \leq j < i \leq n+1$$

$$(\beta) \qquad g_j^T g_i = 0 \qquad \forall 1 \leq j < i \leq n+1$$

$$(\gamma) \qquad s_j^T G s_i = 0 \qquad \forall 1 \leq j < i \leq n$$

Z (β) plyne, že gradienty g_i, $1 \leq i \leq n$, jsou nenulové a vzájemně ortogonální, tudíž lineárně nezávislé, takže nutně $g_{n+1} = 0$.

Pro $i = 1$ platí $s_1^T g_1 = -g_1^T g_1 < 0$ takže $s_1 \neq 0$ a $\alpha_1 \neq 0$ a dále není co dokazovat. Indukční krok:
(a) Nechť $i \leq n$. Podle indukčního předpokladu (γ) platí:

$$s_j^T g_{i+1} = s_j^T g_i + s_j^T y_i = s_j^T g_i + \alpha_i s_j^T G s_i = 0$$

$\forall 1 \leq j < i$, neboť pro kvadratickou funkci $Q(x)$ platí $y_i = g_{i+1} - g_i = \alpha_i G s_i$. 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) a (a) dostaneme

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

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

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

$\forall 1 \leq j < i$ neboť podle předpokladu (γ) platí $s_j^T G s_i = 0 \ \forall 1 \leq j < i$. Dále podle (CGa) platí

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

takže $s_j^T G s_{i+1} = 0 \; \forall 1 \leq j \leq i$.

Poznámka 19 Důkaz byl proveden pro CGa. Věta 15 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 20 Metoda sdružených gradientů s přesným výběrem délky kroku najde minimum kvadratické funkce po nejvýše n krocích. 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 15, je třeba iterační proces přerušit ($s_{n+1} = -g_{n+1}$).

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

Poznámka 21 Přerušovaná metoda sdružených gradientů je globálně konvergentní (stačí modifikovat důkaz věty 8 tak jak je to naznačeno v poznámce 7).

Věta 16 Nechť $x_i \in R^n$, $i \in N$, je posloupnost získaná přerušovanou metodou sdružených gradientů Fletchera a Reevese (CGc) s výběrem délky kroku splňujícím silnou Wolfeho podmínku (S2) a (S3a), kde $\varepsilon_1 < \varepsilon_2 < 1/2$, taková, že $x_i \to x^*$. Nechť funkce $F : R^n \to R$ vyhovuje podmínkám (F3) a (F4). Pak směrové vektory $\| s_i \|$, $i \in N$ jsou stejnoměrně spádové a platí $\| s_i \| \sim \| g_i \|$.

Důkaz (a) Zřejmě $\| e_i \| = O(\| e_{i-1} \|)$ (poznámka 10) a $\| g_i \| \sim \| e_i \|$ (poznámka 2), takže $\| g_i \| = O(\| g_{i-1} \|)$. Existuje tedy konstanta $c < \infty$ tak, že

$$\frac{\| g_i \|}{\| g_{i-1} \|} \leq c \qquad \forall i \notin M$$

Nechť $i \notin M$. Pak podle (CGc) platí

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

takže

$$\frac{\| s_i \|}{\| g_i \|} \leq 1 + \frac{\| g_i \|}{\| g_{i-1} \|} \frac{\| s_{i-1} \|}{\| g_{i-1} \|} \leq 1 + c \frac{\| s_{i-1} \|}{\| g_{i-1} \|}$$

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

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

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

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

což spolu s (a) dává

$$-\frac{s_i^T g_i}{\| s_i \| \| g_i \|} = -\frac{s_i^T g_i}{\| g_i \|^2} \frac{\| g_i \|}{\| s_i \|} \geq -\frac{1}{\overline{c}} \frac{s_i^T g_i}{\| g_i \|^2} \geq \frac{1}{\overline{c}} \frac{1-2\varepsilon_2}{1-\varepsilon_2}$$

takže $-s_i^T g_i \geq \varepsilon_0 \| s_i \| \| g_i \|$ kde $\varepsilon_0 = (1-2\varepsilon_2)/(\overline{c}(1-\varepsilon_2))$.
(c) Použitím Al-Baaliho nerovnosti a Schwartzovy nerovnosti dostaneme

$$\| s_i \| \| g_i \| \geq -s_i^T g_i \geq \frac{1-2\varepsilon_2}{1-\varepsilon_2} \| g_i \|^2$$

což dává $\| s_i \| \geq \underline{c} \| g_i \|$, kde $\underline{c} = (1-2\varepsilon_2)/(1-\varepsilon_2)$.

Tvrzení 17 (n-kroková kvadratická konvergence) Nechť jsou splněny předpoklady věty 16 a nechť výběr délky kroku je asymptoticky přesný. Nechť funkce $F : R^n \to R$ splňuje navíc podmínku (F5). Pak existuje index $k \in M$ a konstanta $\overline{C} < \infty$ tak, že $\forall i \in M$, $i \geq k$ platí

$$\| x_{i+n} - x^* \| \leq \overline{C} \| x_i - x^* \|^2$$

Lemma 18 Nechť jsou splněny předpoklady věty 15. 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

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

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

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

kde $P_{i-1}^*(G)$ je určitý polynom stupně $i-1$ v G. Označme $\overline{P}_i^* = I + G P_{i-1}^*$ (takže $\overline{P}_i^*$ je stupně i a $\overline{P}_i^*(0) = 1$). Jelikož z důkazu věty 15 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

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

pro libovolný polynom $\overline{P}_i$ stupně i takový, že $\overline{P}_i(0) = 0$. Nechť λ_k a v_k $1 \le k \le n$ jsou vlastní čísla (nezáporná) a vlastní vektory (ortonormální) matice G a nechť

$$x_1 - x^* = \sum_{k=1}^{n} \gamma_k v_k$$

Pak

$$Q(x_1) - Q(x^*) = \frac{1}{2}(x_1 - x^*)G(x_1 - x^*) = \frac{1}{2}\left(\sum_{k=1}^{n} \gamma_k v_k\right)^T G \left(\sum_{k=1}^{n} \gamma_k v_k\right) = \frac{1}{2}\sum_{k=1}^{n} \gamma_k^2 \lambda_k$$

a

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

Po vydělení dostaneme

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

Věta 19 Nechť jsou splněny předpoklady věty 15. Nechť $g_i \ne 0$ pro nějaký index $1 \le i \le n$. Pak platí

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

Důkaz Podle lemmatu 18 platí

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

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

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

Tuto vlastnost má Čebyševův polynom transformovaný na interval $\lambda_1 \le \lambda \le \lambda_n$ 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 $|\xi| \le 1$ a $T_i(\xi) = ((\xi + \sqrt{\xi^2 - 1})^i + (\xi - \sqrt{\xi^2 - 1})^i)/2$ pro $|\xi| \ge 1$. Jelikož $|T_i(\xi)| \le 1$ pro $|\xi| \le 1$, platí

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

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

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

Důsledkem Tvrzení 17 (s výsledky získanými při jeho důkazu) a věty 19 je toto tvrzení.

Tvrzení 20 (Asymptotický odhad) Nechť jsou splněny předpoklady věty 16 a nechť výběr délky kroku je asymptoticky přesný. Nechť funkce $F : R^n \to R$ splňuje navíc podmínku (F5). Pak platí

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

Poznámka 22 Odhad $(\sqrt{\kappa}-1)/(\sqrt{\kappa}+1)$ je mnohem příznivější než odhad $(\kappa-1)/(\kappa+1)$ platný pro metodu spádových směrů jak ukazuje tato tabulka:

	SD	CG
$\kappa = 10^2$, $\varepsilon = 10^{-4}$	460	45
$\kappa = 10^4$, $\varepsilon = 10^{-6}$	69077	690
$\kappa = 10^6$, $\varepsilon = 10^{-8}$	9210340	9210

V tabulce je uveden počet iterací potřebný k dosažení požadované přesnosti ε.

Nyní uvedeme několik poznámek k implementaci metod sdružených gradientů.

1) Metody sdružených gradientů vyžadují některé úpravy algoritmu pro výběr délky kroku:
a) Používá se počáteční odhad

$$\alpha = \min\left(1, \frac{2(F_i - F_{i-1})}{s_i^T g_i}, \frac{20(\underline{F} - F_i)}{s_i^T g_i}\right)$$

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

b) Místo slabé Wolfeho podmínky (S3b) se používá silná Wolfeho podmínka (S3a)

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

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

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

která je částí podmínky (S3a).

2) Je výhodné metodu sdružených gradientů škálovat (nejjednodušší předpodmínění). Místo

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

použijeme vzorec

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

kde

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

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

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

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

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

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

3) Používá se řízené přerušování iteračního procesu:

a) Klasický způsob. Iterační proces se přeruší vždy po n iteracích, nebo když $\beta_i \le 0$.

b) Srovnání s (CGc). Iterační proces se přeruší, pokud neplatí

$$\eta_1 \beta_i^{FR} \le \beta_i \le \eta_2 \beta_i^{FR}$$

kde $\beta_i^{FR} = g_{i+1}^T g_{i+1} / g_i^T g_i$ a $\eta_1 \approx 0.1 - 0.3$ a $\eta_2 \approx 1.1 - 1.3$.

c) Test na sdruženost směrů. Iterační proces se přeruší pokud

$$s_{i+1}^T y_i \ge \eta_3 \parallel s_{i+1} \parallel\parallel y_i \parallel$$

kde $\eta_3 \approx 0.04 - 0.05$.

d) Test na ortogonalitu gradientů. Iterační proces se přeruší pokud

$$g_{i+1}^T g_i \ge \eta_4 \parallel g_{i+1} \parallel\parallel g_i \parallel$$

kde $\eta_4 \approx 0.4 - 0.5$.

Algoritmus metody sdružených gradientů lze zhruba popsat takto (čísla a písmena se vztahují k poznámkám k implementaci metod sdružených gradientů):

Algoritmus 2 (CG) Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 10^{-1}$, $\eta_1 = 0.3$, $\eta_2 = 1.2$, $\eta_3 = 0.05$, $\eta_4 = 0.46$, $\underline{\varepsilon} > 0$.

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), (CGb), (CGc) a rozhodneme o přerušení iteračního procesu podle některé ze strategií 3a, 3b, 3c, 3d. Určíme škálovací koeficient γ_i a určíme směrový vektor s_i podle 2.

Krok 3 Určíme délku kroku α_i použitím algoritmu 1 upraveného podle 1a a 1b. 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 24 testovacích funkcí se 100 proměnnými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i celkový čas výpočtu).

Metoda	NIT-NFV	selhání	Čas
CGc + 3a	10152 - 18772	1	2:13.90
CGc + 3c	4508 - 8952	-	1:01.84
CGc + 3d	4140 - 8397	-	1:00.91
CGb + 3a	7568 - 14917	1	1:48.20
CGb + 3b	5690 - 11248	-	1:09.10
CGa + 3a	10357 - 20423	1	1:31.51
CGa + 3b	5140 - 10111	-	1:07.18

2.3. Metody s proměnnou metrikou

Definice 21 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 $U_i \in R^{n\times 2}$, $M_i \in R^{2\times 2}$ (SPD) 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 23 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 charakterizivaná výběrem $U_i = [d_i, H_i y_i]$.

Věta 21 (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

$$(\alpha) \qquad H_i y_j = \lambda_i^j d_j \qquad \forall 1 \leq j < i \leq n+1$$

$$(\beta) \qquad s_j^T g_i = 0 \qquad \forall 1 \le j < i \le n+1$$

$$(\gamma) \qquad s_j^T G s_i = 0 \qquad \forall 1 \le j < i \le n$$

Z (VM1) a (γ) plyne, že s_i, $1 \le i \le n$, 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 $\alpha_1 > 0$ a dále není co dokazovat. Indukční krok:

(a) Nechť $i \le n$. 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 $U_i^T y_j = 0\ \forall 1 \le j < i$. Podle (VM2) a (α) tedy platí

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

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

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

(c) Podle (VM1) je $g_{i+1}^T s_{i+1} = -g_{i+1}^T H_{i+1} g_{i+1} < 0$ takže $s_{i+1} \ne 0$ a $\alpha_{i+1} \ne 0$. Použijeme-li (VM1), (Q), (a), (b) dostaneme

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

$\forall 1 \le j \le i$.

Věta 22 (Aproximace Hessovy matice). Nechť jsou splněny předpoklady věty 21 s $\gamma_i = 1$ a $\rho_i = 1$ $\forall i \in N$. Pak platí $H_{n+1} = G^{-1}$.

Důkaz Z důkazu věty 21 plyne, že

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

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

Poznámka 24 Nyní se budeme zabývat vyšetřováním aktualizace (VM3). Pro zjednodušení budeme index i vynechávat a index $i+1$ nahradíme symbolem $+$.

Věta 23 Nechť $H_+ = \gamma(H + UMU^T)$, kde $U = [d, Hy]$ je matice hodnosti 2. 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

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

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 23.

Poznámka 25 Vztah $H_+ = \gamma(H + UMU^T)$ můžeme roznásobit. Pak platí

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

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

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

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

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

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

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

Z těchto tří konkrétních metod je bez dalších úprav prakticky použitelná pouze metoda BFGS. Metoda DFP vyžaduje přesný výběr délky kroku nebo důsledné škálování, jinak konverguje velmi pomalu (podrobnější zdůvodnění udává poznámka 34). Metoda hodnosti 1 obecně nesplňuje podminku pro pozitivní definitnost matice H_+ (tuto podmínku udává věta 25), takže může dojít ke ztrátě globální konvergence vlivem porušení podmínky spádovosti ($\overline{\text{S1}}$). Metoda hodnosti 1 se často kombinuje s metodou BFGS.

Lemma 24 Nechť H je SPD matice, $a > 0$, $b > 0$, $c > 0$, $ac - b^2 > 0$ a nechť platí (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.

$$\lambda^2 - p\lambda + q = 0$$

kde

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

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

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 $H^{-\frac{1}{2}}U$. Zbylé dva vlastní vektory můžeme vyjádřit ve tvaru $H^{-\frac{1}{2}}Uz$ takže odpovídající vlastní čísla musí vyhovovat rovnici

$$H^{-\frac{1}{2}}U(I + MU^TH^{-1}U)z = \lambda H^{-\frac{1}{2}}Uz$$

neboli (po vynásobení $(U^TH^{-1}U)^{-1}U^TH^{-\frac{1}{2}}$ zleva)

$$((1-\lambda)I + MU^TH^{-1}U)z = 0$$

Dosadíme-li M z věty 23 a

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

můžeme psát

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

což po úpravě dává $\lambda^2 - p\lambda + q$ s koeficienty uvedenými v lemmatu 24.

Věta 25 Nechť jsou splněny předpoklady lemmatu 24. Pak H_+ je SPD právě tehdy, jestliže $\eta(ac - b^2) + b^2 > 0$.

Důkaz Je třeba najít podmínku pro to, aby rovnice $\lambda^2 - p\lambda + q$ s koeficienty uvedenými v lemmatu 24 měla kladné kořeny. Označme λ_1 a λ_2 tyto kořeny. Pak $\lambda_1 + \lambda_2 = p$ a $\lambda_1\lambda_2 = q$ takže $\lambda_1 > 0$ a $\lambda_2 > 0$ právě tehdy. když $p > 0$ a $q > 0$. Z definice p a q plyne, že

$$p = \frac{a}{b}\frac{\gamma}{\rho}q + \frac{\rho}{\gamma}\frac{c}{b}$$

Jelikož předpokládáme $a > 0$, $b > 0$, $c > 0$, $\gamma > 0$, $\rho > 0$, platí $p > 0$ kdykoliv $q > 0$. Z $q > 0$ dostaneme podmínku $\eta(ac - b^2) + b^2 > 0$.

Poznámka 26 Poslední nerovnost lze zapsat ve tvaru $\eta > \eta^*$, kde

$$\eta^* = -\frac{b^2}{ac - b^2} < 0$$

Podmínky $a > 0$, $c > 0$ jsou splněny, je-li matice H SPD a $d \neq 0$, $y \neq 0$. Podmínka $b > 0$ je splněna, vybíráme-li délku kroku podle (S3b), neboť

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

Jestliže $b \leq 0$, není matice H^+ SPD pro žádné hodnoty parametrů γ, ρ a η.

Věta 26 (Aktualizace matice $B = H^{-1}$). Nechť jsou splněny předpoklady lemmatu 24. Nechť $B = H^{-1}$ a nechť H_+ je matice určená pomocí aktualizace (H). Nechť $B_+ = H_+^{-1}$. Pak platí

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

kde

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

Důkaz Inverzí vztahu $H_+ = \gamma(H + UMU^T)$ dostaneme

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

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

$$B_+d = \frac{1}{\gamma}\left(Bd + [Bd, y]\begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix}\begin{bmatrix} c \\ b \end{bmatrix}\right) = \frac{1}{\gamma}(Bd + (k_1c + k_2b)Bd + (k_2c + k_3b)y) = \frac{1}{\rho}y$$

takže nutně

$$k_1 c + k_2 b = -1$$

$$k_2 c + k_3 b = \gamma/\rho$$

Zvolíme $k_2 = -\beta/b$ a zbylé prvky k_1, k_3 určíme řešením 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 $B_+ = (1/\gamma)(B + BUKU^T B)$ dává (B). Vztah svazující β s η lze získat například z rovnosti

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

(nebudeme to provádět).

Poznámka 27 (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 $\beta = 1$:

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

Metodu BFGS dostaneme pro $\beta = 0$:

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

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

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

Poznámka 28 Matice B_+ je SPD právě tehdy, jestliže $\beta > \beta^*$, kde

$$\beta^* = -\frac{b^2}{ac - b^2} < 0$$

Věta 27 Nechť jsou splněny předpoklady lemmatu 24 přičemž

$$\begin{aligned} q &= \frac{\rho}{\gamma}\frac{1}{ab}\left(\eta(ac - b^2) + b^2\right) \geq 0 \\ m &= \frac{1}{ab}\left(\eta\left(\frac{a}{b} - \frac{\rho}{\gamma}\right) + \frac{\rho}{\gamma}\right) \geq 0 \end{aligned}$$

Nechť $H = SS^T$ a nechť H_+ je matice určená pomocí aktualizace (H). Označme $\tilde{U} = S^{-1}U = [S^{-1}d, S^T y] \triangleq [\tilde{d}, \tilde{y}]$. Pak platí $H_+ = S_+ S_+^T$, kde

$$S_+ = \sqrt{\gamma}S(I + \tilde{U}uv^T\tilde{U}^T) \tag{S}$$

a

$$uv^T = \frac{1}{\lambda}\begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix}\begin{bmatrix} \sqrt{q} - b\sqrt{m} \\ -\frac{\gamma}{\rho}\sqrt{q} + c\sqrt{m} \end{bmatrix}^T$$

přičemž

$$\lambda = \left(\frac{\rho}{\gamma}c - b\right) + \left(b - \frac{\gamma}{\rho}a\right)\sqrt{q} + \left(ac - b^2\right)\sqrt{m}$$

Důkaz (a) Dosadíme-li vztah (S) do vztahu $H_+ = S_+ S_+^T$, dostaneme

$$H_+ = \gamma(I + Uuv^T U^T H^{-1}) H (I + Uuv^T U^T H^{-1})^T$$

Porovnáme-li tento vztah se vztahem $H_+ = \gamma(H + UMU^T)$, můžeme psát

$$M = uv^T + vu^T + uv^T U^T H^{-1} U v u^T = [u, v] \begin{bmatrix} v^T U^T H^{-1} U v & , & 1 \\ 1 & , & 0 \end{bmatrix} \begin{bmatrix} u^T \\ v^T \end{bmatrix}$$

a podle věty o násobení determinantů platí

$$(u_1 v_2 - v_1 u_2)^2 = (\det[u, v])^2 = -\det M = m$$

neboť z vyjádření matice M (věta 23) plyne, že

$$\det M = -\frac{1}{ab}\left(\eta\left(\frac{a}{b} - \frac{\rho}{\gamma}\right) + \frac{\rho}{\gamma}\right) = -m$$

Použijeme-li tento výsledek můžeme psát

$$uv^T - vu^T = \begin{bmatrix} 0 & , & u_1 v_2 - v_1 u_2 \\ v_1 u_2 - u_1 v_2 & , & 0 \end{bmatrix} = \begin{bmatrix} 0, & +\sqrt{m} \\ -\sqrt{m}, & 0 \end{bmatrix}$$

(b) Dosadíme-li vztah (S) do vztahu $H_+ = S_+ S_+^T$, dostaneme

$$\frac{1}{\gamma} H_+ = S(I + \tilde{U} uv^T \tilde{U}^T)(I + \tilde{U} vu^T \tilde{U}^T) S^T$$

Jelikož z důkazu lemmatu 24 víme, že platí $\det(H_+/\gamma) = q \det H$, můžeme psát

$$\det(I + \tilde{U} uv^T \tilde{U}^T) = \sqrt{q}$$

takže podle Shermanova-Morrisonova vzorce platí

$$(I + \tilde{U} uv^T \tilde{U}^T)^{-1} = I - \frac{1}{\sqrt{q}} \tilde{U} uv^T \tilde{U}^T$$

a podmínku $H_+ y = \rho d$ můžeme zapsat ve tvaru

$$(I + \tilde{U} vu^T \tilde{U}^T)\tilde{y} = \frac{\rho}{\gamma}\left(I - \frac{1}{\sqrt{q}} \tilde{U} uv^T \tilde{U}^T\right)\tilde{d}$$

vynásobíme-li tuto rovnici zleva maticí $\tilde{U}^T$ a přihlédneme-li k tomu, že platí

$$\tilde{U}^T \tilde{U} = \begin{bmatrix} \tilde{d}^T \tilde{d} & \tilde{d}^T \tilde{y} \\ \tilde{y}^T \tilde{d} & \tilde{y}^T \tilde{y} \end{bmatrix} = \begin{bmatrix} c, & b \\ b, & a \end{bmatrix}$$

dostaneme

$$\begin{bmatrix} b \\ a \end{bmatrix} + \begin{bmatrix} c, & b \\ b, & a \end{bmatrix} vu^T \begin{bmatrix} b \\ a \end{bmatrix} = \frac{\rho}{\gamma}\left(\begin{bmatrix} c \\ b \end{bmatrix} - \frac{1}{\sqrt{q}} \begin{bmatrix} c, & b \\ b, & a \end{bmatrix} uv^T \begin{bmatrix} c \\ b \end{bmatrix}\right)$$

což po úpravě dává

$$vu^T \begin{bmatrix} b \\ a \end{bmatrix} + uv^T \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix} = \frac{1}{ac - b^2} \begin{bmatrix} a, & -b \\ -b, & c \end{bmatrix} \left(\frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix} - \begin{bmatrix} b \\ a \end{bmatrix}\right) = \begin{bmatrix} \frac{\rho}{\gamma} \\ -1 \end{bmatrix}$$

Použijeme-li nyní (a), dostaneme

$$uv^T \left(\begin{bmatrix} b \\ a \end{bmatrix} + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} \begin{bmatrix} c \\ b \end{bmatrix}\right) = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix}$$

Z tohoto vyjádření je patrné, že vektor $u \in R^2$ je skalárním násobkem vektoru na pravé straně poslední rovnosti. Jelikož skalární násobek můžeme zvolit libovolně, položíme

$$u = \begin{bmatrix} \frac{\rho}{\gamma} + a\sqrt{m} \\ -1 - b\sqrt{m} \end{bmatrix}$$

Pak pro vektor $v \in R^2$ dostaneme rovnici

$$v_1 \left(b + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} c \right) + v_2 \left(a + \frac{1}{\sqrt{q}} \frac{\rho}{\gamma} b \right) = 1$$

a z (a) plyne

$$v_1 \left(1 + b\sqrt{m} \right) + v_2 \left(\frac{\rho}{\gamma} + a\sqrt{m} \right) = \sqrt{m}$$

Řešením těchto dvou rovnic je vektor

$$v = \frac{1}{\lambda} \begin{bmatrix} \sqrt{q} - b\sqrt{m} \\ -\frac{\gamma}{\rho}\sqrt{q} + c\sqrt{m} \end{bmatrix}$$

kde

$$\lambda = \left(\frac{\rho}{\gamma} c - b \right) + \left(b - \frac{\gamma}{\rho} a \right) \sqrt{q} + (ac - b^2)\sqrt{m}$$

Poznámka 29 Vzorec (S) můžeme použít tak, že místo vztahu $s = -Hg$ používáme vztahy

$$\tilde{g} = S^T g, \qquad \tilde{s} = -\tilde{g}, \qquad s = S\tilde{s}$$

Jelikož $d = \alpha s$, můžeme položit $\tilde{d} = \alpha \tilde{s}$ a $\tilde{y} = S^T(g_+ - g)$. Matici S pak aktualizujeme podle vzorce

$$S_+ = \sqrt{\gamma} S (I + \tilde{u}\tilde{v}^T)$$

kde $\tilde{u} = u_1 \tilde{d} + u_2 \tilde{y}$ a $\tilde{v} = v_1 \tilde{d} + v_2 \tilde{y}$ (čísla u_1, u_2 a v_1, v_2 jsou určeny větou 27). Tento způsob se používá zejména tehdy, není-li matice S čtvercová, například při řešení úloh s lineárními omezeními, kdy má matice S více řádků než sloupců.

Poznámka 30 Pro metodu DFP platí $\eta = 0$, čili $q = \rho b/(\gamma a)$ a $m = \rho/(\gamma a b)$, takže po dosazení do (S) dostaneme

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

Pro metodu BFGS platí $\eta = 1$, čili $q = \rho c/(\gamma b)$ a $m = 1/b^2$, takže po dosazení do (S) dostaneme

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

Pro metodu hodnosti 1 platí $\eta = (\rho/\gamma)(\rho/\gamma - a/b)$, čili $q = ((\rho/\gamma)c - b)/(b - (\gamma/\rho)a)$, takže po dosazení do (S) dostaneme

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

Poznámka 31 Podle věty 27 platí $H = SS^T$ kde matice S je aktualizována podle vztahu (S). Předpokládejme nyní, že $B = A^T A = (S^{-1})^T S^{-1} = H^{-1}$. Vztah pro aktualizaci matice $A = S^{-1}$ můžeme odvodit pomocí Shermanova-Morrisonova vzorce. Platí

$$A_+ = \left(\sqrt{\gamma} S \left(I + \tilde{U} u v^T \tilde{U}^T\right)\right)^{-1} = \frac{1}{\sqrt{\gamma}} \left(I - \frac{1}{\sqrt{q}} \tilde{U} u v^T \tilde{U}^T\right) A \tag{A}$$

(viz důkaz věty 27). Pro metodu DFP platí

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

Pro metodu BFGS platí

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

Pro metodu hodnosti 1 platí

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

kde $1/q = ((\gamma/\rho)a - b)/(b - (\rho/\gamma)c)$. Aktualizace (A) se používá zejména v úlohách na nejmenší čtverce.

Ukážeme ještě jednu vlastnost metod s proměnnou metrikou.

Věta 28 Nechť W je symetrická pozitivně definitní matice. Pak Frobeniova norma $\| W^{-1/2}(\gamma B_+ - B)W^{-1/2} \|_F$ je minimální na množině všech matic splňujících podmínku

$$(\gamma B_+ - B)d = \frac{\gamma}{\rho} y - Bd \triangleq w$$

právě tehdy, platí-li

$$B_+ = \frac{1}{\gamma} \left(B + \frac{W d w^T + w (W d)^T}{d^T W d} - \frac{w^T d}{d^T W d} \frac{W d (W d)^T}{d^T W d}\right)$$

Důkaz Jelikož matice $\gamma B_+ - B$ je symetrická, můžeme položit $\gamma B_+ - B = X + X^T$, kde X je zatím neznámá čtvercová matice . Nutnost dokážeme pomocí Lagrangeovy funkce

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

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

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

$$
\begin{aligned}
&+\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} e_l^T W^{-1/2} e_i e_i^T W^{-1/2}(X+X^T)W^{-1/2}e_j e_j^T W^{-1/2} e_k + \\
&+\sum_{i=1}^{n}\sum_{j=1}^{n} u_i(e_i^T e_k e_l^T e_j + e_i^T e_l e_k^T e_j)d_j = \\
= \quad & e_k^T W^{-1}(X+X^T)W^{-1}e_l - (u_k d_l + u_l d_k)
\end{aligned}
$$

Podmínka pro stacionaritu Langrangovy funkce má tedy tvar

$$\gamma B_+ - B = W(ud^T + du^T)W$$

Dosadíme-li tento vektor do vztahu $(\gamma B_+ - B)d = w$, dostaneme po úpravě

$$(d^T Wd \cdot W + Wd(Wd)^T)u = w$$

Použijeme-li Shermanův-Morrisonův vzorec můžeme vypočítat inverzi

$$(d^T Wd \cdot W + Wd(Wd)^T)^{-1} = \frac{1}{d^T Wd}\left(W^{-1} - \frac{dd^T}{2d^T Wd}\right)$$

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

$$u = \frac{1}{d^T Wd}\left(W^{-1} - \frac{dd^T}{2d^T Wd}\right)w$$

což po dosazení do vztahu pro $\gamma B_+ - B$ dává tvrzení věty. Postačitelnost plyne z konvexity Frobeniovy normy.

Poznámka 32 Zvolíme-li matici W tak, aby platilo $Wd = \lambda y - Bd$, kde λ se volí tak, že

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

dostaneme aktualizaci (B). Jestliže $\lambda = \infty$ (neboli $Wd = y$), dostaneme metodu DFP. Jestliže $\lambda = \sqrt{\gamma c/(\rho b)}$ dostaneme metodu BFGS. Jestliže $\lambda = \gamma/\rho$ dostaneme metodu hodnosti 1. Zvolíme-li $W = I$, dostaneme metodu, která nepatří do Broydenovy třídy a která se nazývá Powellovou symetrizací Broydenovy metody (PSB).

$$B_+ = \frac{1}{\gamma}\left(B + \frac{dw^T + wd^T}{d^T d} - \frac{w^T d}{d^T d}\frac{dd^T}{d^T d}\right)$$

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 55).

Lemma 29 Nechť x_i, $i \in N$ je posloupnost generovaná 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 (F3) a (F4). Pak platí

$$
\begin{aligned}
Tr(B_{i+1}) \quad &= \quad \frac{1}{\gamma_i}\left(Tr(B_i) + \frac{\gamma_i}{\rho_i}\frac{y_i^T y_i}{y_i^T d_i} + \beta_i \frac{y_i^T y_i}{y_i^T d_i}\frac{d_i^T B_i d_i}{y_i^T d_i} - \right. \\
&\left. -2\beta_i \frac{y_i^T B_i d_i}{y_i^T d_i} - (1-\beta_i)\frac{(B_i d_i)^T B_i d_i}{d_i^T B_i d_i}\right) \leq \\
&\leq \quad \frac{1}{\underline{\gamma}}\left(Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \frac{\overline{G}}{1-\varepsilon_2}\alpha_i + \frac{2\overline{G}}{1-\varepsilon_2}\frac{\alpha_i}{c_i} - \frac{\lambda \underline{G}}{2(1-\varepsilon_1)}\frac{\alpha_i}{c_i^2}\right)
\end{aligned}
\tag{T}
$$

a

$$\det(B_{i+1}) = \frac{1}{\gamma_i^n}\frac{\gamma_i}{\rho_i}\det(B_i)\frac{y_i^T d_i}{d_i^T B d_i}\left(1-\frac{\beta_i}{\beta_i^*}\right) \geq \frac{1}{\overline{\gamma}^n}\frac{\underline{\gamma}}{\overline{\rho}}\det(B_i)\lambda\frac{1-\varepsilon_2}{\alpha_i} \tag{D}$$

kde $\alpha_i > 0$ je délka kroku a $c_i > 0$, je směrový kosinus ($|g_i^T d_i| = c_i \parallel g_i \parallel \parallel d_i \parallel$).

Důkaz 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 první část vztahu (T). Druhou část tohoto vztahu odvodíme pomocí vět o střední hodnotě

$$y_i = g_{i+1} - g_i = \int_0^1 G(x_i + \lambda d_i) d_i d\lambda \triangleq \tilde{G}_i d_i \tag{a}$$

a

$$F_{i+1} - F_i \geq g_i^T d_i + \frac{1}{2}\underline{G} \parallel d_i \parallel^2 \tag{b}$$

a pomocí vztahů $F_{i+1} - F_i \leq \varepsilon_1 g_i^T d_i$ (S3a) a $y_i^T d_i \geq (1-\varepsilon_2)|g_i^T d_i|$ (S3b) a $B_i d_i = -\alpha_i g_i$ (VM1). Podle (a), (S3b) a (VM1) platí

$$\begin{aligned}
\frac{y_i^T y_i}{y_i^T d_i} &= \frac{d_i^T \tilde{G}_i^2 d_i}{d_i^T \tilde{G}_i d_i} \leq \overline{G} \\
\frac{d_i^T B_i d_i}{y_i^T d_i} &= -\frac{\alpha_i g_i^T d_i}{(g_{i+1} - g_i)^T d_i} \leq \frac{\alpha_i}{(1-\varepsilon_2)} \\
\frac{|y_i^T B_i d_i|}{y_i^T d_i} &\leq \frac{\parallel \tilde{G} d_i \parallel \parallel \alpha_i g_i \parallel}{(1-\varepsilon_2)|g_i^T d_i|} \leq \frac{\alpha_i \overline{G}}{(1-\varepsilon_2) c_i}
\end{aligned}$$

Použijeme-li (b) a (S3a) dostaneme

$$(1-\varepsilon_1)|g_i^T d_i| \geq \frac{1}{2}\underline{G} \parallel d_i \parallel^2$$

Neboli $\parallel g_i \parallel / \parallel d_i \parallel \geq \underline{G}/(2c_i(1-\varepsilon_1))$, takže můžeme psát

$$\frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i} = \frac{\alpha_i \parallel g_i \parallel}{\parallel d_i \parallel c_i} \geq \frac{\alpha_i \underline{G}}{2(1-\varepsilon_1)c_i^2}$$

Dosadíme-li získané nerovnosti do první části vztahu (T) a užijeme-li toho, že $\beta_i \leq (1-\lambda) < 1$, dostaneme druhou část tohoto vztahu. Při důkazu vztahu (D) vyjdeme z toho, že platí

$$\det(\gamma_i B_{i+1}) = \det B_i / q_i$$

(lemma 24). Vztah pro $1/q_i$ lze odvodit pomocí duality (poznámka 28) nebo použitím vztahu svazujícího β_i s η_i (věta 26). Dostaneme

$$\frac{1}{q_i} = \frac{\gamma_i}{\rho_i}\frac{y_i^T d_i}{d_i^T B_i d_i}\left(1 - \frac{\beta_i}{\beta_i^*}\right)$$

(poznámka 29), což po dosazení do vztahu pro $\det(\gamma_i B_{i+1})$ dává první část vztahu (D). Druhá část tohoto vztahu plyne z již dokázaných nerovností.

Věta 30 (Globální konvergence) Nechť jsou splněny předpoklady lemmatu 29. Pak

$$\liminf_{i\to\infty} \parallel g_i \parallel = 0$$

Důkaz Bez újmy na obecnosti budeme předpokládat, že $\underline{\gamma} < 1 < \overline{\gamma}$. Podle vztahu (T) dostaneme

$$\begin{aligned} Tr(B_{i+1}) &\leq \frac{1}{\underline{\gamma}}\left(Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \overline{G}\frac{\| d_i \|^2}{y_i^T d_i}\frac{\| B_i d_i \|}{\| d_i \|} + 2\frac{\| y_i \| \| d_i \|}{y_i^T d_i}\frac{\| B_i d_i \|}{\| d_i \|}\right) \leq \\ &\leq \frac{1}{\underline{\gamma}}\left(Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \left(\frac{\overline{G}}{\underline{G}} + 2\frac{\overline{G}}{\underline{G}}\right) \| B_i \|\right) \leq \frac{1}{\underline{\gamma}}\frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \frac{1}{\underline{\gamma}}\left(1 + 3\frac{\overline{G}}{\underline{G}}\right) Tr(B_i) \end{aligned}$$

(používáme věty o střední hodnotě). Tento lineární rekurentní vztah implikuje existenci konstanty $\overline{C} > 1$ takové, že

$$Tr(B_{i+1}) \leq \overline{C}^i$$

Podle druhé části (D) je

$$\det(B_{i+1}) \geq \frac{1}{\overline{\gamma}^n}\frac{\underline{\gamma}}{\overline{\rho}}\det(B_i)\lambda\frac{1-\varepsilon_2}{\alpha_i} \geq \left(\frac{1}{\overline{\gamma}^n}\frac{\underline{\gamma}^2}{\underline{\rho}}\right)^i \det(B_1)\lambda^i(1-\varepsilon_2)^i \prod_{j=1}^{i}\frac{1}{\underline{\gamma}^j\alpha_j}$$

Použitím nerovnosti pro geometrický a aritmetický průměr dostaneme

$$\det(B_{i+1}) \leq \left[\frac{Tr(B_{i+1})}{n}\right]^n \leq (Tr(B_{i+1}))^n \leq (\overline{C}^n)^i$$

takže

$$\prod_{j=1}^{i}\frac{1}{\underline{\gamma}^j\alpha_j} \leq \left(\frac{\overline{\gamma}^n\overline{\rho}}{\underline{\gamma}^2}\right)^i \frac{(\overline{C}^n)^i}{\lambda^i(1-\varepsilon_2)^i(\min(1,\det(B_1))^i} \triangleq \frac{1}{\underline{C}^i}$$

neboli podle nerovnosti pro geometrický a aritmetický průměr

$$\underline{C} \leq \left(\prod_{j=1}^{i}\underline{\gamma}^j\alpha_j\right)^{\frac{1}{i}} \leq \frac{1}{i}\sum_{j=1}^{i}\underline{\gamma}^j\alpha_j \tag{$*$}$$

Označme

$$\xi_i = \frac{\overline{G}}{1-\varepsilon_2} + \frac{2\overline{G}}{(1-\varepsilon_2)c_i} - \frac{\lambda\underline{G}}{2(1-\varepsilon_1)c_i^2}$$

a přepokládejme, že $\liminf_{i\to\infty} \| g_i \| > 0$. Pak podle důkazu věty 8 nutně platí $c_i \to 0$ a tedy $\xi_i \to -\infty$. Existuje tedy číslo $K > 0$ a index $k \in N$ takový, že $\xi_i < -K < 0 \ \forall i \geq k$. Pak ale podle (T) a ($*$) platí

$$\begin{aligned} Tr(B_{i+1}) &\leq \frac{1}{\underline{\gamma}}\left(Tr(B_i) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G} + \alpha_i\xi_i\right) \leq \\ &\leq \frac{1}{\underline{\gamma}^{i+1-k}}\left[Tr(B_k) + \frac{\overline{\gamma}}{\underline{\rho}}\overline{G}\sum_{j=k}^{i}\underline{\gamma}^{j-k} - K\sum_{j=k}^{i}\underline{\gamma}^{j-k}\alpha_j\right] \leq \\ &\leq \frac{1}{\underline{\gamma}^{i+1-k}}\left[Tr(B_k) + \frac{\overline{\gamma}}{\underline{\rho}}\frac{\overline{G}}{1-\underline{\gamma}} - K\left(\sum_{j=1}^{i}\underline{\gamma}^j\alpha_j - \sum_{j=1}^{k-1}\underline{\gamma}^j\alpha_j\right)\right] \leq \\ &\leq \frac{1}{\underline{\gamma}^{i+1-k}}\left[Tr(B_k) + \frac{\overline{\gamma}}{\underline{\rho}}\frac{\overline{G}}{1-\underline{\gamma}} - K\left(i\underline{C} - (k-1)\underline{C}\right)\right] \end{aligned}$$

Zvolíme-li

$$i > \frac{1}{K\underline{C}}\left(Tr(B_k) + \frac{\overline{\gamma}}{\underline{\rho}}\frac{\overline{G}}{1-\underline{\gamma}} + (k-1)K\underline{C}\right)$$

Je výraz v hranaté závorce záporný, což je spor, neboť stopa SPD matice je kladná.

Poznámka 33 Věta 30 je nejobecnějším doposud známým tvrzením o globální konvergenci metod s proměnnou metrikou. Tato věta vyžaduje aby byla splněna podmínka (F4) (exitence konstanty $\underline{C} > 0$), takže ji lze použít pouze pro konvexní funkce.

Poznámka 34 Věta 30 teoreticky zdůvodňuje špatné konvergenční vlastnosti metody DFP (s nepřesným výběrem délky kroku). Metoda DFP odpovídá volbě $\beta = 1$, $i \in N$, takže není splněn předpoklad $\beta_i \leq 1 - \lambda < 1$, $i \in N$. Ve vztahu (T) vymizí poslední člen a nelze pak použít princip důkazu.

Nyní uvedeme tvrzení o superlineární konvergenci metod s proměnnou metrikou. Jelikož superlineární konvergence vyžaduje aby v jistém smyslu platilo $B_i \to G^*$ (věta 13), budeme požadovat splnění předpokladů věty 22 ($\rho_i = 1$, $\gamma_i = 1$, $\forall i \in N$).

Tvrzení 31 (Superlineární konvergence) Nechť jsou splněny předpoklady lemmatu 29 s $\rho_i = 1$, $\gamma_i = 1$, a nechť navíc $x_i \to x^*$, funkce $F : R^n \to R$ splňuje podmínku (F5) a $\alpha_i = 1$ se vybírá vždy, když tato hodnota vyhovuje slabé Wolfeho podmínce. Pak jestliže

$$\sum_{\substack{i=1 \\ \beta_i < 0}}^{\infty} \frac{\beta_i}{\beta_i^*} < \infty$$

platí

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

Poznámka 35 K důkazu tvrzení 31 se používá invariance Broydenovy třídy metod s proměnnou metrikou k transformaci proměnných. Označíme-li $x = T\hat{x}$ pak $\hat{g} = T^T g$, $\hat{G} = T^T G T$ a také $d = T\hat{d}$, $\hat{y} = T^T y$. Označíme-li $\hat{B} = T^T B T$, $\hat{B}^+ = T^T B^+ T$, pak z (B) plyne

$$\hat{B}^+ = \frac{1}{\gamma}\left(\hat{B} + \frac{\gamma}{\rho}\frac{1}{\hat{b}}\hat{y}\hat{y}^T - \frac{1}{\hat{c}}\hat{B}\hat{d}(\hat{B}\hat{d})^T + \frac{\beta}{\hat{c}}\left(\frac{\hat{c}}{\hat{b}}\hat{y} - \hat{B}\hat{d}\right)\left(\frac{\hat{c}}{\hat{b}}\hat{y} - \hat{B}\hat{d}\right)^T\right) \tag{$\hat{\mathrm{B}}$}$$

kde $\hat{a} = \hat{y}^T \hat{B}^{-1}\hat{y}$, $\hat{b} = \hat{y}^T \hat{d}$, $\hat{c} = \hat{d}^T \hat{B}\hat{d}$ a kde parametry β, γ, ρ jsou stejné jako v (B). Pro teoretické účely se pokládá $T = (G^*)^{-\frac{1}{2}}$.

Nyní uvedeme několik poznámek k implementaci metod sproměnnou metrikou.

1) Výběr délky kroku: Metody s proměnnou metrikou nejsou citlivé na výběr délky kroku. Je možné použít algoritmus 1 beze změny. Volí se počáteční odhad

$$\alpha = \min\left(1, \frac{4(\underline{F} - F_i)}{s_i^T g_i}\right)$$

2) Stabilizace (parametr ρ): Označme $\varphi(\lambda) = F(x + \lambda d)$. Parametr ρ se volí tak, aby platilo $d^T B_+ d \approx \varphi''(1)$. Použitím zpětného rozvoje $\varphi(0) = \varphi(1) - \varphi'(1) + \varphi''(\hat{\lambda})/2$, kde $0 \leq \hat{\lambda} \leq 1$, můžeme psát $\varphi''(\hat{\lambda}) = 2(\varphi(0) - \varphi(1) + \varphi'(1))$, takže po dosazení $d^T B_+ d = \varphi''(\hat{\lambda}) \approx \varphi''(1)$ do (VM3) dostaneme $d^T y/\rho = d^T B_+ d = 2(\varphi(0) - \varphi(1) + \varphi'(1))$, což dává

$$\rho = \frac{d^T y}{2(F - F_+ + d^T g_+)}$$

Tuto hodnotu používáme pouze tehdy, jestliže $0.01 \leq \rho \leq 100$, v opačném případě pokládáme $\rho = 1$.

3) Škálování (parametr γ): Jelikož podle (VM3) má platit $B_+ d = y/\rho$, je výhodné volit parametr γ tak, aby Bd/γ bylo co nejblíže k y/ρ. Tedy například

$$d^T B d/\gamma = d^T y/\rho \Rightarrow \gamma/\rho = c/b$$

nebo

$$y^T d/\gamma = y^T B^{-1} y/\rho \Rightarrow \gamma/\rho = b/a$$

Další možnost je geometrický střed $\gamma/\rho = (c/a)^{\frac{1}{2}}$. Tedy k danému ρ najdeme $\gamma = \rho c/b$ nebo $\gamma = \rho b/a$ nebo $\gamma = \rho(c/a)^{\frac{1}{2}}$. Pokud pro takto získanou hodnotu neplatí $0.5 \le \gamma \le 4.0$ pokládáme $\gamma = 1$.

4) Výběr konkretni metody (parametr β): Praktické zkušenosti ukázují, že z jednoduchých metod je nejůčinnější metoda BFGS a že metoda DFP je velmi špatná. Ačkoliv metodu BFGS lze překonat některými složitějšími metodami, stabilizace a škálování rozdíly mezi nimi stírají (s celkovým zlepšením účinnosti), takže lze doporučit stabilizovanou a škálovanou metodu BFGS realizovanou algoritmem 3.

Algoritmus 3 (VM) Data $\varepsilon_1 = 10^{-4}$, $\varepsilon_2 = 0.9$, $\underline{\varepsilon} > 0$, $\underline{\rho} = 0.01$, $\overline{\rho} = 100$, $\underline{\gamma} = 0.5$, $\overline{\gamma} = 4$.

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 $\| g_i \| \le \underline{\varepsilon}$ ukončíme výpočet. V opačném případě položíme $s_i = -H_i g_i$ a určíme délku kroku α_i použitím algoritmu 1. Položíme $x_{i+1} = x_i + \alpha_i s_i$ a vypočteme $F_{i+1} = F(x_{i+1})$, $g_{i+1} = g(x_{i+1})$.

Krok 3 Položíme $d_i = x_{i+1} - x_i$ a $y_i = g_{i+1} - g_i$. Určíme parametr ρ_i podle 2. Jestliže $\rho_i < \underline{\rho}$ nebo $\rho_i > \overline{\rho}$ položíme $\rho_i = 1$. Určíme parametr γ_i podle 3. Jestliže $i > 1$ a současně $\gamma_i < \underline{\gamma}$ nebo $\gamma_i > \overline{\gamma}$, položíme $\gamma_i = 1$. Zvolíme $\eta_i = 1$ (metoda BFGS). Určíme matici H_{i+1} podle (H), zvětšíme i o 1 a přejdeme na krok 2.

Následující tabulka ukazuje srovnání tří metod (CG - metoda sdružených gradientů, VM metoda s proměnnou metrikou, MN modifikovaná Newtonova metoda) při minimalizaci 24 testovacích funkcí s 10, 20, 40 a 80 proměnnými (jsou uvedeny celkové počty iterací NIT a funkčních hodnot NFV, jakož i celkový čas výpočtu).

Metoda	n	NIT - NFV	Čas
CG	10	1211 - 2643	1.54
	20	1346 - 2897	3.52
	40	2105 - 4397	11.59
	80	4035 - 8125	54.37
VM	10	959 - 1125	1.10
	20	1115 - 1261	2.53
	40	1491 - 1666	7.85
	80	2200 - 2433	35.37
MN	10	425 - 527	1.15
	20	407 - 508	3.30
	40	433 - 571	15.54
	80	514 - 686	1:54.24

3. Metody s lokálně omezeným krokem

3.1. Základní vlastnosti metod s lokálně omezeným krokem

Při výkladu metod s lokálně omezeným krokem budeme používat označení

$$Q_i(s) = g_i^T s + \frac{1}{2} s^T B_i s$$

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. 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 22 Řekneme, že základní minimalizační 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

$$\| s_i \| \leq \Delta_i \tag{T1a}$$

$$\| s_i \| < \Delta_i \Rightarrow \| \omega_i(s_i) \| \leq \overline{\omega} \tag{T1b}$$

$$-Q_i(s_i) \geq \underline{\sigma} \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \tag{T1c}$$

kde $0 < \underline{\sigma} < 1$, a $0 \leq \overline{\omega} < 1$ a kde délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq 0 \Rightarrow \alpha_i = 0 \tag{T2a}$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1 \tag{T2b}$$

Přitom posloupnost $\Delta_i > 0$, $i \in N$, se konstruuje tak, že

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta} \| s_i \| \leq \Delta_{i+1} \leq \overline{\beta} \| s_i \| \tag{T3a}$$

$$\rho_i(s_i) \geq \underline{\rho} \Rightarrow \Delta_i \leq \Delta_{i+1} \leq \overline{\gamma} \Delta_i \tag{T3b}$$

kde $0 < \underline{\beta} \leq \overline{\beta} < 1 < \overline{\gamma}$ a $0 < \underline{\rho} < 1$.

Poznámka 36 Jestliže $\overline{\omega} = 0$ nebo $\overline{\omega} > 0$ dostaneme přesné nebo nepřesné metody s lokálně omezeným krokem.

Poznámka 37 Normy v (T1) a (T3) mohou být i jiné než euklidovské. Některé podmínky mohou být oslabeny.

Poznámka 38 Označíme $N_1 \subset N$ množinu indexů takových, že platí (T1b), $N_2 \subset N$ množinu indexů takových, že platí (T2b) a $N_3 \subset N$ množinu indexů takových, že platí (T3b).

Lemma 32 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

$$\| s_i \| \geq \underline{c} m_i / M_i \tag{$*$}$$

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 $i \notin N_3$. Pak podle definice množiny N_3 a funkce $Q_i(s)$ platí

$$F(x_i+s_i)-F(x_i) \geq \underline{\rho} Q_i(s_i) = \underline{\rho}\left(g_i^T s_i + \frac{1}{2} s_i^T B_i s_i\right) \geq \underline{\rho}\left(g_i^T s_i - \frac{1}{2} \| B_i \| \| s_i \|^2\right)$$

Z druhé strany (věta o střední hodnotě) 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

$$\frac{1}{2}(\overline{G}+\underline{\rho} \| B_i \|) \| s_i \|^2 \geq (\underline{\rho}-1) g_i^T s_i$$

z (T1c) dostaneme

$$-\underline{\sigma} \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \geq Q_i(s_i) \geq g_i^T s_i - \frac{1}{2} \| B_i \| \| s_i \|^2$$

což spolu s předchozí nerovností dává

$$\begin{aligned} \frac{1}{2}(\overline{G}+\underline{\rho} \| B_i \|) \| s_i \|^2 &\geq (\underline{\rho}-1) g_i^T s_i \geq \frac{1}{2}(\underline{\rho}-1) \| B_i \| \| s_i \|^2 - \\ &- \underline{\sigma}(\underline{\rho}-1) \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|) \end{aligned}$$

neboli

$$\frac{1}{2}(\overline{G}+ \| B_i \|) \| s_i \|^2 \geq \underline{\sigma}(1-\underline{\rho}) \| g_i \| \min(\| s_i \|, \| g_i \| / \| B_i \|)$$

takže buď $\| s_i \| \geq \| g_i \| / \| B_i \| \geq m_i/M_i$ nebo

$$\frac{1}{2}(\overline{G} M_i / \| B_1 \| + M_i) \| s_i \| \geq \underline{\sigma}(1-\underline{\rho}) \| g_i \|$$

což dává $\| s_i \| \geq [2\underline{\sigma}(1-\underline{\rho}) \| B_1 \| / (\overline{G}+ \| B_1 \|)] m_i/M_i$
(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

$$\| s_i \| \geq \Delta_i \geq \Delta_{k+1} \geq \min(\Delta_k, \underline{\beta} \| s_k \|) \geq \min(\| s_k \|, \underline{\beta} \| s_k \|) \geq \underline{\beta} \| s_k \|$$

takže podle (a)-(c) platí

$$\| s_i \| \geq \underline{\beta} \| s_k \| \geq \underline{c} m_k/M_k \geq \underline{c} m_i/M_i$$

kde

$$\underline{c} = \underline{\beta} \min\left((1-\overline{\omega}), \frac{2\underline{\sigma}(1-\underline{\rho}) \| B_1 \|}{\overline{G}+ \| B_1 \|}, \frac{\| s_1 \| \| B_1 \|}{\| g_1 \|}\right)$$

Věta 33 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 32. 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 $\underline{\varepsilon} > 0$ tak, že $\| g_i \| \geq \underline{\varepsilon}$ $\forall i \in N$. Pak podle lemmatu 32 platí

$$\| s_i \| \geq \frac{\underline{c}\underline{\varepsilon}}{M_i} \qquad (*)$$

$\forall i \in N$. Protože $N_3 \subset N_2$, můžeme psát

$$F_i - F_{i+1} = F(x_i) - F(x_i + s_i) \geq -\underline{\rho} Q_i(s_i) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon} \min(\| s_i \|, \underline{\varepsilon}/M_i) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon}^2 \underline{c}/M_i$$

$\forall i \in N_3$, takže

$$F_1 - \underline{F} \geq \lim_{i\to\infty} (F_1 - F_{i+1}) = \sum_{i=1}^{\infty} (F_i - F_{i+1}) \geq \sum_{i \in N_3} (F_i - F_{i+1}) \geq \underline{\rho}\underline{\sigma}\underline{\varepsilon}^2 \underline{c} \sum_{i \in N_3} \frac{1}{M_i}$$

Platí tedy

$$\sum_{i \in N_3} \frac{1}{M_i} < \infty$$

(b) Nechť $i \in N$, nechť r je přirozené číslo takové, že $\overline{\beta}^{r-1}\overline{\gamma} < 1$ (takové číslo existuje neboť $\overline{\beta} < 1$ a $\overline{\gamma} < \infty$) 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 $[1, i] \cap N_3$). Označme

$$N_4 = \{i \in N : rp(i) < i\}$$

Pak podle (T1a) a (T3) platí

$$\Delta_i \leq \overline{\gamma}^{p(i-1)} \overline{\beta}^{i-1-p(i-1)} \Delta_1 \leq \overline{\gamma}^{(i-1)/r} \overline{\beta}^{(r-1)(i-1)/r} \Delta_1 \leq \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{(i-1)/r}$$

Protože podle předpokladu je $\overline{\gamma}\overline{\beta}^{r-1} < 1$, můžeme psát

$$\sum_{i \in N_4} \Delta_i \leq \sum_{i \in N_4} \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{(i-1)/r} \Delta_1 \leq \sum_{i=1}^{\infty} \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{(i-1)/r} \Delta_1 = \frac{\Delta_1}{1 - \left(\overline{\gamma}\overline{\beta}^{(r-1)}\right)^{1/r}} < \infty$$

Použijeme-li nyní (T1a) a $(*)$, dostaneme

$$\sum_{i \in N_4} \frac{1}{M_i} \leq \frac{1}{\underline{c}\underline{\varepsilon}} \sum_{i \in N_4} \| s_i \| \leq \frac{1}{\underline{c}\underline{\varepsilon}} \sum_{i \in N_4} \Delta_i < \infty$$

(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

$$\begin{aligned} N_3 &= \{i_1, i_2, i_3 \ldots\} \\ N_5 &= \{k_1, k_2, k_3 \ldots\} \end{aligned}$$

(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 39 Předpoklady věty 33 jsou splněny například tehdy, jestliže $\| B_i \| \leq \overline{B}$ $\forall i \in N$. Důkaz tohoto dílčího tvrzení je velmi jednoduchý. Stačí část (a) důkazu věty 33 pozměnit tak, že

$$F_1 - \underline{F} \geq \lim_{i \to \infty} (F_1 - F_{i+1}) = \sum_{i=1}^{\infty} (F_i - F_{i+1}) \geq \sum_{i \in N_3} (F_i - F_{i+1}) \geq \underline{\rho \sigma \varepsilon}^2 \underline{c} \sum_{i \in N_3} \frac{1}{\overline{B}}$$

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 $(*)$.

Věta 34 (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

$$\lim_{i \to \infty} \| \omega_i(s_i) \| = 0 \tag{α}$$

a

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

Pak posloupnost x_i, $i \in N$, konverguje Q-superlineárně k bodu $x^* \in R^n$.

Důkaz (a) Naprosto stejným způsobem jako v důkazu věty 13 se ukáže, že existuje index $k_1 \in N$ takový, že

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

$\forall i \geq k_1$ pokud $\underline{G} < \underline{\lambda}(G^*)$ a $\overline{G} > \overline{\lambda}(G^*)$.
(b) Ukážeme, že existuje index $k_2 \geq k$ takový, že

$$-Q_i(s_i) \geq \frac{1}{2} \underline{G} \| s_i \|^2$$

$\forall i \geq k_2$ pokud $\underline{G} < \underline{\lambda}(G^*)$. Označíme-li tak jako v důkazu věty 13 $\omega_i = (B_i s_i + g_i) / \| g_i \|$ a $\vartheta_i = (G^* - B_i) s_i / \| s_i \|$, můžeme psát

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

Jelikož $\| \omega_i \| \to 0$ a $\| \vartheta_i \| \to 0$ (podle (α) a (β)) a $\underline{\lambda}(G_i) \to \underline{\lambda}(G^*) > \underline{G}$, existuje index $k_2 \geq k_1$ tak, že $-Q_i(s_i) \geq (1/2) \underline{G} \| s_i \|^2$ $\forall i \geq k_2$

(c) Ukážeme, že existuje index $k_3 \geq k_2$ tak, že $i \in N_3$ $\forall i \geq k_3$. Podle věty o střední hodnotě 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

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

Podle (b) však platí

$$\left| \frac{s_i^T (G_i - B_i) s_i + o(\| s_i \|^2)}{2 Q_i(s_i)} \right| \leq \frac{\| \vartheta_i \| \| s_i \|^2 + o(\| s_i \|^2)}{\underline{G} \| s_i \|^2} \rightarrow 0$$

neboť $\| \vartheta_i \| \rightarrow 0$. Platí tedy $\rho_i(s_i) \rightarrow 1$ a jelikož $\underline{\rho} < 1$, existuje index $k_3 \geq k_2$ takový, že $\rho_i(s_i) \geq \underline{\rho}$ $\forall i \geq k_3$.
(d) Ukážeme, že existuje index $k \geq k_3$ takový, že $i \in N_1$ $\forall i \geq k$. Podle věty o střední hodnotě platí (pro $i \in N_2 \subset N_3$)

$$g_{i+1} = g(x_i + s_i) = g_i + G_i s_i + o(\| s_i \|)$$

Označme

$$\lambda_i = \frac{g_{i+1} - g_i - B_i s_i}{\| g_i \|}$$

Pak podle předchozí úvahy platí $\| \lambda_i \| \leq \| \vartheta_i \| / \underline{G} + o(\| s_i \|) / \| s_i \| \rightarrow 0$. Jelikož zároveň $\| \omega_i \| \rightarrow 0$, existuje index $k_4 \geq k_3$ takový, že $\| \lambda_i \| < (\underline{G}/\overline{G})/2$ a $\| \omega_i \| < (\underline{G}/\overline{G})/2$ $\forall i \geq k_4$. Jestliže $i \notin N_1$ $\forall i \geq k_4$, pak $\| s_i \| \geq \Delta_i$ $\forall i \geq k_4$ a jelikož $i \in N_3$ $\forall i \geq k_4$, dostaneme $\| s_i \| \geq \Delta_{k_4}$ $\forall i \geq k_4$, což je spor s tím, že $\| s_i \| \rightarrow 0$. Existuje tedy index $k \geq k_4$ takový, že $k \in N_1$. Předpokládejme, že $i \in N_1$ pro nějaký index $i \geq k$ (platí to jistě pro $i = k$). Pak můžeme psát

$$\begin{aligned} \| s_{i+1} \| &\leq \frac{1}{\underline{G}} \| g_{i+1} \| \leq \frac{1}{\underline{G}} (\| g_{i+1} - g_i - B_i s_i \| + \| B_i s_i + g_i \|) \leq \\ &\leq \frac{\overline{G}}{\underline{G}} (\| \lambda_i \| + \| \omega_i \|) \| s_i \| < \left(\frac{1}{2} + \frac{1}{2} \right) \| s_i \| = \| s_i \| \end{aligned}$$

Jelikož $i \in N_3$ podle (c), 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 $i \in N_1$ $\forall i \geq k$.
(e) 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 $\| \lambda_i \| \rightarrow 0$ a $\| \omega_i \| \rightarrow 0$ dává

$$\lim_{i \rightarrow \infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq \frac{\overline{G}}{\underline{G}} \lim_{i \rightarrow \infty} \frac{\| g_{i+1} \|}{\| g_i \|} = 0$$

3.2. Metody s optimálním lokálně omezeným krokem

Definice 23 Metody s optimálním lokálně omezeným krokem používají směrové vektory $s_i \in R^n$, $i \in N$, takové, že

$$s_i = \arg \min_{\| s \| \leq \Delta_i} Q_i(s) \tag{$\overline{\mathrm{T}}1$}$$

Věta 35 Směrový vektor určený podle ($\overline{\mathrm{T}}1$) vyhovuje podmínkám (T1) s $\overline{\omega} = 0$ a $\underline{\sigma} = 1/2$.

Důkaz Podmínka (T1a) je přímo součástí podmínky ($\overline{\mathrm{T}}1$).
(a) Předpokládejme, že $s_i \in R^n$ je řešením úlohy ($\overline{\mathrm{T}}1$), přičemž $\| s_i \| < \Delta_i$. Pak nutně $Q_i(s)$ je 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 ($\overline{\mathrm{T}}1$) 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 $\| s \| > \Delta_i$, 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 ($\overline{\mathrm{T}}1$) musí být $Q_i(s_i) \leq Q_i(\tilde{s})$ takže nutně

$$-Q_i(s_i) \geq -Q_i(\tilde{s}) \geq \frac{1}{2} \| g_i \| \| s_i \|$$

3.3. Výpočet optimálního lokálně omezeného kroku

Věta 36 Vektor $s_i \in R^n$ je řešením úlohy ($\overline{\mathrm{T}}1$) 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 Dokážeme nejprve 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 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

$$\begin{aligned}
Q_i(s) - Q_i(s_i) &= (s - s_i)^T g_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= (s_i - s)^T (B_i + \lambda_i I) s_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= \frac{1}{2}(s_i - s)^T (B_i + \lambda_i I)(s_i - s) + \frac{1}{2}\lambda_i (s_i^T s_i - s^T s) = \\
&= \frac{1}{2}(s_i - s)^T (B_i + \lambda_i I)(s_i - s) \geq 0
\end{aligned}$$

takže matice $B_i + \lambda_i I$ musí být pozitivně semidefinitní. 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 dříve), že

$$
\begin{aligned}
Q_i(s) - Q_i(s_i) &= (s - s_i)^T g_i + \frac{1}{2} s^T B_i s - \frac{1}{2} s_i^T B_i s_i = \\
&= \frac{1}{2}(s_i - s)^T (B_i + \lambda_i I)(s_i - s) + \frac{1}{2}\lambda_i (s_i^T s_i - s^T s) \geq \\
&\geq \frac{1}{2}(s_i - s)^T (B_i + \lambda_i I)(s_i - s) \geq 0
\end{aligned}
$$

pro všechny vektory $s \in R^n$ takové, že $\| s \| \leq \| s_i \| = \Delta_i$.

Tvrzení věty 36 tvoří základ algoritmu, ve kterém se nelineární rovnice $\phi(\lambda) \triangleq \frac{1}{\Delta_i} - \frac{1}{\| s_i(\lambda) \|}$ řeší pomocí Newtonovy metody (zde vektor $s_i(\lambda)$ je řešením soustavy lineárních rovnic $(B_i + \lambda I) s_i(\lambda) + g_i = 0$):

Algoritmus 4 Data $0 < \underline{\delta} < 1 < \overline{\delta}$ (obvykle $\underline{\delta} = 0.9$ a $\overline{\delta} = 1.1$).

Krok 1 Určíme $\underline{\gamma}$ jako maximální diagonální prvek matice $-B$. Položíme $\underline{\lambda} = 0$ a $\overline{\lambda} = \| g \| / \Delta + \| B \|$. Položíme $\lambda = \max(\underline{\gamma}, \underline{\lambda})$.

Krok 2 Položíme $\underline{\lambda} = \max(\underline{\gamma}, \underline{\lambda})$. Jestliže $\lambda < \underline{\gamma}$ položíme $\lambda = \sqrt{\underline{\lambda}\overline{\lambda}}$.

Krok 3 Je-li matice $B + \lambda I$ SPD, určíme rozklad $R^T R = B + \lambda I$ a přejdeme na krok 4. V opačném případě určíme vektor $v \in R^n$ takový, že $\| v \| = 1$ a $v^T (B + \lambda I) v < 0$, položíme $\underline{\gamma} = \lambda - v^T (B + \lambda I) v$ a přejdeme na krok 2.

Krok 4 Určíme vektor $s \in R^n$ řešením rovnice $R^T R s + g = 0$. Jestliže $\| s \| > \overline{\delta}\Delta$, položíme $\underline{\lambda} = \lambda$ a přejdeme na krok 6. Jestliže $\underline{\delta}\Delta \leq \| s \| \leq \overline{\delta}\Delta$ ukončíme výpočet. Jestliže $\| s \| < \underline{\delta}\Delta$ a $\lambda = 0$ ukončíme výpočet. Jestliže $\| s \| < \underline{\delta}\Delta$ a $\lambda \neq 0$ 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 $v^T s \geq 0$. Určíme číslo $\alpha \geq 0$ tak, aby platilo $\| s + \alpha v \| = \Delta$. 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 $\underline{\gamma} = \lambda - \| Rv \|^2$ 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

3.4. Nepřesné metody s lokálně omezeným krokem

K určení lokálně omezeného kroku můžeme velmi efektivně použít metodu sdružených gradientů aplikovanou na minimalizaci kvadratické funkce

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

(vynecháváme index i). Metoda sdružených gradientů používá rekurentní vztahy

$$
s_1 = 0, \quad g_1 = g \quad p_1 = -g
$$

a

$$
\begin{gathered}
q_i = B p_i, \quad \alpha_i = \| g_i \|^2 / p_i^T q_i \\
s_{i+1} = s_i + \alpha_i p_i
\end{gathered} \tag{CG}
$$

$$g_{i+1} = g_i + \alpha_i q_i, \qquad \beta_i = \| g_{i+1} \|^2 / \| g_i \|^2$$
$$p_{i+1} = -g_{i+1} + \beta_i p_i$$

pro $1 \le i \le n$.

Poznámka 40 Platí $g_i = Bs_i + g$.

Poznámka 41 Hodnota $\alpha_i = \| g_i \|^2 / p_i^T q_i$ realizuje přesný výběr délky kroku, neboť platí

$$p_i^T g_{i+1} = p_i^T g_i + \alpha_i p_i^T q_i = -\| g_i \|^2 + (\| g_i \|^2 / p_i^T g_i) p_i^T q_i = 0$$

Použili jsme indukční krok

$$p_i^T g_i = -\| g_i \|^2 + \beta_{i-1} p_{i-1}^T g_i = -\| g_i \|^2$$

Věta 37 Aplikujeme-li metodu sdružených gradientů na kvadratickou funkci $Q(s)$ a platí-li $p_i^T B p_i > 0$ pro $1 \le i \le m$, pak

$$\begin{aligned}
Q(s_{i+1}) &\le -\frac{1}{2} \| g \|^2 / \| B \| \\
Q(s_{i+1}) &< Q(s_i) \\
\| s_{i+1} \| &> \| s_i \|
\end{aligned}$$

pro $1 \le i \le m$.

Důkaz Z důkazu věty 15 plyne, že

$$p_j^T B p_i = 0 \qquad \forall 1 \le j < i \le m$$
$$g_j^T g_i = 0 \qquad \forall 1 \le j < i \le m+1$$
$$p_j^T g_i = 0 \qquad \forall 1 \le j < i \le m+1$$

použijeme-li (CG) dostaneme

$$\begin{aligned}
Q(s_{i+1}) &= g^T(s_i + \alpha_i p_i) + \frac{1}{2}(s_i + \alpha_i p_i)^T B (s_i + \alpha_i p_i) = \\
&= Q(s_i) + \alpha_i g^T p_i + \alpha_i s_i^T B p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i = \\
&= Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2}\alpha_i^2 p_i^T B p_i = \\
&= Q(s_i) - \frac{\| g_i \|^4}{p_i^T B p_i} + \frac{1}{2}\frac{\| g_i \|^4}{p_i^T B p_i} = \\
&= Q(s_i) - \frac{1}{2}\frac{\| g_i \|^4}{p_i^T B p_i} < Q(s_i)
\end{aligned}$$

Dále platí

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

neboť

$$\begin{aligned}
p_j^T p_i &= p_j^T(-g_i + \beta_{i-1} p_{i-1}) = \beta_{i-1} p_j^T p_{i-1} = \\
&= \left(\prod_{k=j}^{i-1} \beta_k \right) \| p_j \|^2 = (\| g_i \|^2 / \| g_j \|^2) \| p_j \|^2
\end{aligned}$$

Protože

$$s_2 = s_1 + \frac{\| g_1 \|^2}{p_1^T B p_1} p_1 = -\frac{\| g \|^2}{g^T B g} g$$

platí podle části (b) důkazu věty 35

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

což spolu s $Q(s_{i+1}) < Q(s_i)$ pro $1 \leq i \leq m$ dává $Q(s_{i+1}) < -(1/2) \| g \|^2 / \| B \|$.

Poznámka 42 Jestliže $p_i^T B p_i \leq 0$ pak

$$Q(s_i + \alpha_i p_i) = Q(s_i) + \alpha_i g_i^T p_i + \frac{1}{2} \alpha_i^2 p_i^T B p_i \leq Q(s_i) - \alpha_i \| g_i \|^2 < Q(s_i)$$

pro libovolnou hodnotu $\alpha_i \geq 0$. Jestliže $\| s_i \| \leq \Delta$ a $p_i^T B p_i \leq 0$, určíme číslo $\alpha_i \geq 0$ tak, aby platilo $\| s_i + \alpha_i p_i \| = \Delta$ a položíme $s = s_i + \alpha_i p_i$. Podle věty 37 platí $Q(s) \leq -(1/2) \| g \|^2 / \| B \|$ pro $i \geq 2$. Podle části (c) důkazu věty 35 to platí i pro $i = 1$.

Poznámka 43 Číslo $\alpha_i \geq 0$, pro které platí $\| s_i + \alpha_i p_i \| = \Delta$, určujeme podle vzorce

$$\alpha_i = -p_i^T s_i + \sqrt{(p_i^T s_i)^2 + \Delta^2 - \| s_i \|^2}$$

Poznámka 44 Metoda sdružených gradientů generuje cestu v přístupné oblasti, to znamená křivku $s(t) \in R^n$ takovou, že

$$\frac{d \| s(t) \|}{dt} > 0$$

$$\frac{dQ(s(t))}{dt} < 0$$

Tvrzení věty 37 tvoří základ jednoduchého algoritmu:

Algoritmus 5 Data $0 < \omega < 1$, $0 < \Delta$, $m \geq n$

Krok 1 Položíme $s = 0$, $r = -g$, $\sigma = \| r \|^2$, $p = r$ 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 $s := s + \alpha p$, $r := r - \alpha q$ a $\sigma = \| r \|^2$. Jestliže $\sigma \leq \omega^2 \| g \|^2$ 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 $m = n + 3$).

3.5. Využití směru největšího spádu (metody psí nohy)

Nevýhodou metod s optimálním lokálně omezeným krokem je nutnost řešení n-rozměrné úlohy ($\overline{\mathrm{T}}1$) což vyžaduje opakované řešení soustavy rovnic $(B_i + \lambda I) s_i(\lambda) + g_i = 0$. V průměru se tato soustava řeší 2-3 krát v každém iteračním kroku. Proto se rozměrnější úloha ($\overline{\mathrm{T}}1$) často nahražuje úlohou

$$s_i = \arg \min_{\| s(\alpha,\beta) \| \leq \Delta_i} Q_i(s(\alpha, \beta)) \qquad (\tilde{\mathrm{T}}1)$$

kde

$$s(\alpha, \beta) = \alpha g_i + \beta B_i^{-1} g_i$$

Úloha ($\tilde{\mathrm{T}}1$) 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 ($\tilde{\mathrm{T}}1$) 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 ($\tilde{\mathrm{T}}1$) speciálním přibližným výběrem koeficientů α a β, který se nazývá metodou psí nohy. V tomto textu se budeme zabývat obecnějším algoritmem, který přejde na metodu psí nohy pokud $m = 1$.

Věta 38 Nechť jsou splněny předpoklady věty 37 pro $m \geq 1$, přičemž $\| s_{m+1} \| < \Delta$ a $Bs_{m+1} + g \neq 0$. Nechť $s_n \in R^n$ je vektor takový, že $Bs_n + g = 0$. Pak platí

$$\frac{dQ(s_{m+1} + \alpha(s_n - s_{m+1}))}{d\alpha} = (1-\alpha)(s_n - s_{m+1})^T g_{m+1}$$

Důkaz Jelikož

$$\begin{aligned} Q(s_{m+1} + \alpha(s_n - s_{m+1})) &= g^T(s_{m+1} + \alpha(s_n - s_{m+1})) + \\ &+ \frac{1}{2}(s_{m+1} + \alpha(s_n - s_{m+1}))^T B(s_{m+1} + \alpha(s_n - s_{m+1})) \end{aligned}$$

platí

$$\begin{aligned} \frac{dQ(s_{m+1} + \alpha(s_n - s_{m+1}))}{d\alpha} &= (s_n - s_{m+1})^T g + (s_n - s_{m+1})^T B(s_{m+1} + \alpha(s_n - s_{m+1})) = \\ &= (s_n - s_{m+1})^T B(s_{m+1} - s_n + \alpha(s_n - s_{m+1})) = \\ &= (1-\alpha)(s_n - s_{m+1})^T B(s_{m+1} - s_n) = \\ &= (1-\alpha)(s_n - s_{m+1})^T (Bs_{m+1} + g) = \\ &= (1-\alpha)(s_n - s_{m+1})^T g_{m+1} \end{aligned}$$

Tvrzení věty 38 tvoří základ algoritmu, který je kombinací algoritmu 5 a metody psí nohy:

Algoritmus 6 Data $0 < \Delta$, $m < n$

Krok 1 Jako v algoritmu 5.

Krok 2 Jako v algoritmu 5.

Krok 3 Jako v algoritmu 5.

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 $(s^* - s)^T r > 0$, určíme číslo $\alpha \geq 0$ tak, aby platilo $\| s + \alpha(s^* - s) \| = \Delta$. Jestliže $\alpha \geq 1$ položíme $s := s^*$ a ukončíme výpočet. Jestliže $\alpha < 1$ položíme $s := s + \alpha(s^* - s)$ a ukončíme výpočet. Pokud $(s^* - s)^T r \leq 0$, 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 jednoduchou metodu psí nohy).

3.6. Maticové rozklady pro symetrické indefinitní matice

1) 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 $E = 0$). Na začátku i-tého eliminačního kroku máme matici

$$\begin{bmatrix} R_{(i-1),(i-1)}, & R_{(i-1),i}, & R_{(i-1),(n-i)} \\ *, & B_{ii}^{(i-1)}, & B_{i,(n-i)}^{(i-1)} \\ *, & *, & B_{(n-i),(n-i)}^{(i-1)} \end{bmatrix}$$

kde horní index v závorce značí počet již provedených eliminačních kroků a dolní indexy v závorkách značí submatice s $(i-1)$ řádky nebo $(n-i)$ sloupci. Eliminační krok vypadá takto:

$$\gamma_i = \max_{i<j\leq n}(|B_{i,j}^{(i-1)}|)$$

$$\rho_i^2 = \max\left(|B_{ii}^{(i-1)}|, \frac{\gamma_i^2}{\beta^2}, \delta^2\right)$$

$$R_{ii} = \rho_i$$

$$R_{i,(n-i)} = B_{i,(n-i)}^{(i-1)}/R_{ii}$$

$$B_{(n-i),(n-i)}^{(i)} = B_{(n-i),(n-i)}^{(i-1)} - R_{i,(n-i)}^T R_{i,(n-i)}$$

kde δ je malé číslo a $\beta > \sqrt{\| B \|}$). Tento proces se od Choleského rozkladu liší pouze tím že může platit $\rho_i^2 \neq B_{ii}^{(i-1)}$. Bližším rozborem uvedených vztahů se dá dokázat že pro prvky matice E platí

$$E_{ii} = \rho_i^2 - B_{ii}^{(i-1)} = \rho_i^2 + R_{i,(n-i)}R_{i,(n-i)}^T - B_{ii}$$

kde B_{ii} je prvek původní matice.

Věta 39 Nechť $R^T R = B + E$ 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

$$v^T B v = \frac{B_{kk}^{(k-1)}}{\rho_k^2} < 0$$

Důkaz Z rovnice $Rv = e_k$ plyne, že $v_k = 1/\rho_k$. Platí tedy

$$\begin{aligned} v^T B v &= v^T(B+E)v - v^T E v \leq v^T R^T R v - v_k^2 E_{kk} = \\ &= e_k^T e_k - E_{kk}/\rho_k^2 = \frac{\rho_k^2 - E_{kk}}{\rho_k^2} = \frac{B_{kk}^{(k-1)}}{\rho_k^2} \end{aligned}$$

Není-li matice B pozitivně semidefinitní, musí existovat index $1 \leq i \leq n$ tak, že $E_{ii} \neq 0$, neboli $\rho_i^2 \neq B_{ii}^{(i-1)}$. Mohou nastat dva případy. Buď $\rho_i^2 = |B_{ii}^{(i-1)}| \neq B_{ii}^{(i-1)}$, takže $B_{ii}^{(i-1)} < 0$ a tedy i $B_{kk}^{(k-1)} < 0$, nebo $\rho_i^2 = \gamma_i^2/\beta^2$. Ve druhém případě musí existovat index $i < j \leq n$ tak, že $\gamma_i = |B_{ij}^{(i-1)}|$, takže

$$|R_{ij}| = \frac{|B_{ij}^{(i-1)}|}{\rho_i} = \frac{\gamma_i}{\gamma_i/\beta} = \beta$$

což dává

$$B_{ii}^{(i-1)} = \rho_i^2 - E_{ii} = B_{ii} - R_{i,(n-i)}R_{i,(n-1)}^T \leq B_{ii} - \beta^2 < \| B \| - \| B \| = 0$$

2) 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). 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

$$B^{(i-1)} \rightarrow \begin{bmatrix} D_{ii}, & L_{i,(m-i)} \\ *, & B^{(i)} \end{bmatrix}$$

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 40 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 $\underline{\lambda}(B_{ii})$, pokud $\underline{\lambda}(B_{ii}) < 0$. 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

$$v^T B v = \sum_{\underline{\lambda}(D_{ii})<0} \underline{\lambda}(D_{ii}) \leq \underline{\lambda}(D_{kk}) < 0$$

3.7. 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 41 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 33 (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 34).

Nejpoužívanější jsou tyto realizace Newtonovy metody:

1) 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).

2) Newtonova metoda s optimálním lokálně omezeným krokem. Pro tuto realizaci platí obzvláště silné tvrzení:

Tvrzení 42 Nechť jsou splněny předpoklady (F1)-(F3). Nechť x_i, $i \in N$ je posloupnost určená Newtonovou metodou s optimálním lokálně omezeným krokem. Pak existuje hromadný bod $x^* \in R^n$ posloupnosti x_i, $i \in N$ takový, že $g(x^*) = 0$ a $G(x^*) \geq 0$. Nechť navíc bod $x^* \in R^n$ vyhovuje postačujícím podmínkám pro extrém ($g(x^*) = 0$ a $G(x^*) > 0$). 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 $x^* \in R^n$.

Následující tabulka ukazuje srovnání několika realizací Newtonovy metody (S - metoda spádových směrů, T - metoda s lokálně omezeným krokem, G - Gillův-Murrayův rozklad, B - Bunchův-Parlettův rozklad) při minimalizaci 15 testovacích funkcí s 20 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG, jakož i počet selhání a celkový čas výpočtu).

Metoda	NIT - NFV - NFG	selhání	čas
S - G	312 - 508 - 508	1	7.63
S - B	342 - 488 - 488	1	16.03
T - G (optimální)	281 - 321 - 296	-	4.61
T - B (psí noha)	292 - 325 - 307	-	10.38

3.8. Gaussova-Newtonova metoda pro součet čtverců

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

$$B_i = J_i^T J_i = \sum_{k=1}^{m} g_k(x_i)g_k^T(x_i)$$

Zdůvodnění:

1) Úlohy s nulovým reziduem ($F(x^*) = 0$). Z $x_i \to x^*$ plyne $F(x_i) \to F(x^*) = 0$ a tedy $f_k(x_i) \to 0$ $\forall 1 \leq k \leq m$. Jestliže $\| G_k(x) \| \leq \overline{G}$, pak i

$$\| C(x_i) \| = \| \sum_{k=1}^{m} f_k(x_i)G_k(x_i) \| \leq \overline{G} \sum_{k=1}^{m} |f_k(x_i)| \to 0$$

a tedy $\| G(x_i) - B_i \| = \| C(x_i) \| \to 0$ 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 $F(x) \geq 0$ $\forall x \in R^n$. Podmínku (F3) nahradíme podmínkou

$$\| G_k(x) \| \leq \overline{G} \qquad (\overline{\mathrm{F3}})$$

$\forall x \in R^n$, $\forall 1 \leq k \leq m$. Z (F2) a ($\overline{\mathrm{F3}}$) plyne omezenost gradientů i funkčních hodnot

$$\| g_k(x) \| \leq \overline{g}$$

$$|f_k(x)| \leq \overline{f}$$

$\forall x \in \mathcal{L}(F(x_1))$, $\forall 1 \leq k \leq m$, a tudíž i (F3).

Věta 43 Nechť jsou splněny podmínky (F2) a ($\overline{\mathrm{F3}}$). 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 ($\overline{\mathrm{F3}}$) plyne $\| G_k(x) \| \leq \overline{G}$, $\| g_k(x) \| \leq \overline{g}$, $|f_k(x)| \leq \overline{f}$ $\forall x \in R^n$, $\forall 1 \leq k \leq m$. Platí tedy jednak

$$\| G(x) \| \leq \sum_{k=1}^{m} \| g_k(x) \|^2 + \sum_{k=1}^{m} |f_k(x)| \, \| G_k(x) \| \leq m\overline{g}^2 + m\overline{f}\overline{G}$$

(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 33 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 34).

Směrový vektor odpovídající Gaussově-Newtonově metodě můžeme určit třemi možnými 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 - 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 $J_i r_i = 0$. 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. Je to vhodné 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).

3.9. Hybridní metody pro součet čtverců

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:

$F_i \to F^* = 0 \Rightarrow$ Gaussova-Newtonova metoda
$F_i \to F^* > 0 \Rightarrow$ Metoda BFGS (s proměnnou metrikou)

Lemma 44 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

$$\lim_{i\to\infty} \frac{F_i - F_{i+1}}{F_i} = \frac{1}{F^*} \lim_{i\to\infty} (F_i - F_{i+1}) = 0$$

Nejjednodušší hybridní metodu můžeme popsat takto: Nechť $B_1 = J_1^T J_1$. Jestliže $i > 1$ a $(F_i - F_{i+1})/F_i > \underline{\vartheta}$, položíme

$$B_{i+1} = J_{i+1}^T J_{i+1}$$

Jestliže $i > 1$ a $(F_i - F_{i+1})/F_i \le \underline{\vartheta}$, 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 $y_i = g_{i+1} - g_i$ a $d_i = x_{i+1} - x_i$. Obvykle $\underline{\vartheta} = 0.01$ pro metody spádových směrů a $\underline{\vartheta} = 0.0001$ pro metody s lokálně omezeným krokem.

Některé další hybridní metody jsou popsány v oddílu 4.7. Následující tabulka ukazuje srovnání několika metod pro minimalizaci součtu čtverců (S - metoda spádových směrů, T - metoda s lokálně omezeným krokem, GN - Gaussova-Newtonova metoda, VM - metoda s proměnnou metrikou, GN+VM - hybridní metoda) při řešení 30 testovacích problémů s 2-12 proměnnými (jsou uvedeny celkové počty iterací NIT, funkčních hodnot NFV a gradientů NFG, jakož i pocet selhani a celkový čas výpočtu).

Metoda	IT - IF - IG	selhání	čas
S - GN	1917 - 2974 - 2974	3	9.67
S - VM	1543 - 3256 - 3256	2	5.71
S - GN+VM	635 - 1037 - 1037	-	4.12
T - GN	605 - 748 - 634	-	2.64
T - VM	2155 - 2542 - 2183	2	7.58
T - GN+VM	536 - 664 - 565	-	2.58

4. 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.

4.1. Metody s proměnnou metrikou s omezenou pamětí

Metody s proměnnou metrikou s omezenou pamětí jsou založeny na použití omezeného počtu kroků metody BFGS.

Lemma 45 Aktualizace získaná metodou BFGS se dá zapsat ve tvaru

$$H_+ = \gamma V^T H V + \frac{\rho}{b} d d^T$$

kde

$$V = I - \frac{1}{b} y d^T$$

přičemž $y = g_+ - g$, $d = x_+ - x$ a $a = y^T H y$, $b = y^T d$.

Důkaz Roznásobením dokazovaného vztahu dostaneme

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

Tentýž výsledek získáme úpravou vztahu

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

odpovídajícího metodě BFGS.

Definice 24 Řekneme, že základní optimalizační metoda je m-krokovou metodou BFGS s omezenou pamětí, jestliže

$$s_i = -H_i^i g_i$$

kde $H_{i-m}^i = I$, a

$$H_{j+1}^i = \gamma_j^i V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T$$

pro $i - m \le j \le i - 1$. Přitom $\gamma_{i-m}^i = b_i / a_i$ a $\gamma_j^i = 1$ pro $i - m < j \le i - 1$.

Věta 46 Pro m-krokovou metodu BFGS s omezenou pamětí platí

$$H_{j+1}^i = \frac{b_i}{a_i} \left(\prod_{k=i-m}^{j} V_k \right)^T \left(\prod_{k=i-m}^{j} V_k \right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k \right)$$

Důkaz (Indukcí) pro $j = i - m$ to platí (stačí dosadit $\gamma_{i-m}^i = b_i / a_i$ a $H_{i-m}^i = I$ do vztahu pro BFGS). Indukční krok:

$$
\begin{aligned}
H_{j+1}^i &= V_j^T H_j^i V_j + \frac{\rho_j}{b_j} d_j d_j^T = \frac{b_i}{a_i} V_j^T \left(\prod_{k=i-m}^{j-1} V_k \right)^T \left(\prod_{k=i-m}^{j-1} V_k \right) V_j + \\
&+ \sum_{l=i-m}^{j-1} \frac{\rho_l}{b_l} V_j^T \left(\prod_{k=l+1}^{j-1} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j-1} V_k \right) V_j + \frac{\rho_j}{b_j} d_j d_j^T = \\
&= \frac{b_i}{a_i} \left(\prod_{k=i-m}^{j} V_k \right)^T \left(\prod_{k=i-m}^{j} V_k \right) + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k \right)^T d_l d_l^T \left(\prod_{k=l+1}^{j} V_k \right)
\end{aligned}
$$

Tvrzení věty 46 ukazuje, že matici H_i^i můžeme určit pomocí $2m$ vektorů d_j, y_j, $i-m \le j \le i-1$, aniž je třeba uchovávat matice H_j^i, $i-m \le j \le i-1$. Tuto matici však nemusíme konstruovat explicitně, stačí počítat vektor $s_i = -H_i^i g_i$, což se provádí pomocí dvou rekurentních vztahů (Strangova formule). Nejprve se počítají zpětnou rekurzí vektory

$$u_j = -\left(\prod_{k=j}^{i-1} V_k \right) g_i$$

pro $i-m \le j \le i-1$. Protože

$$u_j = V_j u_{j+1} = \left(I - \frac{1}{b_j} y_j d_j^T \right) u_{j+1} = u_{j+1} - \frac{d_j^T u_{j+1}}{b_j} y_j$$

pro $i-m \le j \le i-1$, kde $u_i = -g_i$, 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 $i-m \le j \le i-1$. Potom počítáme přímou rekurzí vektory

$$v_{j+1} = \frac{b_i}{a_i} \left(\prod_{k=i-m}^{j} V_k \right)^T u_{i-m} + \sum_{l=i-m}^{j} \frac{\rho_l}{b_l} \left(\prod_{k=l+1}^{j} V_k \right)^T d_l d_l^T u_{l+1}$$

pro $i-m \le j \le i-1$. Protože

$$v_{j+1} = V_j^T v_j + \frac{\rho_j}{b_j} d_j d_j^T u_{j+1} = \left(I - \frac{1}{b_j} d_j y_j^T \right) v_j + \rho_j \sigma_j d_j = v_j + (\rho_j \sigma_j - y_j^T v_j / b_j) d_j$$

kde $v_{i-m} = (b_i/a_i) u_{i-m}$, můžeme psát

$$v_{i-m} = (b_i/a_i) u_{i-m}$$

a

$$v_{j+1} = v_j + (\rho_j \sigma_j - y_j^T v_j) d_j \tag{R2}$$

pro $i-m \le j \le i-1$. Nakonec položíme $s_i = v_i$.

Poznámka 45 Je třeba uchovávat pouze čísla σ_j, $i-m \le j \le i-1$. Všechny vektory u_j, v_j, $i-m \le j \le i$ mohou být uloženy na témže místě v paměti počítače. Celkem tedy potřebujeme uložit $2m+3$ vektorů (d_j, y_j, $i-m \le j \le i-1$ a 3 vektory pro základní optimalizační metodu) a použijeme $O(mn)$ numerických operací.

Tvrzení 47 (Kvadratické ukončení). Nechť x_i, $i \in N$ je posloupnost generovaná m-krokovou metodou BFGS 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 existuje index $k \le n$ tak, že $g_{k+1} = 0$ a $x_{k+1} = x^*$.

Strangova formule (R1) a (R2) je nejstarší a nejjednodušší realizací metody BFGS s omezenou pamětí. Pro některé aplikace jsou výhodnější kompaktní schemata, která nyní odvodíme.

Lemma 48 Nechť $N = -M^{-1}$, kde M je matice vystupující ve větě 23 s $\rho = 1$, $\gamma = 1$. Pak platí

$$N = \begin{bmatrix} \frac{(\eta-1)b^2}{\eta a + (1-\eta)b}, & \frac{\eta a b}{\eta a + (1-\eta)b} \\ \frac{\eta a b}{\eta a + (1-\eta)b}, & a + \frac{\eta a b}{\eta a + (1-\eta)b} \end{bmatrix} \tag{N}$$

Důkaz Z vyjádření matice M (věta 23) plyne

$$N = -M^{-1} = -\frac{1}{\det M} \begin{bmatrix} \frac{\eta - 1}{a}, & \frac{\eta}{b} \\ \frac{\eta}{b}, & \frac{1}{b}(\eta \frac{a}{b} + 1) \end{bmatrix}$$

Dosadíme-li za $-\det M$ vyjádření m vystupující ve větě 27, dostaneme po úpravě tvrzení lemmatu.

Poznámka 46 Pro metodu DFP je $\eta = 0$, takže

$$N = \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix} \tag{ND}$$

Pro metodu BFGS je $\eta = 1$, takže

$$N = \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix} \tag{NB}$$

Lemma 49 Nechť B je čtvercová regulární matice. Pak platí

$$[A, a] \begin{bmatrix} B, & b \\ b^T, & \beta \end{bmatrix}^{-1} \begin{bmatrix} A^T \\ a^T \end{bmatrix} = AB^{-1}A^T + (AB^{-1}b - a)(\beta - b^T B^{-1} b)^{-1}(b^T B^{-1} A^T - a^T)$$

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

$$H_{k+1} = H_k - [d_k, H_k y_k] N_k^{-1} \begin{bmatrix} d_k^T \\ y_k^T H_k \end{bmatrix} \tag{H}$$

kde N_k je matice specifikující konkrétní metodu s proměnnou metrikou. Budeme se snažit nalézt vyjádření

$$H_{k+1} = H_1 - [D_k, H_1 Y_k] \overline{N}_k^{-1} \begin{bmatrix} D_k \\ Y_k^T H_1 \end{bmatrix} \tag{$\overline{\text{H}}$}$$

kde $D_k = [d_1, \dots, d_k]$, $Y_k = [y_1, \dots, y_k]$ a kde $\overline{N}_k$ je symetrická matice řadu $2k$. 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 $(R_k)_{ij} = 0$, $i > j$, a C_k pro diagonální matici řádu k takovou, že $(C_k)_{ij} = d_i^T y_j$, $i = j$, a $(C_k)_{ij} = 0$, $i \ne j$. Abychom zjednodušili zápis budeme v důkazech index k vynechávat a index $k+1$ nahradíme symbolem $+$. V této souvislosti budeme používat označení $D = [d_1, \ldots, d_{k-1}]$, $Y = [y_1, \ldots, y_{k-1}]$ a $R = R_{k-1}$, $C = C_{k-1}$, 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 ($\overline{\mathrm{H}}$) můžeme indukční krok, používaný v důkazech, zapsat ve tvaru

$$\begin{aligned} H_+ &= H - \left[d, H_1 y - [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} y\right] \cdot \\ & \quad N^{-1} \begin{bmatrix} d^T \\ y^T H_1 - y^T [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \end{bmatrix} = \\ &= H - \left([d, H_1 y] - [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right) \cdot \\ & \quad N^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix}\right) \end{aligned} \tag{$*$}$$

kde $\overline{N} = \overline{N}_{k-1}$ a $N = N_k$.

Věta 50 Nechť H_1 je SPD matice a nechť platí (H) s $\rho_k = 1$, $\gamma_k = 1$ a $\eta_k = 0$ (metoda DFP), pro libovolný index $1 \le k \le m$. Pak lze psát

$$H_{k+1} = H_1 - [D_k, H_1 Y_k] \begin{bmatrix} -C_k, & R_k - C_k \\ (R_k C_k)^T, & Y_k^T H_1 Y_k \end{bmatrix}^{-1} \begin{bmatrix} D_k^T \\ Y_k^T H_1 \end{bmatrix} \tag{$\overline{\mathrm{HD}}$}$$

Důkaz Pro $k = 1$ je ($\overline{\mathrm{HD}}$) ekvivalentní s (HD) (s (H) kde matice N je určena pomocí (ND)). Dále budeme postupovat matematickou indukcí. Předpokládejme, že ($\overline{\mathrm{HD}}$) platí pro všechny indexy menší než k. Pro index k můžeme ($\overline{\mathrm{HD}}$) zapsat ve tvaru

$$H_+ = H_1 - [D, H_1 Y, d, H_1 y] \begin{bmatrix} -C, & R - C, & 0, & D^T y \\ (R - C)^T, & Y^T H_1 Y, & 0, & Y^T H_1 y \\ 0, & 0, & -d^T y, & 0 \\ y^T D, & y^T H_1 Y, & 0, & y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \\ d^T \\ y^T H_1 \end{bmatrix}$$

použijeme-li lemma 49 a označíme-li

$$\overline{N} = \begin{bmatrix} -C, & R - C \\ (R - C)^T, & Y^T H_1 Y \end{bmatrix}$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} - \\ & \quad \left([d, H_1 y] - [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right) \cdot \\ & \quad \left(\begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right)^{-1} \cdot \\ & \quad \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix}\right) = \\ &= H - \left([d, H_1 y] - [D, H_1 Y]\, \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix}\right) \cdot \\ & \quad \begin{bmatrix} -d^T y, & 0 \\ 0, & y^T H y \end{bmatrix}^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix}\right) \end{aligned}$$

což je právě vztah $(*)$ s maticí N určenou pomocí (ND) (poznámka 46)

Věta 51 Nechť H_1 je SPD matice a nechť platí (H) s $\rho_k = 1$, $\gamma_k = 1$a $\eta_k = 1$ (metoda BFGS), pro libovolný index $1 \leq k \leq m$. Pak lze psát

$$H_{k+1} = H_1 - [D_k, H_1 Y_k] \begin{bmatrix} 0, & R_k \\ R_k^T, & C_k + Y_k^T H_1 Y_k \end{bmatrix}^{-1} \begin{bmatrix} D_k \\ Y_k^T H_1 \end{bmatrix} \tag{$\overline{\text{HB}}$}$$

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 ve tvaru

$$H_+ = H_1 - [D, H_1 Y, d, H_1 y] \begin{bmatrix} 0, & R, & 0, & D^T y \\ R^T, & C + Y^T H_1 Y, & 0, & Y^T H_1 y \\ 0, & 0, & 0, & d^T y \\ y^T D, & y^T H_1 Y, & d^T y, & d^T y + y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \\ d^T \\ y^T H_1 \end{bmatrix}$$

Použijeme-li lemma 49 a označíme-li

$$\overline{N} = \begin{bmatrix} 0, & R \\ R^T, & C + Y^T H_1 Y \end{bmatrix}$$

dostaneme

$$\begin{aligned} H_+ &= H_1 - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} - \\ & \left([d, H_1 y] - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ & \left(\begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H_1 y \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right)^{-1} \cdot \\ & \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right) = \\ &= H - \left([d, H_1 y] - [D, H_1 Y] \overline{N}^{-1} \begin{bmatrix} 0, & D^T y \\ 0, & Y^T H_1 y \end{bmatrix} \right) \cdot \\ & \begin{bmatrix} 0, & d^T y \\ d^T y, & d^T y + y^T H y \end{bmatrix}^{-1} \left(\begin{bmatrix} d^T \\ y^T H_1 \end{bmatrix} - \begin{bmatrix} 0, & 0 \\ y^T D, & y^T H_1 Y \end{bmatrix} \overline{N}^{-1} \begin{bmatrix} D^T \\ Y^T H_1 \end{bmatrix} \right) \end{aligned}$$

což je právě vztah $(*)$ s maticí N určenou pomocí (ND) (poznámka 46).

Věta 52 Nechť H_1 je SPD matice a nechť platí (H) s $\rho_k = 1$, $\gamma_k = 1$ a $\eta_k = d_k^T y_k / (d_k^T y_k - y_k^T H_k y_k)$ (metoda hodnosti 1), pro libovolný index $1 \leq k \leq m$. Pak lze psát

$$H_{k+1} = H_1 + (D_k - H_1 Y_k)(R_k + R_k^T - C_k - Y_k^T H_1 Y_k)^{-1} (D_k - H_1 Y_k)^T \tag{$\overline{\text{HR}}$}$$

Důkaz Vztah (HR) (poznámka 25) lze zapsat ve tvaru

$$H_+ = H + (d - Hy)(d^T y - y^T H y)^{-1} (d - Hy)^T$$

což 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

$$H_+ = H_1 - [D - H_1 Y, d - H_1 y] \begin{bmatrix} R + R^T - C - Y^T H_1 Y, & D^T y - Y^T H_1 y \\ y^T D - y^T H_1 Y, & d^T y - y^T H_1 y \end{bmatrix}^{-1} \begin{bmatrix} D^T - Y^T H_1 \\ d^T - y^T H_1 \end{bmatrix}$$

Použijeme-li lemma 49 a označíme-li

$$\overline{N} = R + R^T - C - Y^T H_1 Y$$

dostaneme

$$\begin{aligned} H_+ &= H_1 + (D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T + \\ &\quad \left[(D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T y - (d - H_1 y)\right] \cdot \\ &\quad \left(d^T y - y^T H_1 y - y^T (D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T y\right)^{-1} \cdot \\ &\quad \left[(D - H_1 Y)\,\overline{N}^{-1}\,(D - H_1 Y)^T y - (d - H_1 y)\right]^T = \\ &= H + (d - Hy)\left(d^T y - y^T H y\right)^{-1}(d - Hy)^T \end{aligned}$$

což je právě vztah (HR).

Poznámka 47 Podobná kompaktní schemata můžeme odvodit pro matici $B = H^{-1}$. Lze k tomu použít dualitu (poznámka 27). 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 $(L_k)_{ij} = s_i^T y_j$, $i \geq j$. Pro metodu DFP dostaneme

$$B_{k+1} = B_1 - [Y_k, B_1 D_k] \begin{bmatrix} 0, & L_k^T \\ L_k, & C_k + D_k^T B_1 D_k \end{bmatrix}^{-1} \begin{bmatrix} Y_k^T \\ D_k^T B_1 \end{bmatrix} \tag{$\overline{\mathrm{BD}}$}$$

Pro metodu BFGS dostaneme

$$B_{k+1} = B_1 - [Y_k, B_1 D_k] \begin{bmatrix} -C_k, & (L_k - C_k)^T \\ L_k - C_k, & D_k^T B_1 D_k \end{bmatrix}^{-1} \begin{bmatrix} Y_k^T \\ D_k^T B_1 \end{bmatrix} \tag{$\overline{\mathrm{BB}}$}$$

Pro metodu hodnosti 1 dostaneme

$$B_{k+1} = B_1 + (Y_k - B_1 D_k)\left(L_k + L_k^T - C_k - D_k^T B_1 D_k\right)^{-1}(Y_k - B_1 D_k)^T \tag{$\overline{\mathrm{BR}}$}$$

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

$$H_{k+1} = \gamma_k I + [D_k, \gamma_k Y_k] \begin{bmatrix} (R_k^{-1})^T (C_k + \gamma_k Y_k^T Y_k) R_k^{-1}, & -(R_k^{-1})^T \\ -R_k^{-1}, & 0 \end{bmatrix} \begin{bmatrix} D_k^T \\ \gamma_k Y_k^T \end{bmatrix}$$

Lze se o tom přesvědčit explicitním invertováním tak, jako v důkazu lematu 49. 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

$$\begin{aligned} B_{k+1} &= \frac{1}{\gamma_k} I - \left[Y_k, \frac{1}{\gamma_k} D_k\right] \begin{bmatrix} -C_k^{-\frac{1}{2}}, & (\overline{L}_k^{-1}(L_k - C_k))^T \\ 0, & (\overline{L}_k^{-1})^T \end{bmatrix} \cdot \\ &\quad \begin{bmatrix} C_k^{-\frac{1}{2}}, & 0 \\ -\overline{L}_k^{-1}(L_k - C_k), & \overline{L}_k^{-1} \end{bmatrix} \begin{bmatrix} Y_k^T \\ \frac{1}{\gamma_k} D_k^T \end{bmatrix} \end{aligned}$$

kde

$$\overline{L}_k\overline{L}_k^T = (L_k - C_k)^T C_k^{-\frac{1}{2}}(L_k - C_k) + \frac{1}{\gamma_k} D_k^T D_k$$

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.

4.2. Diferenční verze nepřesné Newtonovy metody

Diferenční verze nepřesné Newtonovy metody jsou v podstatě nepřesné metody s lokálně omezeným krokem (algoritmus 5), 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 ($\delta = \sqrt{\varepsilon_M} / \parallel p \parallel$, kde ε_M je strojová přesnost). Jinak se algoritmus 5 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.

4.3. Diferenční verze Newtonovy metody pro řídké úlohy

Diferenční verze Newtonovy metody jsou založeny na aproximaci sloupců Ge_i, $1 \leq i \leq n$, Hessovy matice G pomocí diferenčních vzorců

$$G(x)e_i \approx \frac{g(x + \delta e_i) - g(x)}{\delta}, \qquad 1 \leq i \leq n$$

kde δ je malá diference ($\varepsilon = \sqrt{\varepsilon_M}$). 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 $j \in \{1, \dots, n\} \backslash (S_1 \cup \dots \cup S_{i-1})$ se přidá do skupiny $\mathcal{S}_i$ pouze tehdy, neporuší-li se přitom požadavek, aby submatice $G_i(\mathcal{S}_i)$ měla v každém řádku nanejvýš jeden nenulový prvek.

Na přerovnání sloupců submatice G_i obvykle dosti záleží. Následující matice se liší pouze pořadím řádků a sloupců (nenulové prvky jsou znázorněny symbolem $*$).

$$\begin{bmatrix} * & & & & * \\ & * & & & * \\ & & * & & * \\ & & & * & * \\ * & * & * & * & * \end{bmatrix} \qquad \begin{bmatrix} * & * & * & * & * \\ * & * & & & \\ * & & * & & \\ * & & & * & \\ * & & & & * \end{bmatrix}$$

Probíráme-li sloupce první matice sekvenčně podle vzrůstajících indexů, potřebujeme k určení všech nenulových prvků celkem pět diferencí gradientů. Probíráme-li sloupce druhé matice sekvenčně podle vzrůstajících indexů, stačí k určení všech nenulových prvků pouze dvě diference gradientů.

Zatím jsme se zabývali přímými metodami pro výpočet prvku řídké Hessovy matice pomocí diferencí. Nyní obrátíme pozornost na substituční metody, které obvykle vyžadují menší počet diferencí než přímé metody. Uvažujme opět matici (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 diference gradientů (přímá metoda používá tři diference gradientů).

Postup, který jsme použili v tomto konkrétním případě můžeme snadno zobecnit. Nechť G_U je horní trojúhelníková matice, jejíž horní trojúhelníková část má stejnou strukturu (rozložení nenulových prvků) jako horní trojúhelníková část matice G. Rozdělme sloupce matice G_U do k disjunktních skupin $\mathcal{S}_i \subset \{1, \dots, n\}$, $1 \le i \le k$, 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, $1 \le i \le k$ jsou vektory obsahující pouze nuly a jednotky takové, že

$$(v_i)_j = e_j^T v_i = 1 \Longleftrightarrow j \in \mathcal{S}_i$$

(pomocí vektoru v_i určíme prvky submatice $G_U(\mathcal{S}_i)$, ostatní prvky submatice $G(\mathcal{S}_i)$ jsou určeny symetrií matice G). Při určování prvků matice G je nutné postupovat podle vzrůstajících indexů:

Určujeme-li prvky v j-tém řádku matice G_U, je nutné od prvku označeného kroužkem odečíst prvky označené křížkem, jež se v důsledku symetrie rovnají již určeným prvkům ležícím v j-tém sloupci matice G_U.

Smyslem těchto úvah bylo ukázat, že určení Hessovy matice pomocí diferencí gradientů může být časově nenáročné , je-li tato matice řídká. To staví diferenční verze Newtonovy metody do zcela jiného světla, neboť pro řídké úlohy mohou konkurovat metodám s proměnnou metrikou a metodám sdružených gradientů nebo je i překonat.

Diferenční verze Newtonovy metody pro řídké úlohy se obvykle realizují jako metody s optimálním lokálně omezeným krokem (algoritmus 4) nebo jako nepřesné metody s lokálně omezeným krokem (algoritmus 5). Metody s optimálním lokálně omezeným krokem vyžadují opakované řešení soustavy rovnic $(G + \lambda I)s + g = 0$ (pro různé hodnoty parametru $\lambda \ge 0$). Používá se přitom řídký Choleského rozklad

$$R^T R = P(G + \lambda I)P^T$$

kde R je regulární horní trojúhelníková matice a P je permutační matice, jejíž jediným účelem je přerovnat řádky a sloupce matice $G + \lambda I$ tak, aby počet nově vzniklých nenulových prvků byl co nejmenší. Nalezení permutační matice P a následné určení struktury horní trojúhelníkové matice R se nazývá symbolickou faktorizací. Symbolická faktorizace se provádí pouze jednou (na začátku iteračního procesu) a proto je možné používat časově náročnější složité postupy, které minimalizují počet nově vzniklých nenulových prvků. Tyto postupy mají kombinatorický charakter a jejich popis se vymyká rozsahu tohoto textu (jsou jim věnovány samostatné monografie). Výpočet prvků horní trojúhelníkové matice R (numerická faktorizace) se provádí podle vzorců uvedených v oddílu 3.6.

Nepřesné metody s lokálně omezeným krokem používají metodu sdružených gradientů (CG), kde se řídká Hessova matice G používá pouze k výpočtu součinů $q_i = Gp_i$, $1 \le i \le n$, a není ji tudíž třeba rozkládat. V souvislosti s diferenčními verzemi Newtonovy metody je často výhodné používat předpodmíněnou metodu sdružených gradientů

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

a

$$q_i = Bp_i, \qquad \alpha_i = g_i^T C^{-1} g_i / p_i^T g_i \tag{$\overline{\text{CG}}$}$$

$$s_{i+1} = s_i + \alpha_i p_i$$

$$g_{i+1} = g_i + \alpha_i q_i, \qquad \beta_i = g_{i+1}^T C^{-1} g_{i+1} / g_i^T C^{-1} g_i$$

$$p_{i+1} = -C^{-1} g_{i+1} + \beta_i p_i$$

kterou dostaneme, použijeme-li metodu sdružených gradientů (CG) na minimalizaci kvadratické funkce

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

kde $\tilde{g} = C^{-\frac{1}{2}} g$, $\tilde{s} = C^{\frac{1}{2}} s$, $\tilde{B} = C^{-\frac{1}{2}} B C^{-\frac{1}{2}}$ a položíme-li $g_i = C^{\frac{1}{2}} \tilde{g}_i$, $s_i = C^{-\frac{1}{2}} \tilde{s}_i$, $p_i = C^{-\frac{1}{2}} \tilde{p}_i$, $q_i = C^{\frac{1}{2}} \tilde{q}_i$, $1 \le i \le n$ (vzorce pro $\tilde{g}_i$ a $\tilde{q}_i$ musíme vynásobit maticí $C^{\frac{1}{2}}$ a vzorce pro $\tilde{s}_i$ a $\tilde{p}_i$ musíme vynásobit maticí $C^{-\frac{1}{2}}$).
Význam předpodmínění spočívá v tom, že matice C se volí tak, aby matice $\tilde{B}$ byla lépe podmíněná než matice B. Metoda ($\overline{\text{CG}}$) pak konverguje rychleji než metoda (CG). Ideální volba je $C = B$. V tomto případě najdeme minimum kvadratické funkce již v prvním kroku.

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 25 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

$$\begin{aligned} num\ G_U &= [G_{11}, G_{12}, G_{13}, G_{14}, G_{15}, G_{22}, G_{33}, G_{44}, G_{55}]^T \\ ind\ G_U &= [1,2,3,4,5,2,3,4,5]^T \\ ord\ G_U &= [1,6,7,8,9,10]^T \end{aligned}$$

4.4. 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

$$\begin{aligned} \mathcal{P}_Q B &= \min_{B^+ \in \mathcal{V}_Q} \| B^+ - B \|_F \\ \mathcal{P}_S B &= \min_{B^+ \in \mathcal{V}_S} \| B^+ - B \|_F \\ \mathcal{P}_G B &= \min_{B^+ \in \mathcal{V}_G} \| B^+ - B \|_F \end{aligned}$$

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 53 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 28. Vztah pro $\mathcal{P}_Q B$ odvodíme pomocí Lagrangeovy funkce

$$\begin{aligned} L &= \frac{1}{2} \| B^+ - B \|_F^2 + u^T (w - (B^+ - B)d) = \\ &= \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n (B_{ij}^+ - B_{ij})^2 + \sum_{i=1}^n u_i \left(y_i - \sum_{j=1}^n B_{ij}^+ d_j \right) \end{aligned}$$

Derivováním Lagrangeovy funkce podle prvků matice B_{ij}^+ 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

$$\frac{1}{2} \| B^+ - B \|_F^2 = \frac{1}{2} \sum_{i=1}^n (B_{ii}^+ - B_{ii})^2 + \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=i+1}^n ((B_{ij}^+ - B_{ij})^2 + (B_{ij}^+ - B_{ji})^2)$$

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

$$\frac{1}{2} \| B^+ - B \|_F^2 = \frac{1}{2} \sum_{G_{ij} \neq 0} (B_{ij}^+ - B_{ij})^2 + \frac{1}{2} \sum_{G_{ij} = 0} B_{ij}^2$$

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

$$\begin{aligned} d^i_j &= d_j \ , \ G_{ij} \neq 0 \\ d^i_j &= 0 \ \ , \ G_{ij} = 0 \end{aligned}$$

a místo standardní kvazinewtonovské podmínky $B^+ d = y$ použijeme řídkou kvazinewtonovskou podmínku

$$\sum_{j=1}^{n} (B^+_{ij} - (\mathcal{P}_G B)_{ij}) d^i_j = w_i, \qquad 1 \leq i \leq n$$

kde $w = y - (\mathcal{P}_G B) d$.

Věta 54 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 28 (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

$$L = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (B^+_{ij} - (\mathcal{P}_G B)_{ij})^2 + \sum_{i=1}^{n} u_i \left(w_i - \sum_{j=1}^{n} (B^+_{ij} - (\mathcal{P}_G B)_{ij}) d^i_j \right)$$

obsahující řídkou kvazinewtonovskou podmínku. Derivováním Lagrangeovy funkce podle prvků matice B^+_{ij} dostaneme

$$\frac{\partial L}{\partial B^+_{ij}} = (B^+_{ij} - (\mathcal{P}_G B)_{ij}) - u_i d^i_j$$

$\forall 1 \leq i \leq n$, $\forall 1 \leq j \leq n$. Podmínka pro stacionaritu Lagrangeovy funkce má tedy tvar $B^+_{ij} - (\mathcal{P}_G B)_{ij} = u_i d^i_j$, $1 \leq i \leq n$, $1 \leq j \leq n$, což jsme měli dokázat. Dosadíme-li toto vyjádření do řídké kvazinewtonovské podmínky, dostaneme

$$\sum_{j=1}^{n} u_i d^i_j d^i_j = w_i, \qquad 1 \leq i \leq n$$

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 55 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

$$Q = \mathcal{P}_G(dd^T) + \sum_{i=1}^{n} \| d^i \|^2 e_i e_i^T$$

která má stejnou strukturu jako matice G.

Důkaz Zřejmě platí $\mathcal{P}_{QSG}B = \mathcal{P}_{QSG}\mathcal{P}_G B$. Jelikož matice $B^+ - \mathcal{P}_G B$ 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_{ij} 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

$$\begin{aligned} v^T Q v &= \sum_{i=1}^{n}\sum_{j=1}^{n} d_i^j d_j^i v_i v_j + \sum_{i=1}^{n} \| d^i \|^2 v_i^2 = \sum_{G_{ij}\neq 0} d_i d_j v_i v_j + \sum_{G_{ij}\neq 0} d_j^2 v_i^2 = \\ &= \frac{1}{2}\sum_{G_{ij}\neq 0}(d_i v_j + d_j v_i)^2 \ge 0 \end{aligned} \tag{$*$}$$

(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

$$d_i v_j + d_j v_i = 0 \quad \forall G_{ij} \neq 0$$

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 $1 \leq i \leq n$ 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 29). 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 56 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 57 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 $\mathcal{V}_A \in R^{n \times n}$, $\mathcal{V}_B \in R^{n \times n}$ 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 $\hat{G} \in \mathcal{V}_B$ a $\mathcal{P}_B$ je operátor ortogonální projekce, můžeme použít Pythagorovu větu

$$\| \mathcal{P}_B \mathcal{P}_A B - \hat{G} \|_F^2 = \| \mathcal{P}_A B - \hat{G} \|_F^2 - \| \mathcal{P}_B \mathcal{P}_A B - \mathcal{P}_A B \|_F^2 \leq \| \mathcal{P}_A B - \hat{G} \|_F^2$$

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 $\hat{G} \in \mathcal{V}_A$ a $\mathcal{P}_A$ je operátor ortogonální projekce, můžeme psát

$$\| \mathcal{P}_A B - \hat{G} \|_F^2 = \| B - \hat{G} \|_F^2 - \| \mathcal{P}_A B - B \|_F^2$$

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 58 Nechť $x_i \in R^n$, $i \in N$, je posloupnost bodů generovaná metodou s lokálně omezeným krokem (definice 22). 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 38). 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 platí $\| B_i \| \leq C_i$, $i \in N$, kde C_i, $i \in N$, jsou čísla splňující rekurentní nerovnosti

$$C_{i+1} \leq C_i + \overline{C} \| d_i \|, \qquad i \in N \tag{C}$$

Nechť $i \in N_2$ a nechť $\hat{G}_i$ je matice definovaná vztahem (a) v důkazu lemmatu 29. Pak platí

$$\begin{aligned}
\| \hat{G}_i - G_i \|_F &= \| \int_0^1 (G(x_i + \lambda d_i) - G(x_i)) d\lambda \|_F \leq \sqrt{n} \int_0^1 \| G(x_i + \lambda d_i) - G(x_i) \| d\lambda \leq \\
&\leq \overline{L} \sqrt{n} \| d_i \| \int_0^1 \lambda d\lambda = \frac{1}{2} \overline{L} \sqrt{n} \| d_i \|
\end{aligned}$$

(používáme předpoklad (F5) a skutečnost, že Frobeniova norma není větší než $\sqrt{n}$ násobek spektrální normy). Podobným způsobem dostaneme

$$\| \hat{G}_i - G_{i+1} \|_F \leq \frac{1}{2} \overline{L} \sqrt{n} \| d_i \|$$

Použijeme-li nerovnost $\| B_{i+1} - \hat{G}_i \|_F \leq \| B_i - \hat{G}_i \|_F$, která plyne z lemmatu 57, můžeme psát

$$\begin{aligned}
\| B_{i+1} - G_{i+1} \|_F &\leq \| B_{i+1} - \hat{G}_i \|_F + \| \hat{G}_i - G_{i+1} \|_F \leq \| B_i - \hat{G}_i \|_F + \| \hat{G}_i - G_{i+1} \|_F \leq \\
&\leq \| B_i - G_i \|_F + \| \hat{G}_i - G_i \|_F + \| \hat{G}_i - G_{i+1} \|_F \leq \\
&\leq \| B_i - G_i \|_F + \overline{L} \sqrt{n} \| d_i \|
\end{aligned}$$

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

$$\| B_{i+1} - G_i \|_F \leq \| B_1 - G_1 \|_F + \overline{L} \sqrt{n} \sum_{j=1}^{i} \| d_j \|$$

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) Z předpokladu (F2) plyne existence konstanty $\overline{\Delta}$ takové, že $\| d_i \| \leq \overline{\Delta}$, $i \in N$. Zvolme tuto konstantu tak velkou, aby platilo $C_1 \leq \overline{C\Delta}$. Jelikož $\| d_i \| \leq \overline{\Delta}$, $i \in N$, a jelikož platí (C), můžeme psát $C_i \leq i\overline{C\Delta}$, $i \in N$. Označíme-li

$$M_i = \max_{1 \leq j \leq i} \| B_j \|$$

můžeme psát $M_i \leq C_i$, $i \in N$, takže dostaneme

$$\sum_{i=1}^{\infty} \frac{1}{M_i} \geq \sum_{i=1}^{\infty} \frac{1}{C_i} \geq \frac{1}{\overline{C}\overline{\Delta}} \sum_{i=1}^{\infty} \frac{1}{i} = \infty$$

a můžeme použít větu 33.

Tvrzení 59 Nechť jsou splněny předpoklady věty 58 a nechť $x_i \to x^*$. Nechť funkce $F : R^n \to R$, splňuje podmínky (F3), (F4), (F5). Pak platí

$$\sum_{i=1}^{\infty} \| d_i \| < \infty$$

Poznámka 48 Z tohoto tvrzení, jehož důkaz je formálně dosti komplikovaný, plyne existence konstanty $\overline{B}$ takové, že $\| B_i \| \leq \overline{B}\ \forall i \in N$. Z (C) totiž snadno dostaneme

$$\| B_i \| \leq C_i \leq C_1 + \overline{C} \sum_{j=1}^{i-1} \| d_i \| \leq C_1 + \overline{C} \sum_{j=1}^{\infty} \| d_i \| \stackrel{\Delta}{=} \overline{B}$$

Věta 60 Nechť jsou splněny předpoklady Tvrzení 59, přičemž $\| \omega_i(s_i) \| \to 0$. Pak $x_i \to x^*$ Q-superlineárně.

Důkaz Nechť $i \in N_2$. Použijeme-li lemma 57 a první dvě nerovnosti z důkazu věty 58, můžeme psát

$$\begin{aligned}
\| B_{i+1} - G_{i+1} \|_F^2 &\leq \left(\| B_{i+1} - \hat{G}_i \|_F + \| G_{i+1} - \hat{G}_i \|_F \right)^2 \leq \\
&\leq \| B_{i+1} - \tilde{G}_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_{i+1} - \hat{G}_i \|_F \| d_i \| \leq \\
&\leq \| B_i - \hat{G}_i \|_F^2 - \frac{\| y_i - B_i d_i \|^2}{\| d_i \|} + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}\sqrt{n}\,(\overline{B} + \overline{G}) \right) \| d_i \|
\end{aligned}$$

a

$$\begin{aligned}
\| B_i - \hat{G}_i \|_F^2 &\leq \left(\| B_i - G_i \|_F + \| G_i - \hat{G}_i \|_F \right)^2 \leq \\
&\leq \| B_i - G_i \|_F^2 + \frac{1}{4}\overline{L}^2 n \| d_i \|^2 + \overline{L}\sqrt{n} \| B_i - G_i \|_F \| d_i \| \leq \\
&\leq \| B_i - G_i \|_F^2 + \left(\frac{1}{4}\overline{L}^2 n\overline{\Delta} + \overline{L}\sqrt{n}\,(\overline{B} + \overline{G}) \right) \| d_i \|
\end{aligned}$$

(existence konstanty $\overline{\Delta}$ plyne z tvrzení 59, existence konstanty $\overline{B}$ plyne z poznámky 48, existence konstanty $\overline{G}$ plyne z (F3)). Spojením obou nerovností dosteneme

$$\frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} \leq \| B_i - G_i \|_F^2 - \| B_{i+1} - G_{i+1} \|_F^2 + \overline{M} \| d_i \|$$

kde $\overline{M} = \frac{1}{2}\overline{L}^2 n\overline{\Delta} + 2\overline{L}\sqrt{n}(\overline{B} + \overline{G})$. 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 Tvrzení 59, dostaneme

$$\begin{aligned}
\sum_{i=1}^{\infty} \frac{\| y_i - B_i d_i \|^2}{\| d_i \|^2} &\leq \sum_{i=1}^{\infty} \left(\| B_i - G_i \|_F^2 - \| B_{i+1} - G_{i+1} \|_F^2 \right) + \overline{M} \sum_{i=1}^{\infty} \| d_i \| \leq \\
&\leq \| B_1 - G_1 \|_F^2 + \overline{M} \sum_{i=1}^{\infty} \| d_i \|
\end{aligned} \tag{$*$}$$

Dále podle (F5) platí

$$\begin{aligned}
\frac{\| (G^* - B_i) d_i \|}{\| d_i \|} &\leq \frac{\| G^* d_i - y_i \|}{\| d_i \|} + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\
&\leq \| G^* - G_i \| + \| \hat{G}_i - G_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|} \leq \\
&\leq \overline{L} \| x^* - x_i \| + \frac{1}{2} \overline{L} \sqrt{n} \| d_i \| + \frac{\| y_i - B_i d_i \|}{\| d_i \|}
\end{aligned}$$

(viz důkaz věty 58), takže

$$\frac{\| (G^* - B_i) d_i \|}{\| d_i \|} \to 0$$

neboť $x_i \to x^*$ podle předpokladu, $\| d_i \| \to 0$ podle tvrzení 59 a $\| y_i - B_i d_i \| / \| d_i \| \to 0$ podle $(*)$. Jelikož $\| \omega_i(s_i) \| \to 0$ jsou splněny předpoklady věty 34 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 (oddíl 3.4).

Věta 61 Aplikujeme-li metodu sdružených gradientů (CG) na kvadratickou funkci $Q(s) = (1/2)s^T Bs + g^T s$, kde $\underline{B} \leq \underline{\lambda}(B) \leq \overline{\lambda}(B) \leq \overline{B}$, a platí-li $p_i^T B p_i > 0$ pro $1 \leq i \leq m$, pak, položíme-li $s = s_{m+1}$, je splněna podmínka

$$-s^T g \geq \varepsilon_0 \| s \| \| g \|$$

s

$$\varepsilon_0 = \frac{1}{m} \frac{\underline{B}}{\overline{B}}$$

Důkaz Z důkazu věty 15 plyne, že

$$\begin{aligned}
p_j^T B p_i &= 0 \qquad \forall 1 \leq j < i \leq m \\
g_j^T g_i &= 0 \qquad \forall 1 \leq j < i \leq m+1 \\
p_j^T g_i &= 0 \qquad \forall 1 \leq j < i \leq m+1
\end{aligned}$$

Použijeme-li (CG), dostaneme

$$g_i^T p_i = \left(g + \sum_{j=1}^{i-1} \alpha_j B p_j \right)^T p_i = g^T p_i$$

a

$$\alpha_i = \frac{g_i^T g_i}{p_i^T B p_i} = -\frac{g_i^T (p_i - \beta_{i-1} p_{i-1})}{p_i^T B p_i} = -\frac{g^T p_i}{p_i^T B p_i}$$

$\forall 1 \leq i \leq m$, takže

$$-g^T s = -\sum_{i=1}^{m} \alpha_i g^T p_i = \sum_{i=1}^{m} \frac{(g^T p_i)^2}{p_i^T B p_i} \geq \frac{(g^T p_1)^2}{p_1^T B p_1} = \frac{g^T g}{g^T B g} \| g \|^2 \geq \| g \|^2 / \overline{B}$$

Dále platí

$$s = \sum_{i=1}^{m} \alpha_i p_i = -\sum_{i=1}^{m} \frac{p_i p_i^T}{p_i^T B p_i} g$$

takže

$$\| s \| \leq \sum_{i=1}^{m} \frac{\| p_i p_i^T \|}{p_i^T B p_i} \| g \| = \sum_{i=1}^{m} \frac{p_i^T p_i}{p_i^T B p_i} \| g \| \leq m \| g \| / \underline{B}$$

Spojíme-li obě nerovnosti, dostaneme

$$-\frac{s^T g}{\| s \| \| g \|} \geq \frac{\| g \|^2}{\overline{B}} \frac{\underline{B}}{m \| g \|^2} = \frac{1}{m} \frac{\underline{B}}{\overline{B}}$$

Tvrzení věty 61 tvoří základ jednoduchého algoritmu.

Algoritmus 7 Data $0 < \omega < 1$, $0 < \overline{\varepsilon} < 1$, $m \geq n$.

Krok 1 Položíme-li $s = 0$, $r = -g$, $\sigma = \| r \|^2$, $p = r$, $\delta = \| r \|^2$, $k = 1$

Krok 2 Položíme $\rho = \sigma$, vypočteme vektor $q = Bp$ a čísla $\tau = p^T q$, $\gamma = \tau / \delta$ a pokud $k = 1$ položíme $\overline{\gamma} = \gamma$. Jestliže $k = 1$ a $\gamma \leq 0$ položíme $s = r$ a ukončíme výpočet. Jestliže $k > 1$ a $\gamma \leq \overline{\varepsilon}\overline{\gamma}$ ukončíme výpočet.

Krok 3 Položíme $\alpha = \rho / \tau$. Položíme $s := s + \alpha p$, $r := r - \alpha q$ a $\sigma = \| r \|^2$. Jestliže $\sigma \leq \omega^2 \| g \|^2$ nebo $k \geq m$, ukončíme výpočet.

Krok 4 Položíme $\beta = \sigma / \rho$, $p = r + \beta p$, $\delta := \sigma + \beta^2 \delta$, $k := k + 1$ a přejdeme na krok 2.

(obvykle volíme $m = n + 3$).

Použití metody sdružených gradientů je velmi výhodné, neboť tato metoda, aplikovaná na kvadratickou funkci $Q(s)$ s pozitivně definitní maticí B, dává spádové směry bez ohledu na to, jak přesně se řeší soustava rovnic $Bs + g$ (obecná věta 6 vyžaduje, aby platilo $\| Bs + g \| \leq \overline{\omega} \| g \|$ a $\overline{\omega} \leq \underline{B} / \overline{B}$). I když konvergenční teorie, kterou jsme se dosud zabývali, není aplikovatelná na metody s proměnnou metrikou realizované jako metody spádových směrů, 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 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	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

4.5. 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, $1 \le k \le m$). Hessova matice $G(x)$ má stejnou strukturu jako matice $J^T(x)J(x)$. Struktura řídké úlohy je tedy plně určena strukturou Jacobiovy matice.

Definice 26 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

$$(ord\ J)_k = 1 + \sum_{i=1}^{k-1} n_k, \qquad 1 \le k \le m+1$$

V dalším výkladu budeme používat spakované gradienty $\hat{g}_k(x) \in R^{n_k}$, které obsahují pouze nenulové prvky gradientů $g_k(x) \in R^n$, $1 \le k \le m$, a spakované Hessovy matice $\hat{G}_k(x) \in R^{n_k \times n_k}$, které obsahují pouze nenulové prvky Hessových matic $G_k(x) \in R^{n \times n}$, $1 \le k \le m$. 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 spakovaných 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 $\hat{e}_j$, $1 \le j \le n_k$, jsou sloupce jednotkové matice řádu n_k. K určení prvků spakovaných Hessových matic je tedy zapotřebí

$$\sum_{k=1}^{m} n_k^2 = mO(1) = O(n)$$

operací.

Poznámka 49 Řídká Jacobiova matice J a numerické hodnoty spakovaných Hessových matic $\hat{G}_k(x)$, $1 \leq k \leq m$, jednoznačně určují gradient g a řídkou Hessovu matici G. Známe-li řídkou reprezentaci Jacobiovy matice (definice 26) a numerické hodnoty spakovaných Hessových matic, můžeme snadno určit řídkou reprezentaci Hessovy matice (definice 25). Spakované 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 3.3) nebo realizaci ve formě nepřesné metody s lokálně omezeným krokem (oddíl 3.4).

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.

4.6. Metody s proměnnou metrikou pro separovatelné úlohy

Metody s proměnnou metrikou pro separovatelné úlohy používají místo spakovaných Hessových matic $\hat{G}_k(x)$, $1 \leq k \leq m$, jejich aproximace $\hat{B}_k$, $1 \leq k \leq m$, které se aktualizují pomocí metod s proměnnou metrikou.

$$\hat{B}_k^+ = \frac{1}{\hat{\gamma}_k}\left(\hat{B}_k + \frac{\hat{\gamma}_k}{\hat{\rho}_k}\frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T + \frac{\hat{\beta}_k}{\hat{c}_k}\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)\left(\frac{\hat{c}_k}{\hat{b}_k}\hat{y}_k - \hat{B}_k\hat{d}_k\right)^T\right)$$

kde $\hat{y}_k = \hat{g}_k^+ - \hat{g}_k$ a $\hat{d}_k \in R^{n_k}$ je vektor, který má stejnou dimenzi jako vektor $\hat{y}_k$ a který obsahuje prvky vektoru d s indexy odpovídajícími prvkům vektoru $\hat{y}_k$ (vše pro $1 \leq k \leq m$). Přitom $\hat{b}_k = \hat{y}_k^T\hat{d}_k$, $\hat{c}_k = \hat{d}_k^T\hat{B}_k\hat{d}_k$, $1 \leq k \leq m$, a $\hat{\gamma}_k$, $\hat{\rho}_k$, $\hat{\beta}_k$, $1 \leq k \leq m$ 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 spakované matice $\hat{B}_k$, $1 \leq k \leq m$, se aktualizují pomocí vektorů y_k, d_k, $1 \leq k \leq m$, je účelné aby platilo $\hat{B}_k \rightarrow \hat{G}_k$, $1 \leq k \leq m$, takže se obyčejně pokládá $\gamma_k = 1$, $\rho_k = 1$, $1 \leq k \leq m$ (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 $\hat{y}_k^T\hat{d}_k > 0$, $1 \leq k \leq n$, takže matice $\hat{B}_k^+$, $1 \leq k \leq m$, nemusí být pozitivně definitní.

Popíšeme nyní efektivní realizaci metod s proměnnou metrikou pro separovatelné úlohy. Tato realizace je metodou spádových směrů (definice 17) a aktualizace se provádí podle vzorců

$$\begin{aligned}
\hat{B}_k^+ &= \hat{B}_k + \frac{1}{\hat{b}_k}\hat{y}_k\hat{y}_k^T - \frac{1}{\hat{c}_k}\hat{B}_k\hat{d}_k\left(\hat{B}_k\hat{d}_k\right)^T, & \hat{y}_k^T\hat{d}_k > 0 \\
\hat{B}_k^+ &= \hat{B}_k, & \hat{y}_k^T\hat{d}_k \leq 0
\end{aligned}$$

kde $1 \le k \le m$ (metoda BFGS). Tyto vzorce zaručují pozitivní definitnost matic $\hat{B}_k^+$, $1 \le k \le m$. Je-li matice $\hat{B}_k$ pozitivně definitní a platí-li $\hat{y}_k^T \hat{d}_k \le 0$, 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 25.

Známe-li aproximace $\hat{B}_k$, $1 \le k \le m$ spakovaných Hessových matic $\hat{G}_k$, $1 \le k \le m$, můžeme podle poznámky 49 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 $\hat{B}_k$, $1 \le k \le m$. 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, separovatelná 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

4.7 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 $f_k : R^n \to R$, $1 \le k \le m$, 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 3.9). V praxi se používají tři hybridní metody pro řídký součet čtverců.

1) Kombinace GN metody s Marwilovou metodou. V prvním iteračním kroku pokládáme $B = J^T J$. Po skončení každého iteračního kroku pokládáme

$$
\begin{aligned}
B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F \\
B_+ &= \mathcal{P}_S \mathcal{P}_{QS} B \quad , \quad F - F_+ \le \underline{\vartheta} F
\end{aligned}
$$

(viz (BM) v oddílu 4.4), kde $J_+ = J(x_+)$. Globální konvergence této metody plyne z věty 43 a věty 58. Superlineární konvergence této metody plyne z věty 43 a věty 60.

2) Kombinace GN metody s diferenční verzí Newtonovy metody. V prvním iteračním kroku pokládáme $B = J^T J$. Po skončení každého iteračního kroku pokládáme

$$
\begin{aligned}
B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F \\
B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ G_k^+ \quad , \quad F - F_+ \le \underline{\vartheta} F
\end{aligned}
$$

kde $J_+ = J(x_+)$ a $f_k^+ = f_k(x_+)$, $G_k^+ = G_k(x_+)$, $1 \le k \le m$ ($G_k(x_+)$ je diferenční aproximace Hessovy matice funkce $f_k(x_+)$). Globální a superlineární konvergence této metody plyne z věty 41 a věty 43.

3) Kombinace GN metody s metodou hodnosti 1. V prvním iteračním kroku pokládáme $B = J^T J$ a $B_k = I_k$, $1 \le k \le m$ (B_k je aproximace Hessovy matice G_k a I_k se od jednotkové matice liší pouze tím, že $(I_k)_{ii} = 0$, pokud $(G_k)_{ii} = 0$). Po skončení každého iteračního kroku pokládáme

$$
\begin{aligned}
B_k^+ &= B_k + \frac{w_k w_k^T}{d_k^T w_k} \quad , \quad |d_k^T w_k| > \underline{\delta} \\
B_k^+ &= B_k \quad , \quad |d_k^T w_k| \le \underline{\delta}
\end{aligned}
$$

pro $1 \le k \le m$, a

$$
\begin{aligned}
B_+ &= J_+^T J_+ \quad , \quad F - F_+ > \underline{\vartheta} F \\
B_+ &= J_+^T J_+ + \sum_{k=1}^{m} f_k^+ B_k^+ \quad , \quad F - F_+ \le \underline{\vartheta} F
\end{aligned}
$$

Přitom $w_k = y_k - B_k d_k$ a $y_k = g_k(x_+) - g_k(x)$, $d_k = x_+ - x$, $1 \le k \le m$. 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, $1 \le k \le m$ (ve skutečnosti se pracuje se spakovanými maticemi $\hat{B}_k$, $1 \le k \le m$).

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.

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 27 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 oddíle 3.4. 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 symetrického Lanczosova procesu.

Definice 28 Nechť $B \in R^{n \times n}$ je symetrická pozitivně definitní (SPD) matice a $g \in R^n$. Pak iterační proces používající rekurentní vztahy

$$q_0 = 0, \quad \beta_1 q_1 = g$$

a

$$\alpha_i = q_i^T B q_i \tag{SL}$$

$$\beta_{i+1} q_{i+1} = B q_i - \alpha_i q_i - \beta_i q_{i-1}$$

pro $1 \le i \le n$, kde koeficienty β_i, $1 \le i \le n$ se volí tak, aby vektory q_i, $1 \le i \le n$ měly jednotkovou normu, nazveme symetrickým Lanczosovým procesem určeným maticí $B \in R^{n \times n}$ a vektorem $g \in R^n$.

Poznámka 50 Nechť $\beta_i \neq 0$, $1 \le i \le k$, pro nějaký index $1 \le k \le n$. Pak podle (SL) platí $g = Q_k(\beta_1 e_1)$ a

$$B Q_k = Q_k T_k + \beta_{k+1} q_{k+1} e_k^T \tag{$\overline{\text{SL}}$}$$

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

$$T_k = \begin{bmatrix} \alpha_1, & \beta_2, & \ldots, & 0, & 0 \\ \beta_2, & \alpha_2, & \ldots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \ldots, & \alpha_{k-1}, & \beta_k \\ 0, & 0, & \ldots, & \beta_k, & \alpha_k \end{bmatrix}$$

(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 SPD maticí $B \in R^{n \times n}$ a vektorem $g \in R^n$. Nechť $\beta_i \neq 0$, $1 \le i \le k$ pro nějaký index $1 \le k \le n$. Pak vektory q_i, $1 \le i \le k$, tvoří ortonormální bázi v Krylovově podprostoru $\mathcal{K}_k = \text{span}\,(g, Bg, \ldots, B^{k-1} g)$.

Důkaz (indukcí). Pro $k = 1$ je tvrzení zřejmé, neboť $q_1 = g / \parallel g \parallel$. Předpokládejme, že tvrzení platí pro nějaký index $1 \le k < n$ a že $\beta_{k+1} \neq 0$. Podle indukčního předpokladu platí $Q_k^T Q_k = I$, takže $Q_k^T B Q_k = T_k + \beta_{k+1} Q_k^T q_{k+1} e_k^T$. Matice $Q_k^T B Q_k$ je symetrická stejně jako matice T_k, takže nutně $Q_{k-1}^T q_{k+1} = 0$. Dále podle (SL) platí $\beta_{k+1} q_k^T q_{k+1} = q_k^T B q_k - \alpha_k = \alpha_k - \alpha_k = 0$. Vektor q_{k+1} je tedy ortogonální k vektorům q_i, $1 \le i \le k$, a má jednotkovou normu. Podle (SL) leží vektory q_i, $1 \le i \le k+1$ v Krylovově podprostoru $\mathcal{K}_{k+1}$ a jelikož jsou vzájemně ortogonální a mají jednotkovou normu, tvoří tam ortonormální bázi.

Poznámka 51 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 52 Symetrický Lanczosův proces můžeme použít k řešení soustavy rovnic $Bs + g = 0$. Pokládáme $s_1 = 0$ a vektory s_{i+1}, $1 \le i \le k$, hledáme tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \left(\frac{1}{2} s^T B s + g^T s \right)$$

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

$$z_i = \arg \min_{z \in R^i} \left(\frac{1}{2} z^T T_i z + \beta_1 e_1^T z \right)$$

(plyne to ze vztahů $g = Q_i(\beta_1 e_1)$ a $Q_i^T Q_i = I$). Pokud $\beta_{k+1} = 0$, je vektor $s_{k+1} \in \mathcal{K}_k$ řešením soustavy rovnic $Bs + g = 0$. Podle ($\overline{\text{SL}}$) totiž platí $B Q_k = Q_k T_k$ a jelikož matice T_k je regulární, lze položit $z_k = -T_k^{-1}(\beta_1 e_1)$, což dává $B_k s_{k+1} = B Q_k z_k = -Q_k T_k T_k^{-1}(\beta_1 e_1) = -Q_k(\beta_1 e_1) = -g$.

Věta 63 Vektory s_{i+1}, $1 \le i \le k$ definované v poznámce 52 jsou shodné s vektory $\overline{s}_{i+1}$, $1 \le i \le k$, generovanými metodou CG:

$$\overline{s}_1 = 0, \quad \overline{g}_1 = g, \quad \overline{p}_1 = -\overline{g}_1$$

a

$$\overline{q}_i = B\overline{p}_i \qquad \overline{\alpha}_i = \| \overline{g}_i \|^2 / \overline{p}_i^T \overline{q}_i$$

$$\overline{s}_{i+1} = \overline{s}_i + \overline{\alpha}_i \overline{p}_i$$

$$\overline{g}_{i+1} = \overline{g}_i + \overline{\alpha}_i \overline{q}_i, \qquad \overline{\beta}_i = \| \overline{g}_{i+1} \|^2 / \| \overline{g}_i \|^2$$

$$\overline{p}_{i+1} = -\overline{g}_{i+1} + \overline{\beta}_i \overline{p}_i$$

pro $1 \le i \le k$. Navíc platí $\alpha_1 = 1/\overline{\alpha}_1$ a

$$\alpha_{i+1} = \frac{\overline{\beta}_i}{\overline{\alpha}_i} + \frac{1}{\overline{\alpha}_{i+1}}$$

$$\beta_{i+1} = \frac{\sqrt{\overline{\beta}_i}}{\overline{\alpha}_i}$$

$$q_i = (-1)^{i+1} \overline{g}_i / \| \overline{g}_i \|$$

pro $1 \le i \le k$.

Důkaz Z důkazu Věty 15 plyne, že vektory $\overline{s}_{i+1}$, $1 \le i \le k$, určené metodou CG, leží v Krylovových podprostorech $\mathcal{K}_i$, $1 \le i \le k$, 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}, $1 \le i \le k$, v poznámce 52. Jelikož vektory $\overline{g}_i$, $1 \le i \le k$ jsou vzájemně ortogonální a leží v Krylovových podprostorech $\mathcal{K}_i$, $1 \le i \le k$, musí být kolineární s vektory q_i, $1 \le i \le k$, neboli

$$\overline{G}_k = Q_k D_k$$

kde $\overline{G}_k = [\overline{g}_1, \ldots, \overline{g}_k]$ a $D_k = \text{diag}\ (\varepsilon_1 \| \overline{g}_1 \|, \ldots, \varepsilon_k \| \overline{g}_k \|)$ (čísla ε_i, $1 \le i \le k$, mohou nabývat hodnot ± 1). Položme $\overline{P}_k = [\overline{p}_1, \ldots, \overline{p}_k]$. Pak z rekurentních vztahů metody CG plyne

$$\overline{G}_k = \overline{P}_k \overline{B}_k$$

kde

$$\overline{B}_k = \begin{bmatrix} -1, & \overline{\beta}_1, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}$$

je horní bidiagonální matice. Z důkazu věty 15 plyne, že matice $\overline{P}_k^T B \overline{P}_k$ je diagonální. Použijeme-li rekurentní vztahy metody CG, dostaneme $P_k^T B P_k = \text{diag}(\| \overline{g}_1 \|^2 / \overline{\alpha}_1, \ldots, \| \overline{g}_k \|^2 / \overline{\alpha}_k) = D_k \text{diag}(1/\overline{\alpha}_1, \ldots, 1/\overline{\alpha}_k) D_k$, takže

$$\begin{aligned} T_k &= Q_k^T B Q_k = D_k^{-1} \overline{G}_k^T B \overline{G}_k D_k^{-1} = D_k^{-1} \overline{B}_k^T \overline{P}_k^T B \overline{P}_k \overline{B}_k D_k^{-1} = \\ &= D_k^{-1} \overline{B}_k^T D_k \text{diag}\ (1/\overline{\alpha}_1, \ldots, 1/\overline{\alpha}_k) D_k \overline{B}_k D_k^{-1} \end{aligned}$$

Ale

$$D_k B_k D_k^{-1} = \begin{bmatrix} -1, & \overline{\beta}_1 \frac{\varepsilon_1 \| \overline{g}_1 \|}{\varepsilon_2 \| \overline{g}_2 \|}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix} = \begin{bmatrix} -1, & \varepsilon_1 \varepsilon_2 \sqrt{\overline{\beta}_1}, & \ldots, & 0 \\ 0, & -1, & \ldots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \ldots, & -1 \end{bmatrix}$$

Dosadíme-li tento vztah do vyjádření pro matici T_k, můžeme psát

$$T_k = \begin{bmatrix} \frac{1}{\overline{\alpha}_1}, & \frac{-\varepsilon_1\varepsilon_2\sqrt{\overline{\beta}_1}}{\overline{\alpha}_1}, & \dots, & 0 \\ \frac{-\varepsilon_1\varepsilon_2\sqrt{\overline{\beta}_1}}{\overline{\alpha}_1}, & \frac{\overline{\beta}_1}{\overline{\alpha}_1} + \frac{1}{\overline{\alpha}_2}, & \dots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \dots, & \frac{\overline{\beta}_{k-1}}{\overline{\alpha}_{k-1}} + \frac{1}{\overline{\alpha}_k} \end{bmatrix}$$

což porovnáním se $(\overline{\mathrm{SL}})$ dává $\alpha_1 = 1/\overline{\alpha}_1$ a

$$\begin{aligned} \alpha_{i+1} &= \frac{\overline{\beta}_i}{\overline{\alpha}_i} + \frac{1}{\overline{\alpha}_{i+1}} \\ \beta_{i+1} &= -\frac{\varepsilon_i\varepsilon_{i+1}\sqrt{\overline{\beta}_i}}{\overline{\alpha}_i} = \frac{\sqrt{\overline{\beta}_i}}{\overline{\alpha}_i} \end{aligned}$$

pro $1 \le i \le k$ (jelikož $\overline{\alpha}_i > 0$, $\overline{\beta}_i \ge 0$ $\beta_{i+1} \ge 0$, musí platit $\varepsilon_i\varepsilon_{i+1} = -1$). Protože podle (SL) platí $\beta_1 q_1 = g = \overline{g}_1$ a $\beta_1 \ge 0$, dostaneme $\varepsilon_1 = 1$, což spolu s $\varepsilon_1\varepsilon_{i+1} = -1$, $1 \le i \le k$, dává $\varepsilon_i = (-1)^{i+1}$, takže

$$q_i = (-1)^{i+1}\overline{g}_i / \parallel \overline{g}_i \parallel$$

Definice 29 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

$$\delta_{i+1}u_{i+1} = Jq_i - \gamma_i u_i \tag{BL}$$

$$\gamma_{i+1}q_{i+1} = J^T u_{i+1} - \delta_{i+1}q_i$$

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 53 Nechť $\gamma_i \ne 0$, $\delta_i \ne 0$, $1 \le i \le k$ pro nějaký index $1 \le k \le n$. Pak podle (BL) platí $f = U_{k+1}(\delta_1 e_1)$ a

$$JQ_k = U_{k+1}B_k$$

$$J^T U_{k+1} = Q_k B_k^T + \gamma_{k+1}q_{k+1}e_{k+1}^T \tag{$\overline{\mathrm{BL}}$}$$

kde $Q_k = [q_1, q_2, \dots, q_k]$, $U_{k+1} = [u_1, u_2, \dots, u_k, u_{k+1}]$, $e_1^T = [1, 0, \dots, 0, 0]$, $e_{k+1}^T = [0, 0, \dots, 0, 1]$ a

$$B_k = \begin{bmatrix} \gamma_1, & 0, & \dots, & 0 \\ \delta_2, & \gamma_2, & \dots, & 0 \\ -- & -- & -- & -- \\ 0, & 0, & \dots, & \gamma_k \\ 0, & 0, & \dots, & \delta_{k+1} \end{bmatrix}$$

(matice $B_k \in R^{(k+1)\times k}$ je bidiagonální).

Věta 64 Uvažujme bidiagonalizační Lanczosův proces určený maticí $J \in R^{m\times n}$ a vektorem $f \in R^m$. Nechť $\gamma_i \ne 0$, $\delta_i \ne 0$, $1 \le i \le k$, pro nějaký index $1 \le k \le n$. Pak vektory q_i, $1 \le i \le k$, tvoří ortonormální bázi v Krylovově podprostoru $\mathcal{K}_i = \mathrm{span}\,(g, Bg, \dots, B^{k-1}g)$, kde $B = J^T J$ a $g = J^T f$, a vektory u_i, $1 \le i \le k$, jsou vzájemně ortogonální a mají jednotkovou normu.

Důkaz (indukcí). Pro $k = 1$ je tvrzení zřejmé, neboť $q_1 = J^T f / \| J^T f \|$ a $u_1 = f / \| f \|$. Předpokládejme, že tvrzení platí pro nějaký index $1 \le k < n$ a že $\gamma_{k+1} \ne 0$, $\delta_{k+1} \ne 0$. 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

$$T_k = B_k^T B_k = \begin{bmatrix} \gamma_1^2 + \delta_2^2, & \gamma_2\delta_2, & \dots, & 0, & 0 \\ \gamma_2\delta_2, & \gamma_2^2 + \delta_3^2, & \dots, & 0, & 0 \\ -- & -- & -- & -- & -- \\ 0, & 0, & \dots, & \gamma_{k-1}^2 + \delta_k^2, & \gamma_k\delta_k \\ 0, & 0, & \dots, & \gamma_k\delta_k, & \gamma_k^2 + \delta_{k+1}^2 \end{bmatrix}$$

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 $\mathcal{K}_i = \text{span}\,(g, Bg, \dots, B^{k-1} g)$, 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 ($\overline{\text{BL}}$) 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, $1 \le i \le k+1$, jsou vzájemně ortogonální a mají jednotkovou normu).

Poznámka 54 Z důkazu věty 64 plyne, že symetrický Lanczosův proces určený SPD maticí $B \in R^{n \times n}$ a vektorem $g \in R^n$ je ekvivalentní bidiagonalizačnímu Lanczosovu 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, $1 \le i \le k$, a platí $\alpha_i = \gamma_i^2 + \delta_{i+1}^2$, $\beta_i = \gamma_i \delta_i$, $1 \le i \le k$, kde k je index takový, že $\alpha_i \ne 0$, $\beta_i \ne 0$, $\gamma_i \ne 0$, $\delta_i \ne 0$, $1 \le i \le k$.

Poznámka 55 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}, $1 \le i \le k$, hledáme tak, aby platilo

$$s_{i+1} = \arg \min_{s \in \mathcal{K}_i} \| Js + f \|$$

Jelikož $s \in \mathcal{K}_i$ právě tehdy, jestliže $s = Q_i z$, kde $z \in R^i$, můžeme psát $s_{i+1} = Q_i z_i$, kde

$$z_i = \arg \min_{z \in R^i} \| B_i z + \delta_1 e_1 \|$$

(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 52 a poznámky 54).

Poznámka 56 Vektory s_{i+1}, $1 \le i \le k$, definované v poznámce 55 jsou shodné s vektory s_{i+1}, $1 \le i \le k$, generovanými metodou CGNE určenou maticí $J \in R^{m \times n}$ a vektorem $f \in R^m$ (plyne to z věty 63 a poznámky 54).

Výhodou bidiagonalizačního Lanczosova procesu je skutečnost, že vektory s_{i+1}, $1 \le i \le k$, 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

$$R_i = \begin{bmatrix} \rho_1, & \sigma_2, & \dots, & 0 \\ 0, & \rho_2, & \dots, & 0 \\ 0, & 0, & \dots, & \rho_i \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_i \end{bmatrix}$$

Přitom $P_i \in R^{i \times i}$, $1 \le i \le k$, jsou ortogonální matice (součiny Givensových elementárních rotací) a $R_i \in R^{i \times i}$, $1 \le i \le k$, 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

$$\begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\rho}_1, & 0 \\ \delta_2, & \gamma_2 \\ 0, & \delta_3 \end{bmatrix} = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \overline{\rho}_2 \\ 0, & \delta_3 \end{bmatrix}, \qquad \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix} \begin{bmatrix} \overline{\eta}_1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} \eta_1 \\ \overline{\eta}_2 \\ 0 \end{bmatrix}$$

a můžeme položit

$$P_2 = \begin{bmatrix} 1, & 0 \\ 0, & \overline{P}_2 \end{bmatrix} \begin{bmatrix} P_1, & 0 \\ 0, & 1 \end{bmatrix}, \qquad \overline{P}_2 = \frac{1}{\sqrt{\overline{\rho}_2^2 + \delta_3^2}} \begin{bmatrix} \overline{\rho}_2, & \delta_3 \\ -\delta_3, & \overline{\rho}_2 \end{bmatrix}$$

Pokračujeme-li takto dále, dostaneme rekurentní vztahy

$$\overline{\rho}_1 = \gamma_1, \qquad \overline{\eta}_1 = \delta_1$$

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 $1 \le i \le k$. Nyní odvodíme rekurentní vztahy pro vektory s_{i+1}, $1 \le i \le k$. 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

$$z_i = \arg \min_{z \in R^i} \| R_i z + h_i \|$$

Jelikož matice $R_i \in R^{i \times i}$ je regulární, musí platit $R_i z_i + h_i = 0$. Vzhledem k jednoduché struktuře matic R_i, $1 \le i \le k$, můžeme vektory z_i, $1 \le i \le k$, a tudíž i vektory s_{i+1}, $1 \le i \le k$, určovat rekurentně. Ukážeme nejprve dva kroky tohoto procesu. Na začátku prvního kroku platí

$$R_1 = [\rho_1], \qquad h_1 = [\eta_1]$$

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

$$s_1 = 0, \qquad p_1 = -q_1$$

Na začátku druhého kroku platí

$$R_2 = \begin{bmatrix} \rho_1, & \sigma_2 \\ 0, & \rho_2 \end{bmatrix}, \qquad h_2 = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix}$$

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

$$s_1 = 0, \qquad p_1 = -q_1$$

a

$$\begin{aligned} s_{i+1} &= s_i + \frac{\eta_i}{\rho_i} p_i \\ p_{i+1} &= -q_{i+1} + \frac{\sigma_{i+1}}{\rho_i} p_i \end{aligned}$$

pro $1 \le i \le k$.

Definice 30 Nechť $J \in R^{m \times n}$ je matice s lineárně nezávislými sloupci a $f \in R^m$. Pak iterační proces používající rekurentní vztahy

$$s_1 = 0, \qquad \delta_1 u_1 = f, \qquad \gamma_1 q_1 = J^T u_1, \qquad p_1 = q_1$$

a

$$\delta_{i+1} u_{i+1} = J q_i - \gamma_i u_i$$

$$\gamma_{i+1} q_{i+1} = J^T u_{i+1} - \delta_{i+1} q_i$$

$$\rho_i = \sqrt{\overline{\rho}_i^2 + \delta_{i+1}^2}, \qquad \lambda_i = \frac{\overline{\rho}_i}{\rho_i}, \qquad \tau_i = \frac{\delta_{i+1}}{\rho_i}$$

$$\overline{\rho}_{i+1} = \lambda_i \gamma_{i+1}, \qquad \sigma_{i+1} = \tau_i \gamma_{i+1}$$

$$\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i$$

$$s_{i+1} = s_i + \frac{\eta_i}{\rho_i} p_i$$

$$p_{i+1} = -q_{i+1} + \frac{\sigma_{i+1}}{\rho_i} p_i$$

pro $1 \le i \le n$, 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 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 56 generují obě metody stejné vektory s_{i+1}, $1 \le i \le k$, kde $k \le n$ a $J^T J s_{k+1} + J^T f = 0$. Ukážeme ještě, jak je možné odhadovat přesnost řešení.

Věta 65 Nechť $s_{i+1} \in R_n$, γ_{i+1}, δ_{i+1}, $\rho_i > 0$, η_i, $1 \le i \le k$, jsou veličiny generované metodou LSQR. Pak pro $1 \le i \le k$ platí

$$\| J^T (J s_{i+1} + f) \| = \gamma_{i+1} \delta_{i+1} \frac{|\eta_i|}{\rho_i}$$

Důkaz Nechť $\gamma_{i+1} \ne 0$, $\delta_{i+1} \ne 0$. Pak použitím ($\overline{\text{BL}}$) a poznámky 55 dostaneme

$$\begin{aligned} J^T (J s_{i+1} + f) &= J^T (J Q_i z_i + f) = J^T U_{i+1} (B_i z_i + \delta_1 e_1) = \\ &= (Q_i B_i^T + \gamma_{i+1} q_{i+1} e_{i+1}^T)(B_i z_i + \delta_1 e_1) = \gamma_{i+1} q_{i+1} e_{i+1}^T B_i z_i = \\ &= \gamma_{i+1} \delta_{i+1} q_{i+1} e_i^T z_i \end{aligned}$$

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

$$\| J^T (J s_{i+1} + f) \| = \gamma_{i+1} \delta_{i+1} |q_i^T s_{i+1}|$$

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 $\| J^T (J s_{i+1} + f) \| = 0$ (poznámka 55).

Větu 65 můžeme využít k zastavení iteračního procesu (není třeba počítat reziduum $\| J^T (J s_{i+1} + f) \|$).

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.

5. Metody pro řešení soustav nelineárních rovnic

Nechť $f : R^n \to R^n$ je spojitě diferencovatelná funkce. 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.

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 případ soustavy rovnic, neboť Jacobiova matice funkce $g : R^n \to R^n$ je Hessovou maticí funkce $F : R^n \to R$ a je tedy symetrická (za standardních podmínek kladených na funkci $F : R^n \to R$).
- Řešení soustavy rovnic můžeme převést na minimalizaci funkce $F : R^n \to R$ definované vztahem $F(x) = (1/2) \parallel f(x) \parallel^2$ (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 $F(x) = (1/2) \parallel f(x) \parallel^2$ a řešením soustavy rovnic $f(x) = 0$ udává tato věta.

Věta 66 Nechť $f : R^n \to R^n \in C^1$ a nechť bod $x^* \in R^n$ je lokálním minimem funkce $F(x) = (1/2) \parallel f(x) \parallel^2$, přičemž Jacobiova matice $J(x^*)$ funkce $f : R^n \to R^n$ 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 $F : R^n \to R$.

V souvislosti s řešením soustavy rovnic $f(x) = 0$, budeme používat tyto standardní podmínky kladené na Jacobiovu matici funkce $f : R^n \to R^n$.

1. Stejnoměrná omezenost:

$$\parallel J(x) \parallel \leq \overline{J} \tag{J3}$$

$\forall x \in R^n$

2. Stejnoměrná regularita:

$$\| J^{-1}(x) \| \leq 1/\underline{J} \tag{J4}$$

$\forall x \in R^n$

3. Lipschitzovskost:

$$\| J(x+d) - J(x) \| \leq \overline{L} \| d \| \tag{J5}$$

$\forall x \in R^n$, $\forall d \in R^n$.

Tyto 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 31 Řekneme, že základní metoda pro řešení soustav nelineárních rovnic je globálně konvergentní, jestliže pro libovolný počáteční vektor $x_1 \in R^n$ platí

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

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

$$(J^T(x_i) J(x_i))^{-1} J^T(x_i) = J^{-1}(x_i)$$

Podmínka (J4) má význam nejen pro vyloučení lokálních extrémů funkce $F(x) = (1/2) \| f(x) \|^2$ (věta 65) ale i pro konstrukci globálně konvergentních metod. Je-li podmínka (J4) splněna, je matice $B_i = J^T(x_i) J(x_i)$ pozitivně definitní, takže Newtonovu metodu pro řešení soustav nelineárních rovnic můžeme realizovat (na rozdíl od Newtonovy metody pro nepodmíněnou minimalizaci popsané v oddílu 3.7) jako metodu spádových směrů.

5.1. Metody spádových směrů

Metody spádových směrů pro řešení soustav nelineárních rovnic používají stejnoměrně spádové směrové vektory (definice 15). Přitom stejnosměrná spádovost se vztahuje k funkci $F : R^n \to R$ ($\triangleq (1/2) \| f \|^2$), neboli předpokládáme existenci konstanty $0 < \varepsilon_0 \leq 1$ takové, že platí

$$-s_i^T g_i \geq \varepsilon_0 \| s_i \| \| g_i \| \tag{S1}$$

$\forall i \in N$, kde $g_i = J_i^T f_i$, $i \in N$, jsou gradienty funkce $F : R^n \to R$ v bodech x_i, $i \in N$. Směrové vektory $s_i \in R^n$, $i \in N$, se obvykle určují přibližným řešením soustav lineárních rovnic $A_i s_i + f_i = 0$, $i \in N$, kde matice A_i, $i \in N$, jsou aproximace Jacobiových matic J_i, $i \in N$. Ukážeme nejprve, jak přesně je třeba řešit soustavy rovnic $A_i s_i + f_i = 0$, $i \in N$, a jak přesně je třeba aproximovat matice J_i, $i \in N$.

Věta 67 Nechť jsou splněny předpoklady (J3)-(J5). Nechť $\| A_i s_i + f_i \| \le \overline{\omega} \| f_i \|$, $\| (A_i - J_i) s_i \| \le \overline{\lambda} \| f_i \|$, $\forall i \in N$, kde $0 \le \overline{\lambda} + \overline{\omega} < 1$. Pak směrové vektory $s_i \in R^n$, $i \in N$, jsou stejnoměrně spádové a platí (S1) s

$$\varepsilon_0 = \frac{\underline{J}}{\overline{J}} \frac{1 - \overline{\lambda} - \overline{\omega}}{1 + \overline{\lambda} + \overline{\omega}}$$

Důkaz Zřejmě platí

$$\begin{aligned} -g_i^T s_i &= -f_i^T J_i s_i = f_i^T ((A_i - J_i) s_i - (A_i s_i + f_i) + f_i) \ge \\ &\ge \| f_i \|^2 - \| f_i \| \| (A_i - J_i) s_i \| - \| f_i \| \| A_i s_i + f_i \| \ge \\ &\ge (1 - \overline{\lambda} - \overline{\omega}) \| f_i \|^2 \end{aligned}$$

$\forall i \in N$. Dále platí

$$J_i s_i = A_i s_i + f_i - (A_i - J_i) s_i - f_i$$

takže

$$\| s_i \| \le \| J_i^{-1} \| \| (A_i s_i + f_i - (A_i - J_i) s_i - f_i) \| \le \frac{1}{\underline{J}} (1 + \overline{\lambda} + \overline{\omega}) \| f_i \|$$

$\forall i \in N$ a

$$\| g_i \| = \| J_i^T f_i \| \le \overline{J} \| f_i \|$$

$\forall i \in N$. Použijeme-li všechny tyto nerovnosti, dostaneme

$$-\frac{s_i^T g_i}{\| s_i \| \| g_i \|} \ge \frac{\underline{J}}{\overline{J}} \frac{1 - \overline{\lambda} - \overline{\omega}}{1 + \overline{\lambda} + \overline{\omega}}$$

Poznámka 57 Předpoklad (J4) je velmi podstatný. Předně je třeba, aby matice J_i, $i \in N$, měly omezené číslo podmíněnosti (koeficient $\overline{J}/\underline{J}$). Dále je třeba, aby pro $0 \le \overline{\omega} < 1$ existovaly směrové vektory s_i, $i \in N$, takové, že $\| A_i s_i + f_i \| \le \overline{\omega} \| f_i \|$, $i \in N$ (to není možné například tehdy, jestliže $A_i = J_i$, a $J_i^T f_i = 0$, neboť v tomto případě platí $\| A_i s_i + f_i \|^2 = s_i^T A_i^T A_i s_i + f_i^T f_i \ge \| f_i \|^2$). Předpoklad (J5) je ve větě 67 nadbytečný, je však nutný pro další využití této věty.

Poznámka 58 Porovnáme-li větu 67 s větou 6, vidíme, že v případě soustav nelineárních rovnic nestačí dostatečně přesně řešit soustavy lineárních rovnic $A_i s_i + f_i = 0$, $i \in N$, ale je též třeba dostatečně přesně aproximovat Jacobiovy matice J_i, $i \in N$. Je to způsobeno tím, že ve vztazích pro gradienty funkce $F : R^n \to R$ vystupují Jacobiovy matice, které jsou různé od matic A_i, $i \in N$ (nepřesná aproximace Jacobiovy matice implikuje nepřesnost gradientu).

Poznámka 59 Porovnáme-li větu 67 s větou 6 vidíme, že v případě Newtonovy metody pro řešení soustav nelineárních rovnic mohou být soustavy lineárních rovnic $J_i s_i + f_i = 0$, $i \in N$, řešeny značně nepřesně. Vyžadujeme pouze, aby platilo $\overline{\omega} < 1$ a nikoliv $\overline{\omega} < \underline{G}/\overline{G}$ jako v případě Newtonovy metody pro nepodmíněnou optimalizaci. To je velká výhoda, neboť můžeme soustavy lineárních rovnic $J_i s_i + f_i = 0$, $i \in N$, řešit rozličnými iteračními metodami, které nemusí být tak dokonalé jako metoda sdružených gradientů (věta 61).

Jestliže $A_i \ne J_i$, $i \in N$, není možné určit gradienty $g_i = J_i^T f_i$, $i \in N$, funkce $F : R^n \to R$. Místo toho můžeme používat přibližné gradienty $\tilde{g}_i = A_i^T f_i$, $i \in N$. V tomto případě je výhodné používat

Armijův výběr délky kroku, kdy $\alpha_i > 0$ je první člen posloupnosti α_i^j, $j \in N$, takové, že $\alpha_i^1 = 1$ a $\underline{\beta}\alpha_i^j \le \alpha_i^{j+1} < \alpha_i^j$, $0 < \underline{\beta} \le 1/2$, pro který platí

$$F_{i+1} - F_i \le \tilde{\varepsilon}_1 \alpha_i s_i^T \tilde{g}_i \tag{$\overline{\mathrm{S2}}$}$$

Věta 68 Nechť jsou splněny předpoklady věty 67 a nechť platí $2\tilde{\varepsilon}_1 < (1-\overline{\lambda}-\overline{\omega})/(1-\overline{\omega})$. Pak existují délky kroku $\alpha_i > 0$, $i \in N$ vyhovující podmínce ($\overline{\mathrm{S2}}$) a použijeme-li Armijův výběr délky kroku s ($\overline{\mathrm{S2}}$), platí

$$\lim_{i\to\infty} \| f_i \| = 0$$

Důkaz (a) Nechť $0 < 2\varepsilon_1 < 1$. Jelikož směrové vektory $s_i \in R^n$, $i \in N$ jsou podle věty 67 stejnoměrně spádové, existují podle lemmatu 7 délky kroku $\alpha_i > 0$, $i \in N$, tak, že platí (S2). Z definice přibližného gradientu plyne

$$\tilde{g}_i^T s_i = f_i^T A_i s_i = f_i^T J_i s_i + f_i^T (A_i - J_i) s_i = g_i^T s_i + f_i^T (A_i - J_i) s_i$$

takže

$$\tilde{g}_i^T s_i \ge g_i^T s_i - \overline{\lambda} \| f_i \|^2$$

a použijeme-li nerovnost z důkazu věty 67, dostaneme

$$-(1-\overline{\lambda}-\overline{\omega}) \| f_i \|^2 \ge g_i^T s_i$$

což dohromady dává

$$\tilde{g}_i^T s_i \ge g_i^T s_i + \frac{\overline{\lambda}}{1-\overline{\lambda}-\overline{\omega}} g_i^T s_i = \frac{1-\overline{\omega}}{1-\overline{\lambda}-\overline{\omega}} g_i^T s_i$$

Dosadíme-li toto vyjádření do podmínky (S2), dostaneme

$$F_{i+1} - F_i \le \varepsilon_1 \frac{1-\overline{\lambda}-\overline{\omega}}{1-\overline{\omega}} \alpha_i s_i^T \tilde{g}_i \triangleq \tilde{\varepsilon}_1 \alpha_i s_i^T \tilde{g}_i$$

a jelikož $0 < 2\varepsilon_1 < 1$, platí $0 < 2\tilde{\varepsilon}_1 < (1-\overline{\omega}-\overline{\lambda})/(1-\overline{\omega}) \le 1$.
(b) Používáme-li Armijův výběr délky kroku, platí buď $\alpha_i = \alpha_i^1 = 1$ nebo $\alpha_i = \alpha_i^k = \beta\alpha_i^{k-1}$, $\underline{\beta} \le \beta < 1$, a $F(x_i + \alpha_i^{k-1} s_i) - F(x_i) \ge \tilde{\varepsilon}_1 \alpha_i^{k-1} s_i^T \tilde{g}_i$. V druhém případě ($\alpha_i \ne 1$) dostaneme

$$F(x_i + \frac{\alpha_i}{\beta} s_i) - F(x_i) \ge \frac{\tilde{\varepsilon}_1}{\beta} \alpha_i s_i^T \tilde{g}_i \ge \frac{\tilde{\varepsilon}_1}{\beta} \frac{1-\overline{\omega}}{1-\overline{\lambda}-\overline{\omega}} \alpha_i s_i^T g_i \triangleq \frac{\tilde{\varepsilon}_2}{\beta} \alpha_i s_i^T g_i$$

kde $0 < 2\tilde{\varepsilon}_2 < 1$, neboť $0 < 2\tilde{\varepsilon}_1 < (1-\overline{\lambda}-\overline{\omega})/(1-\overline{\omega})$. Označíme-li $d_i = (\alpha_i/\beta) s_i$ a použijeme-li větu o střední hodnotě, můžeme psát

$$\tilde{\varepsilon}_2 d_i^T g_i \le F(x_i + d_i) - F(x_i) \le d_i^T g(x_i + \mu d_i) \le d_i^T g_i + \|d_i\| \|g(x_i + \mu d_i) - g(x_i)\|$$

kde $0 \le \lambda \le 1$. Použijeme-li (J5) a větu o střední hodnotě, můžeme psát

$$\begin{aligned}
\|g(x_i + \lambda d_i) - g(x_i)\| &= \|J^T(x_i + \lambda d_i) f(x_i + \lambda d_i) - J^T(x_i) f(x_i)\| \le \\
&\le \|J^T(x_i + \lambda d_i)(f(x_i + \lambda d_i) - f(x_i))\| + \\
&\quad + \|(J^T(x_i + \lambda d_i) - J^T(x_i)) f(x_i)\| \le \\
&\le \overline{J} \|(f(x_i + \lambda d_i) - f(x_i)\| + \overline{L}\lambda \|d_i\| \|f_i\| = \\
&= \overline{J} \| \int_0^1 J(x_i + \tau\lambda d_i) \lambda d_i d\tau \| + \overline{L}\lambda \|d_i\| \|f_i\| \le \\
&\le \frac{1}{2}(\overline{J}^2 + 2\overline{LF}) \|d_i\| \le \frac{1}{2} \frac{\alpha_i}{\underline{\beta}} (\overline{J}^2 + 2\overline{LF}) \|s_i\|.
\end{aligned}$$

kde $\overline{F}$ je horní mez pro normu funkce $f(x)$ (například $\overline{F} > \| f_1 \|$). Platí tedy

$$\tilde{\varepsilon}_2 s_i^T g_i \leq s_i^T g_i + \frac{1}{2}\frac{\alpha_i}{\underline{\beta}}(\overline{J}^2 + 2\overline{LF})\|s_i\|^2$$

což po úpravě dává

$$\alpha_i \geq -\frac{2\underline{\beta}(1-\tilde{\varepsilon}_2)s_i^T g_i}{(\overline{J}^2 + 2\overline{LF}) \| s_i \|^2}$$

a použijeme-li odhady získané v důkazu věty 67, dostaneme

$$\alpha_i \geq \frac{2\underline{\beta}(1-\tilde{\varepsilon}_2)\underline{J}^2(1-\overline{\lambda}-\overline{\omega})}{(\overline{J}^2 + 2\overline{LF})(1+\overline{\lambda}+\overline{\omega})^2}$$

(c) Předpokládejme, že je splněna podmínka ($\overline{\mathrm{S2}}$). Zřejmě platí

$$\tilde{g}_i^T s_i = f_i^T A_i s_i = f_i^T (A_i s_i + f_i - f_i) \leq -(1-\overline{\omega}) \| f_i \|^2$$

a

$$g_i^T s_i = f_i^T J_i s_i = f_i^T (A_i s_i + f_i - (A_i - J_i)s_i - f_i) \geq -(1+\overline{\lambda}+\overline{\omega}) \| f_i \|^2$$

což dohromady dává

$$\tilde{g}_i^T s_i \leq \frac{1-\overline{\omega}}{1+\overline{\lambda}+\overline{\omega}} g_i^T s_i$$

a po dosazení do ($\overline{\mathrm{S2}}$) dostaneme

$$F_{i+1} - F_i \leq \tilde{\varepsilon}_1 \frac{1-\overline{\omega}}{1+\overline{\lambda}+\overline{\omega}} \alpha_i s_i^T g_i$$

Použijeme-li tuto nerovnost spolu s odhadem získaným v důkazu věty 67, dostaneme

$$\begin{aligned}
\| f_i \| (\| f_{i+1} \| - \| f_i \|) &\leq \frac{1}{2}(\| f_{i+1} \| + \| f_i \|)(\| f_{i+1} \| - \| f_i \|) = \\
&= F_{i+1} - F_i \leq \tilde{\varepsilon}_1 \frac{1-\overline{\omega}}{1+\overline{\lambda}+\overline{\omega}} \alpha_i s_i^T g_i \leq \\
&\leq -\tilde{\varepsilon}_1 (1-\overline{\omega}) \frac{1-\overline{\lambda}-\overline{\omega}}{1+\overline{\lambda}+\overline{\omega}} \alpha_i \| f_i \|^2
\end{aligned}$$

takže

$$\| f_{i+1} \| \leq \left(1 - \tilde{\varepsilon}_1(1-\overline{\omega})\frac{1-\overline{\lambda}-\overline{\omega}}{1+\overline{\lambda}+\overline{\omega}}\alpha_i\right) \| f_i \|$$

Platí tedy

$$\| f_{i+1} \| \leq q \| f_i \|$$

kde

$$0 < q = \left(1 - (1-\overline{\omega})\frac{1-\overline{\omega}-\overline{\lambda}}{1+\overline{\lambda}+\overline{\omega}} \min\left(\tilde{\varepsilon}_1, \frac{2\underline{\beta}\tilde{\varepsilon}_1(1-\tilde{\varepsilon}_2)\underline{J}^2(1-\overline{\lambda}-\overline{\omega})}{(\overline{J}^2 + 2\overline{LF})(1+\overline{\lambda}+\overline{\omega})^2}\right)\right) < 1$$

neboť $0 < 2\tilde{\varepsilon}_1 < 1$, $0 < 2\tilde{\varepsilon}_2 < 1$, $\underline{J}/\overline{J} \leq 1$ (takže $\underline{J}^2/(\overline{J}^2 + 2\overline{LF}) \leq 1$), $(1-\overline{\omega})(1-\overline{\lambda}-\overline{\omega})/(1+\overline{\lambda}+\overline{\omega}) \leq 1$, $(1-\overline{\lambda}-\overline{\omega})/(1+\overline{\lambda}+\overline{\omega})^2 \leq 1$.

Armijův výběr délky kroku můžeme realizovat také tak, že není nutné používat přibližné gradienty $\tilde{g}_i = A_i^T f_i$, $i \in N$. V tomto případě používáme podmínku

$$\| f_{i+1} \| \leq (1 - \tilde{\varepsilon}_1 \alpha_i) \| f_i \| \tag{$\overline{\text{S4}}$}$$

která je důsledkem podmínky ($\overline{\text{S2}}$) a $\alpha_i > 0$ je první člen posloupnosti α_i^j, $j \in N$, takové, že $\alpha_i^1 = 1$ a $\underline{\beta}\alpha_i^j \leq \alpha_i^{j+1} < \alpha_i^j$, $0 < \underline{\beta} \leq 1/2$, pro který platí ($\overline{\text{S4}}$) (číslo $\tilde{\varepsilon}_1$ je zde jiné než ve větě 68).

Věta 69 Nechť jsou splněny předpoklady věty 67 a nechť platí $2\tilde{\varepsilon}_1 < (1 - \overline{\lambda} - \overline{\omega})$. Pak existují délky kroku $\alpha_i > 0$, $i \in N$ vyhovující podmínce ($\overline{\text{S4}}$) a použijeme-li Armijův výběr délky kroku s ($\overline{\text{S4}}$), platí

$$\lim_{i \to \infty} \| f_i \| = 0$$

Důkaz (a) Nechť $0 < 2\varepsilon_1 < 1$. Jelikož směrové vektory $s_i \in R^n$, $i \in N$, jsou podle věty 67 stejnoměrně spádové, existují podle lemmatu 7 délky kroku $\alpha_i > 0$, $i \in N$, tak, že platí (S2). Použijeme-li (S2) spolu s odhadem získaným v důkazu věty 67, dostaneme

$$\begin{aligned} \| f_i \| (\| f_{i+1} \| - \| f_i \|) &\leq \frac{1}{2}(\| f_{i+1} \| + \| f_i \|)(\| f_{i+1} \| - \| f_i \|) = \\ &= F_{i+1} - F_i \leq \varepsilon_1 \alpha_i s_i^T g_i \leq -\varepsilon_1 (1 - \overline{\lambda} - \overline{\omega}) \alpha_i \| f_i \|^2 \end{aligned}$$

takže

$$\| f_{i+1} \| \leq \left(1 - \varepsilon_1 (1 - \overline{\lambda} - \overline{\omega}) \alpha_i\right) \| f_i \| \stackrel{\Delta}{=} (1 - \tilde{\varepsilon}_1 \alpha_i) \| f_i \|$$

a jelikož $0 < 2\varepsilon_1 < 1$, platí $0 < 2\tilde{\varepsilon}_1 < (1 - \overline{\lambda} - \overline{\omega}) \leq 1$.
(b) Používáme-li Armijův výběr délky kroku platí buď $\alpha_i = \alpha_i^1 = 1$ nebo $\alpha_i = \alpha_i^k = \beta \alpha_i^{k-1}$, $\underline{\beta} \leq \beta < 1$, a $\|f(x_i + \alpha_i^{k-1} s_i)\| \geq (1 - \tilde{\varepsilon}_1 \alpha_i^{k-1}) \|f_i\|$.
V druhém případě dostaneme

$$\|f(x_i + \frac{\alpha_i}{\beta} s_i)\| \geq \left(1 - \frac{\tilde{\varepsilon}_1}{\beta} \alpha_i\right) \|f_i\|$$

Označíme-li $d_i = (\alpha_i / \beta) s_i$ a použijeme-li nerovnost z důkazu Věty 67, můžeme psát

$$(\|f(x_i + d_i)\| - \|f_i\|)\|f_i\| \geq -\frac{\tilde{\varepsilon}_1}{\beta} \alpha_i \|f_i\|^2 \geq \frac{\tilde{\varepsilon}_1}{1 - \overline{\lambda} - \overline{\omega}} \frac{\alpha_i}{\beta} s_i^T g_i = \frac{\tilde{\varepsilon}_1}{1 - \overline{\lambda} - \overline{\omega}} d_i^T g_i \stackrel{\Delta}{=} \tilde{\varepsilon}_2 d_i^T g_i$$

kde $0 < 2\tilde{\varepsilon}_2 < 1$, neboť $0 < 2\tilde{\varepsilon}_1 < (1 - \overline{\lambda} - \overline{\omega})$. Pokud $\|f(x_i + d_i)\| \geq \|f_i\|$, je také $F(x_i + d_i) \geq F(x_i)$ a zřejmě platí

$$F(x_i + d_i) - F(x_i) \geq \tilde{\varepsilon}_2 d_i^T g_i$$

Pokud $\|f(x_i + d_i)\| \leq \|f_i\|$, můžeme psát

$$(\|f(x_i + d_i)\| - \|f_i\|)\|f_i\| \leq \frac{1}{2}(\|f(x_i + d_i)\| - \|f_i\|)(\|f(x_i + d_i)\| + \|f_i\|) = F(x_i + d_i) - F(x_i)$$

takže opět

$$F(x_i + d_i) - F(x_i) \geq \tilde{\varepsilon}_2 d_i^T g_i$$

a použijeme-li stejný postup jako v důkazu věty 68, dostaneme

$$\alpha_i \geq \frac{2\underline{\beta}(1 - \tilde{\varepsilon}_2)\underline{J}^2(1 - \overline{\lambda} - \overline{\omega})}{(\overline{J}^2 + 2\overline{LF})(1 + \overline{\lambda} + \overline{\omega})^2}$$

což po dosazení do ($\overline{\text{S4}}$) dává $\| f_{i+1} \| \leq q \| f_i \|$, kde

$$0 < q = \left(1 - \min\left(\tilde{\varepsilon}_1, \frac{2\underline{\beta}\tilde{\varepsilon}_1(1-\tilde{\varepsilon}_2)\underline{J}^2(1-\overline{\lambda}-\overline{\omega})}{(\overline{J}^2 + 2\overline{LF})(1+\overline{\lambda}+\overline{\omega})^2}\right)\right) < 1$$

neboť $0 < 2\tilde{\varepsilon}_1 < 1$, $0 < 2\tilde{\varepsilon}_2 < 1$, $\underline{J}/\overline{J} \le 1$ (takže $\underline{J}^2/(\overline{J}^2 + 2\overline{LF}) \le 1$), $(1-\overline{\lambda}-\overline{\omega})/(1+\overline{\lambda}+\overline{\omega})^2 \le 1$.

Poznámka 60 Jestliže $x_i \to x^*$ je rychlost konvergence R-lineární. Plyne to z odhadu $\| f_{i+1} \| \le q \| f_i \|$, $i \in N$, kde $0 < q < 1$, použitého v důkazu věty 68 a věty 69.

Věta 70 (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 ($\overline{S2}$) (nebo ($\overline{S4}$)). Nechť platí

$$\lim_{i\to\infty} \frac{\| A_i s_i + f_i \|}{\| f_i \|} = 0 \tag{α}$$

a

$$\lim_{i\to\infty} \frac{\| (A_i - J_i) s_i \|}{\| s_i \|} = 0 \tag{β}$$

Pak existuje index $k \in N$ takový, že $\alpha_i = 1$, $\forall i \ge k$ a posloupnost x_i, $i \in N$, konverguje superlineárně k bodu $x^* \in R^n$.

Důkaz Důkaz povedeme poněkud obecněji, neboť získané výsledky použijeme v důkazu věty 87. To znamená, že v částech (a)-(c) budeme předpokládat pouze, že

$$\limsup_{i\to\infty} \frac{\| A_i s_i + f_i \|}{\| f_i \|} \le \overline{\omega} < 1 \tag{$\overline{\alpha}$}$$

Platí-li (α), můžeme ve všech vzorcích položit $\overline{\omega} = 0$
(a) Ukážeme, že existuje index $k_1 \in N$ tak, že

$$\| f_i \| (1-\overline{\omega})/\overline{J} \le \| s_i \| \le \| f_i \| (1+\overline{\omega})/\underline{J}$$

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 ($\overline{\alpha}$) 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 ($\overline{\alpha}$) 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 $k_2 \ge k_1$ takový, že $-s_i^T g_i \ge \varepsilon_0 \| s_i \| \| g_i \|$ $\forall i \ge k_2$, pokud $\varepsilon_0 < (1-\overline{\omega})/\kappa(J^*)$. Zvolíme $\underline{J} < 1/ \| (J^*)^{-1} \|$ a $\overline{J} > \| J^* \|$ tak, aby platilo $\varepsilon_0 = (1-\overline{\omega})\underline{J}/\overline{J}$. Z definice ω_i a ϑ_i a z (a) plyne, že

$$
\begin{aligned}
-s_i^T g_i &\geq -f_i^T J_i s_i = -f_i^T (A_i s_i + f_i) + f_i^T (A_i - J_i) s_i + f_i^T f_i \geq \\
&\geq (1 - \| \omega_i \| - \| \vartheta_i \| (1+\overline{\omega})/\underline{J}) \| f_i \|^2 \geq \\
&\geq \frac{\underline{J}}{\| J_i \|} (1 - \| \omega_i \| - \| \vartheta_i \| (1+\overline{\omega})/\underline{J}) \| g_i \| \| s_i \|
\end{aligned}
$$

a jelikož $\| \omega_i \| \leq \overline{\omega}$ a $\| \vartheta_i \| \to 0$ (podle $(\overline{\alpha})$ a (β)) a $\| J_i \| \to \| J^* \| < \overline{J}$, existuje index $k_2 \geq k_1$ tak, že $-s_1^T g_i \geq (\underline{J}/\overline{J})(1-\overline{\omega}) \| s_i \| \| g_i \| \; \forall i \geq k_2$.
(c) Ukážeme, že existuje index $k \geq k_2$ tak, že hodnota $\alpha_i = 1$ vyhovuje podmínce $(\overline{\mathrm{S4}})$, pokud $\overline{\varepsilon}_1 < 1 - \overline{\omega}$, nebo podmínce $(\overline{\mathrm{S2}})$, pokud $\overline{\varepsilon}_1 < (1-\overline{\omega})/2$. 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

$$
\frac{\| f(x_i + s_i) \|}{\| f_i \|} \leq \| \omega_i \| + \| \vartheta_i \| (1+\overline{\omega})/\underline{J} + o(\| f_i \|)/\| f_i \|
$$

takže $\limsup_{i\to\infty} \| f(x_i + s_i) \| / \| f_i \| \leq \overline{\omega}$ (podle $(\overline{\alpha})$ a (β)). Pokud $\overline{\varepsilon}_1 < 1 - \overline{\omega}$, existuje index $k \geq k_2$ tak, že podmínka $(\overline{\mathrm{S4}})$ s $\alpha_i = 1$ je splněna $\forall i \geq k$ (platí-li (α), může být číslo $0 < \overline{\varepsilon}_1 < 1/2$ libovolné, neboť $\| f(x_i + s_i) \| / \| f_i \| \to 0$). Podobně (za předpokladu, že $\| f(x_i + s_i) \| \leq \| f_i \|$) dostaneme

$$
\begin{aligned}
-\frac{2(F(x_i + s_i) - F_i)}{\| f_i \|^2} &= -\frac{(\| f(x_i+s_i) \| - \| f_i \|)(\| f(x_i+s_i) \| + \| f_i \|)}{\| f_i \|^2} \geq \\
&\geq -\frac{\| f(x_i+s_i) \| - \| f_i \|)}{\| f_i \|} \geq \\
&\geq 1 - \| \omega_i \| - \| \vartheta_i \| (1+\overline{\omega})/\underline{J} - o(\| f_i \|)/\| f_i \|
\end{aligned}
$$

a jelikož platí

$$
\tilde{g}^T s_i = f_i^T A_i s_i = f_i^T (A_i s_i + f_i) - f_i^T f_i \leq -(1 - \| \omega_i \|) \| f_i \|^2
$$

můžeme psát

$$
\frac{F(x_i+s_i) - F_i}{s_i^T \tilde{g}_i} \geq -\frac{F(x_i+s_i) - F_i)}{\| f_i \|^2 (1 - \| \omega_i \|)} \geq \frac{1}{2} (1 - \| \omega_i \| - \| \vartheta_i \| (1+\overline{\omega})/\underline{J} + o(\| f_i \|)/\| f_i \|)
$$

takže $\liminf_{i\to\infty} (F(x_i + s_i) - F_i)/s_i^T \tilde{g}_i \geq (1-\overline{\omega})/2$ (podle $(\overline{\alpha})$ a (β)). Pokud $\overline{\varepsilon}_1 < (1-\overline{\omega})/2$, existuje index $k \geq k_2$ tak, že podmínka $(\overline{\mathrm{S2}})$ s $\alpha_i = 1$ je splněna $\forall i \geq k$ (platí-li (α), může být číslo $0 < \overline{\varepsilon}_1 < 1/2$ libovolné, neboť $(F(x_i + s_i) - F_i)/s_i^T \tilde{g}_i \to 1/2$).
(d) Předpokládejme nyní že platí (α). Pomocí vět o střední hodnotě dostaneme

$$
\frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq \frac{\overline{J}}{\underline{J}} \frac{\| f_{i+1} \|}{\| f_i \|}
$$

takže podle (α), (β) a (c) 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.

5.2 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 $\| f(x_i + s) \| - \| f(x_i) \|$, označení

$$\omega_i(s) = (A_i s + f_i) / \| f_i \|$$

pro přesnost určení směrového vektoru, a označení

$$\rho_i(s) = \frac{\| f(x_i + s) \| - \| f_i(x_i) \|}{L_i(s)}$$

pro podíl skutečného a předpověděného poklesu normy funkce $f : R^n \to R^n$.

Definice 32 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 směrové vektory $s_i \in R^n$, $i \in N$, se určují tak, že

$$\| s_i \| \leq \Delta_i \tag{$\overline{\mathrm{T1a}}$}$$

$$\| s_i \| < \Delta_i \Rightarrow \| \omega_i(s_i) \| \leq \overline{\omega} \tag{$\overline{\mathrm{T1b}}$}$$

$$-L_i(s_i) \geq 2\underline{\sigma} \min(\| s_i \|, \| f_i \|) \tag{$\overline{\mathrm{T1c}}$}$$

kde $0 < \underline{\sigma} < 1/2$ a $0 \leq \overline{\omega} < 1$ a kde délky kroku $\alpha_i \geq 0$, $i \in N$, se vybírají tak, že

$$\rho_i(s_i) \leq 0 \Rightarrow \alpha_i = 0 \tag{$\overline{\mathrm{T2a}}$}$$

$$\rho_i(s_i) > 0 \Rightarrow \alpha_i = 1 \tag{$\overline{\mathrm{T2b}}$}$$

Přitom posloupnost $\Delta_i > 0$, $i \in N$, se konstruuje tak, že

$$\rho_i(s_i) < \underline{\rho} \Rightarrow \underline{\beta} \| s_i \| \leq \Delta_{i+1} \leq \overline{\beta} \| s_i \| \tag{$\overline{\mathrm{T3a}}$}$$

$$\rho_i(s_i) \geq \underline{\rho} \Rightarrow \Delta_i \leq \Delta_{i+1} \leq \overline{\gamma} \Delta_i \tag{$\overline{\mathrm{T3b}}$}$$

kde $0 < \underline{\beta} < \overline{\beta} < 1 < \overline{\gamma}$ a $0 < \underline{\rho} < 1/2$, a matice A_i, $i \in N$, vyhovují podmínce

$$\| (J_i - A_i) s_i \| \leq \overline{\vartheta} \| s_i \| \tag{$\overline{\mathrm{T4}}$}$$

kde $0 < \overline{\vartheta} < (1 - 2\underline{\rho})\underline{\sigma}$.

Poznámka 61 Definice 32 se liší od definice 22 tvarem podmínky ($\overline{\mathrm{T1c}}$) a přidáním nové podmínky ($\overline{\mathrm{T4}}$). Podmínka ($\overline{\mathrm{T4}}$) specifikuje přesnost aproximace Jacobiovy matice a má jiný tvar než odpovídající podmínka u metod spádových směrů. Z podmínky ($\overline{\mathrm{T4}}$) plyne

$$| \| A_i s_i \| - \| J_i s_i \| | \leq \overline{\vartheta} \| s_i \|$$

takže buď $\| A_i s_i \| \leq \| J_i s_i \|$ nebo $\| A_i s_i \| \geq \| J_i s_i \|$ a $\| A_i s_i \| \leq \| J_i s_i \| + \overline{\vartheta} \| s_i \|$. Spojením těchto nerovností dostaneme $\| A_i s_i \| \leq \| J_i s_i \| + \overline{\vartheta} \| s_i \|$ a je-li splněna podmínka (J3) platí

$$\| A_i s_i \| \leq (\overline{J} + \overline{\vartheta}) \| s_i \|$$

Lemma 71 Aplikujeme-li metodu s lokálně omezeným krokem ($\overline{\mathrm{T1}}$) – ($\overline{\mathrm{T4}}$) na funkci $f : R^n \to R^n$, která splňuje podmínky (J3)-(J5), existuje konstanta $0 < \underline{c} < 1$ tak, že

$$\| s_i \| \geq \underline{c} m_i$$

kde

$$m_i = \min_{1\le j\le i} \| f_j \|$$

Důkaz Opět budeme používat označení N_1, N_2, N_3 pro množiny indexů takové, že platí ($\overline{\text{T1b}}$), ($\overline{\text{T2b}}$), ($\overline{\text{T3b}}$).
(a) Nechť $i \in N_1$. Pak podle (T1b) platí

$$|\| A_i s_i \| - \| f_i \|| \le \| A_i s_i + f_i \| \le \overline{\omega} \| f_i \|$$

takže buď $\| f_i \| \le \| A_i s_i \|$ nebo $\| f_i \| \ge \| A_i s_i \|$ a zároveň $(1-\overline{\omega}) \| f_i \| \le \| A_i s_i \|$. Spojením těchto nerovností a použitím nerovnosti z poznámky 61 dostaneme $(1-\overline{\omega}) \| f_i \| \le \| A_i s_i \| \le (\overline{J} + \overline{\vartheta}) \| s_i \|$, což dává

$$\| s_i \| \ge \frac{1-\overline{\omega}}{\overline{J}+\overline{\vartheta}} m_i$$

(b) Nechť $i \notin N_3$. Podle ($\overline{\text{T1c}}$) je $L_i(s_i) \le 0$, takže

$$\begin{aligned}
\frac{1}{2} L_i(s_i) \| f_i \| &= \frac{1}{2} (\| A_i s_i + f_i \| - \| f_i \|) \| f_i \| \ge \\
&\ge \frac{1}{2} (\| A_i s_i + f_i \| - \| f_i \|)(\| A_i s_i + f_i \| + \| f_i \|) = \\
&= \frac{1}{2} (\| A_i s_i + f_i \|^2 - \| f_i \|^2) = f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i \triangleq Q_i(s_i)
\end{aligned}$$

Jesliže $\| f(x_i + s_i) \| \le \| f(x_i) \|$, můžeme podle ($\overline{\text{T3a}}$) psát

$$\begin{aligned}
F(x_i + s_i) - F(x_i) &= \frac{1}{2} (\| f(x_i + s_i) \|^2 - \| f(x_i) \|^2) = \\
&= \frac{1}{2} (\| f(x_i + s_i) \| - \| f(x_i) \|)(\| f(x_i + s_i) \| + \| f(x_i) \|) \ge \\
&\ge (\| f(x_i + s_i) \| - \| f(x_i) \|) \| f(x_i) \| \ge \\
&\ge \underline{\rho} L_i(s_i) \| f_i \| \ge 2\underline{\rho} Q_i(s_i)
\end{aligned}$$

Jestliže $\| f(x_i + s_i) \| \ge \| f(x_i) \|$, platí tato nerovnost triviálně. Můžeme tedy psát

$$F(x_i + s_i) - F(x_i) \ge 2\underline{\rho} Q_i(s_i) = 2\underline{\rho}(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i)$$

Z druhé strany s použitím ($\overline{\text{T4}}$) (jako v důkazu věty 68) dostaneme

$$\begin{aligned}
F(x_i + s_i) - F(x_i) &\le s_i^T g_i + \| s_i \| \| g(x_i + \lambda s_i) - g(x_i) \| \le s_i^T g_i + \frac{1}{2} (\overline{J} + 2\overline{LF}) \| s_i \|^2 = \\
&= f_i^T A_i s_i + f_i^T (J_i - A_i) s_i + \frac{1}{2} (\overline{J} + 2\overline{LF}) \| s_i \| \le \\
&\le f_i^T A_i s_i + \overline{\vartheta} \| s_i \| \| f_i \| + \frac{1}{2} (\overline{J} + 2\overline{LF}) \| s_i \|^2
\end{aligned}$$

Spojíme-li obě nerovnosti dostaneme

$$2\underline{\rho}(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i) \le f_i^T A_i s_i + \overline{\vartheta} \| s_i \| \| f_i \| + \frac{1}{2} (\overline{J}^2 + 2\overline{LF}) \| s_i \|^2$$

neboli

$$-(1-2\underline{\rho}) Q_i(s_i) = (2\underline{\rho} - 1)(f_i^T A_i s_i + \frac{1}{2} s_i^T A_i^T A_i s_i) \le \overline{\vartheta} \| s_i \| \| f_i \| + \frac{1}{2} (\overline{J}^2 + 2\overline{LF} + (\overline{J} + \overline{\vartheta})^2) \| s_i \|^2$$

(poznámka 61). Z ($\overline{\text{T1c}}$) však plyne

$$\begin{aligned} -Q_i(s_i) &= -\frac{1}{2}(\| A_i s_i + f_i \| - \| f_i \|)(\| A_i s_i + f_i \| + \| f_i \|) \geq \\ &\geq -\frac{1}{2}(\| A_i s_i + f_i \| - \| f_i \|) \| f_i \| = -\frac{1}{2} L_i(s_i) \| f_i \| \geq \underline{\sigma} \min(\| s_i \|, \| f_i \|) \| f_i \| \end{aligned}$$

takže buď

$$\| s_i \| \geq \| f_i \| \geq m_i$$

nebo $\| s_i \| \leq \| f_i \|$ a platí

$$-Q_i(s_i) \geq \underline{\sigma} \| s_i \| \| f_i \|$$

Použijeme-li poslední nerovnost, dostaneme

$$(1-2\underline{\rho})\underline{\sigma} \| s_i \| \| f_i \| \leq -(1-2\underline{\rho}) Q_i(s_i) \leq \overline{\vartheta} \| s_i \| \| f_i \| + \frac{1}{2}(\overline{J}^2 + 2\overline{LF} + (\overline{J} + \overline{\vartheta})^2) \| s_i \|^2$$

neboli

$$\| s_i \| \geq \frac{2((1-2\underline{\rho})\underline{\sigma} - \overline{\vartheta})}{(\overline{J}^2 + 2\overline{LF} + (\overline{J} + \overline{\vartheta})^2)} \| f_i \| \geq \frac{2((1-2\underline{\rho})\underline{\sigma} - \overline{\vartheta})}{(\overline{J}^2 + 2\overline{LF} + (\overline{J} + \overline{\vartheta})^2)} m_i$$

(čitatel je kladný neboť $\overline{\vartheta} < (1-2\underline{\rho})\underline{\sigma}$).
(c) Nechť $i = 1$. Pokud $\| f_1 \| = 0$, platí zřejmě $\| s_1 \| \geq \underline{c} \| f_1 \| \geq \underline{c} m_1$ pro libovolnou konstantu $\underline{c} > 0$. Pokud $\| f_1 \| \neq 0$, dostaneme

$$\| s_1 \| = \frac{\| s_1 \|}{\| f_1 \|} \| f_1 \| \geq \frac{\| s_1 \|}{\| f_1 \|} m_1$$

(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 $(\overline{\text{T3}})$ a $(\overline{\text{T1a}})$ můžeme psát

$$\| s_i \| \geq \Delta_i \geq \Delta_{k+1} \geq \min(\Delta_k, \underline{\beta} \| s_k \|) \geq \min(\| s_k \|, \underline{\beta} \| s_k \|) = \underline{\beta} \| s_k \|$$

takže podle (a)-(c) dostaneme

$$\| s_i \| \geq \underline{\beta} \| s_k \| \geq \underline{c} m_k \geq \underline{c} m_i$$

kde

$$\underline{c} = \underline{\beta} \min\left(\frac{1-\overline{\omega}}{\overline{J} + \overline{\vartheta}}, 1, \frac{2((1-2\underline{\rho})\underline{\sigma} - \overline{\vartheta})}{\overline{J}^2 + 2\overline{LF} + (\overline{J} + \overline{\vartheta})^2)}, \frac{\| s_1 \|}{\| f_1 \|} \right) < 1$$

Věta 72 Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem $(\overline{\text{T1}}) - (\overline{\text{T4}})$. Nechť funkce $f : R^n \to R^n$ splňuje podmínky (J3)-(J5). Pak platí

$$\liminf_{i \to \infty} \| f_i \| = 0$$

Důkaz Předpokládejme, že existuje číslo $\underline{\varepsilon} > 0$ tak, že $\| f_i \| \geq \underline{\varepsilon}$ $\forall i \in N$. Pak podle lemmatu 71 platí

$$\| s_i \| \geq \underline{c\varepsilon}$$

$\forall i \in N$. Předpokládejme nejprve, že množina N_3 je nekonečná. Protože $N_3 \subset N_2$, můžeme psát

$$\begin{aligned} \| f_i \| - \| f_{i+1} \| &= \| f(x_i) \| - \| f(x_i + s_i) \| \geq -\underline{\rho} L_i(s_i) \geq 2\underline{\rho}\underline{\sigma} \min(\| s_i \|, \| f_i \|) \geq \\ &\geq 2\underline{\rho}\underline{\sigma} \min(\underline{c\varepsilon}, \underline{\varepsilon},) = 2\underline{\rho}\underline{\sigma}\underline{c\varepsilon} \end{aligned}$$

$\forall \in N_3$, takže

$$
\begin{aligned}
\| f_1 \| &\geq \lim_{i\to\infty} (\| f_1 \| - \| f_{i+1} \|) = \sum_{i=1}^{\infty} (\| f_i \| - \| f_{i+1} \|) \geq \sum_{i\in N_3} (\| f_i \| - \| f_{i+1} \|) \geq \\
&\geq \sum_{i\in N_3} 2\underline{\rho}\underline{\sigma}\underline{c}\underline{\varepsilon} = \infty
\end{aligned}
$$

což je spor. Předpokládejme nyní, že množina N_3 není nekonečná. Pak podle ($\overline{\text{T3a}}$) platí $\Delta_i \to 0$, což spolu s ($\overline{\text{T1a}}$) dává $\| s_i \| \to 0$. To je ale ve sporu s předpokladem, že $\| s_i \| \geq \underline{c}\underline{\varepsilon}\ \forall \in N$.

Věta 73 (superlineární konvergence). Nechť $x_i \in R^n$, $i \in N$, je posloupnost generovaná metodou s lokálně omezeným krokem ($\overline{\text{T1}}$) – ($\overline{\text{T4}}$) taková, že $x_i \to x^*$. Nechť funkce $f : R^n \to R^n$ splňuje podmínky (J3)-(J5). Nechť

$$
\lim_{i\to\infty} \| \omega_i(s_i) \| = 0 \qquad (\alpha)
$$

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) Stejným způsobem jako v důkazu věty 70 se ukáže, že existuje index $k_1 \in N$ takový, že

$$
\| f_i \| / \overline{J} \leq \| s_i \| \leq \| f_i \| / \underline{J}
$$

$\forall i \geq k_1$ pokud $\| J^* \| \leq \overline{J}$ a $\| (J^*)^{-1} \| \leq 1/\underline{J}$ (nyní $\overline{\omega} = 0$ podle (α)).
(b) Ukážeme, že existuje index $k_2 \geq k_1$ takový, že

$$
-L_i(s_i) \geq 2\underline{\sigma} \| f_i \|
$$

$\forall i \geq k_2$ pokud $0 < \underline{\sigma} < 1/2$. Podle definice funkce $L_i(s_i)$ můžeme psát

$$
-L_i(s_i) = \| f_i \| - \| A_i s_i + f_i \| = \| f_i \| (1 - \| \omega_i(s_i) \|)
$$

a jelikož $\| \omega_i(s_i) \| \to 0$, existuje index $k_2 \geq k_1$ tak, že $(1 - \| \omega_i(s_i) \|) \geq 2\underline{\sigma}$ pokud $0 < \underline{\sigma} < 1/2$.
(c) Ukážeme, že existuje index $k_3 \geq k_2$ tak, 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

$$
\begin{aligned}
\rho_i(s_i) &= \frac{\| f(x_i) \| - \| f(x_i + s_i) \|}{-L_i(s_i)} \geq \frac{(1 - \| \omega_i \| - \| \vartheta_i \| / \underline{J}) \| f_i \| + o(\| f_i \|)}{(1 - \| \omega_i \|) \| f_i \|} = \\
&= \frac{1 - \| \omega_i \| - \| \vartheta_i \| / \underline{J}}{1 - \| \omega_i \|} + o(\| f_i \|) / \| f_i \|
\end{aligned}
$$

a jelikož $\| \omega_i \| \to 0$ a $\| \vartheta_i \| \to 0$ (podle (α) a (β)), dostaneme $\rho_i(s_i) \to 1$. Existuje tedy index $k_3 \geq k_2$ tak, že $\rho_i(s_i) \geq 2\underline{\rho}$ pokud $0 < \underline{\rho} < 1/2$.
(d) Ukážeme, že existuje index $\overline{k} \geq k_3$ takový, že $i \in N_1\ \forall i \geq \overline{k}$. Podle věty o střední hodnotě platí (pro $i \in N_3 \subset N_2$)

$$
f_{i+1} = f(x_i + s_i) = f_i + J_i s_i + o(\| s_i \|)
$$

(viz (c)). Označme

$$
\lambda_i = \frac{f_{i+1} - f_i - A_i s_i}{\| f_i \|}
$$

Pak podle předchozích úvah platí $\| \lambda_i \| \leq \| \vartheta_i \| / \overline{J} + o(\| f_i \|) / \| f_i \| \to 0$. Jelikož zároveň $\| \omega_i \| \to 0$, existuje index $k_4 \geq k_3$ takový, že $\| \lambda_i \| < (\underline{J}/\overline{J})/2$ a $\| \omega_i \| < (\underline{J}/\overline{J})/2 \ \forall i \geq k_4$. Jestliže $i \notin N_1 \ \forall i \geq k_4$, pak $\| s_i \| \geq \Delta_i \ \forall i \geq k_4$ a jelikož $i \in N_3 \ \forall i \geq k_4$, dostaneme $\| s_i \| \geq \Delta_{k_4}, \forall i \geq k_4$, což je ve sporu s tím, že $\| s_i \| \to 0$. Existuje tedy index $k \geq k_4$ takový, že $k \in N_1$. Předpokládejme, že $i \in N_1$ pro nějaký index $i \geq k$ (platí to jistě pro $i = k$). Pak můžeme psát

$$\begin{aligned} \| s_{i+1} \| &\leq \frac{1}{\underline{J}} \| f_{i+1} \| \leq \frac{1}{\underline{J}} (\| f_{i+1} - f_i - A_i s_i \| + \| A_i s_i + f_i \|) \leq \\ &\leq \frac{\overline{J}}{\underline{J}} (\| \lambda_i \| + \| \omega_i \|) \| s_i \| < \left(\frac{1}{2} + \frac{1}{2} \right) \| s_i \| = \| s_i \| \end{aligned}$$

Jelikož $i \in N_3$ podle (c), platí $\Delta_{i+1} \geq \Delta_i$, což dává $\| s_{i+1} \| < \| s_i \| \leq \Delta_i \leq \Delta_{i+1}$, takže $i \in N_1$. Indukcí dostaneme $i \in N_1 \ \forall i \geq k$.
(e) Superlineární konvergence. Platí

$$\frac{\| f_{i+1} \|}{\| f_i \|} \leq \frac{\| f_{i+1} - f_i - A_i s_i \| + \| A_i s_i + g_i \|}{\| f_i \|} \leq \| \lambda_i \| + \| \omega_i \|$$

což spolu s $\| \lambda_i \| \to 0$ a $\| \omega_i \| \to 0$ dává

$$\lim_{i \to \infty} \frac{\| x_{i+1} - x^* \|}{\| x_i - x^* \|} \leq \frac{\overline{J}}{\underline{J}} \frac{\| f_{i+1} \|}{\| f_i \|} = 0$$

5.3. Newtonova metoda

Newtonova metoda používá matice $A_i = J(x_i) \ \forall i \in N$, takže $(A_i - J_i) s_i / \| f_i \| = 0 \ \forall i \in N$.

Věta 74 Nechť jsou splněny podmínky (J3)-(J5) a nechť $\| \omega_i \| = \| A_i s_i + f_i \| / \| f_i \| \leq \overline{\omega} < 1 \ \forall i \in N$. 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 68 (nebo věty 69) a věty 72, neboť $\lambda_i = 0 \ \forall i \in N$. Superlineární konvergence plyne bezprostředně z věty 70 a věty 73, neboť $\vartheta_i = 0 \ \forall i \in N$.

Poznámka 63 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 75 Nechť je splněn předpoklad (J5) a nechť platí

$$A e_j = \frac{f(x + \delta e_j) - f(x)}{\delta} \tag{D}$$

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

$$f(x + \delta e_j) = f(x) + J(x) \delta e_j + \int_0^1 (J(x + \tau \delta e_j) - J(x)) \delta e_j d\tau$$

takže

$$
\begin{aligned}
\| (A - J(x))e_j \| &= \| \frac{f(x+\delta e_j) - f(x)}{\delta} - J(x)e_j \| \le \frac{1}{\delta} \| \int_0^1 (J(x+\tau\delta e_j) - J(x))\delta e_j d\tau \| \le \\
&\le \frac{1}{\delta}\frac{1}{2}\overline{L}\delta^2 \| e_j \|^2 = \frac{1}{2}\overline{L}\delta
\end{aligned}
$$

Nechť $s \in R^n$ je libovolný vektor s jednotkovou normou. Pak platí

$$
\begin{aligned}
\| (A - J(x))s \| &= \| \sum_{j=1}^{n} (A - J(x))e_j e_j^T s \| \le \sum_{j=1}^{n} |e_j^T s| \, \| (A - J(x))e_j \| \le \frac{1}{2}\overline{L}\delta \sum_{j=1}^{n} |e_j^T s| \le \\
&\le \frac{1}{2}\overline{L}\sqrt{n}\delta \| s \| = \frac{1}{2}\overline{L}\sqrt{n}\delta
\end{aligned}
$$

a jelikož

$$
\| A - J(x) \| = \max_{\|s\|=1} \| (A - J(x))s \|
$$

dostaneme tvrzení lemmatu.

Věta 76 Nechť $\| As + f \| \le \overline{\omega} \| f \|$, kde $0 \le \overline{\omega} < 1$, a matice A je určena podle vzorce (D), kde

$$
\delta < \frac{\underline{J}(1 - \overline{\omega})}{\overline{L}\sqrt{n}}
$$

Pak existuje číslo $\overline{\lambda}$, $0 < \overline{\omega} + \overline{\lambda} < 1$, takové, že $\| (A - J)s \| \le \overline{\lambda} \| f \|$.

Důkaz Z nerovností $\| As + f \| \le \overline{\omega} \| f \|$ a $\| (A - J)s \| \le \| A - J \| \| s \| \le \frac{1}{2}\overline{L}\sqrt{n}\delta \| s \|$ dostaneme (podobným způsobem jako v poznámce 61) odhad

$$
\| As \| \le (1 + \overline{\omega}) \| f \|
$$

a

$$
\left(\underline{J} - \frac{1}{2}\overline{L}\sqrt{n}\delta \right) \| s \| \le \| As \|
$$

Platí tedy

$$
\frac{1}{2}\overline{L}\sqrt{n}\delta \| s \| \le \frac{\frac{1}{2}\overline{L}\sqrt{n}\delta(1 + \overline{\omega})}{\underline{J} - \frac{1}{2}\overline{L}\sqrt{n}\delta} \| f \| \stackrel{\Delta}{=} \overline{\lambda} \| f \|
$$

Z nerovnosti $\overline{\omega} + \overline{\lambda} < 1$ dostaneme

$$
\frac{1}{2}\overline{L}\sqrt{n}\delta(1 + \overline{\omega}) < \left(\underline{J} - \frac{1}{2}\overline{L}\sqrt{n}\delta \right)(1 - \overline{\omega})
$$

takže

$$
\overline{L}\sqrt{n}\delta < \underline{J}(1 - \overline{\omega})
$$

neboli $\delta < \underline{J}(1 - \overline{\omega}) / \left(\overline{L}\sqrt{n}\right)$.

Poznámka 64 Věta 76 ukazuje, že lze zvolit diferenci $\delta > 0$ tak, aby matice určená podle vztahu (D) splňovala podmínku pro globální konvergenci metody spádových směrů. Je vidět, že diferenci δ je třeba zvolit tím menší, čím menší je číslo $\underline{J}$ v (J4) a čím větší je číslo $\overline{L}$ v (J5).

Poznámka 65 Používáme-li metodu s lokálně omezeným krokem $(\overline{\mathrm{T1}}) - (\overline{\mathrm{T4}})$, stačí volit

$$
\delta < \frac{2(1 - 2\underline{\rho})\underline{\sigma}}{\overline{L}\sqrt{n}}
$$

Pak platí $\| (A - J(x))s \| \leq \overline{\vartheta} \| s \|$, kde $0 < \overline{\vartheta} < (1 - \underline{\rho})\underline{\sigma}$.

5.4 Kvazinewtonovské metody

Definice 33 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 66 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 77 Nechť $A_+ = A + uv^T$ a $Ad \neq y$. Pak $A_+ d = y$ právě tehdy, jestliže $v^T d \neq 0$ a $u = (y - Ad)/v^T d$, takže

$$A_+ = A + \frac{(y - Ad)v^T}{v^T d} \tag{$\overline{\text{A}}$}$$

Jestliže $Ad = y$ stačí položit $u = v = 0$, takže $A_+ = A$.

Důkaz Z podmínky $A_+ d = y$ dostaneme $A_+ d = Ad + uv^T d = y$. Jestliže $Ad = y$, stačí položit $u = v = 0$, takže $A_+ = A$. Jestliže $Ad \neq y$, musí platit $v^T d \neq 0$ a $u = (y - Ad)/v^T d$.

Poznámka 67 Položíme-li $v = d$ dostaneme Broydenovu dobrou metodu

$$A_+ = A + \frac{(y - Ad)d^T}{d^T d} \tag{$\overline{\text{AG}}$}$$

Položíme-li $v = A^T y$, dostaneme Broydenovu špatnou metodu

$$A_+ = A + \frac{(y - Ad)y^T A}{y^T A d} \tag{$\overline{\text{AB}}$}$$

Nechť

$$e_k^T d = \max_{1 \leq i \leq n} e_i^T d$$

Položíme-li $v = e_k$, dostaneme přímou metodu aktualizace sloupců

$$A_+ = A + \frac{(y - Ad)e_k^T}{e_k^T d} \tag{$\overline{\text{AD}}$}$$

která aktualizuje vždy pouze jeden sloupec matice A.

Věta 78 Nechť A je regulární matice a nechť platí ($\overline{\text{A}}$). Pak matice A_+ je regulární právě tehdy, jestliže $v^T A^{-1} y \neq 0$.

Důkaz Nechť $A_+ = A + uv^T$. Pak podle Shermanova-Morrisonova vzorce platí

$$A_+^{-1} = A^{-1} - \frac{A^{-1} u v^T A^{-1}}{1 + v^T A^{-1} u}$$

takže A_+ je regulární právě tehdy, jestliže $1 + v^T A^{-1} u \neq 0$. Dosadíme-li do této nerovnosti $u = (y - Ad)/v^T d$, dostaneme

$$1 + v^T A^{-1} u = 1 + \frac{v^T A^{-1} y - v^T d}{v^T d} = v^T A^{-1} y$$

takže A_+ je regulární právě tehdy, jestliže $v^T A^{-1} y \neq 0$.

Poznámka 68 Věta 78 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 79 (Aktualizace matice $S = A^{-1}$). Nechť jsou splněny předpoklady věty 78. Nechť $S = A^{-1}$ a nechť A_+ je matice určená podle aktualizace ($\overline{\mathrm{A}}$), kde $v^T A^{-1} y \neq 0$. Nechť $S_+ = A_+^{-1}$. Pak platí

$$S_+ = S + \frac{(d - Sy) v^T S}{v^T S y} \tag{$\overline{\mathrm{S}}$}$$

Důkaz Podle Shermanova-Morrisonova vzorce (důkaz věty 78) platí

$$S_+ = S - \frac{S u v^T S}{\delta} = S + \frac{(d - Sy) v^T S}{\delta}$$

kde δ je zatím neznámé číslo. Z rovnice $S_+ y = d$ však plyne

$$S_+ y = Sy + \frac{v^T S y}{\delta}(d - Sy) = d$$

takže nutně $\delta = v^T S y$.

Poznámka 69 Položíme-li $v = d$, dostaneme Broydenovu dobrou metodu

$$S_+ = S + \frac{(d - Sy) d^T S}{d^T S y} \tag{$\overline{\mathrm{SG}}$}$$

Položíme-li $v = (S^{-1})^T y$, dostaneme Broydenovu špatnou metodu

$$S_+ = S + \frac{(d - Sy) y^T}{y^T y} \tag{$\overline{\mathrm{SB}}$}$$

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

$$S_+ = S + \frac{(d - Sy) e_k}{e_k^T y} \tag{$\overline{\mathrm{SI}}$}$$

Poznámka 70 (Dualita). Vztah ($\overline{\mathrm{S}}$) dostaneme ze vztahu ($\overline{\mathrm{A}}$) 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 71 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

$$x_{i+1} = x_i - S_i f_i \tag{α}$$

a

$$S_{i+1} = S_i + \frac{(d_i - S_i y_i) z_i^T}{z_i^T y_i} \tag{β}$$

$\forall i \in N$, 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 80 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

$$y_i = f_{i+1} - f_i = J^* d_i = -J^* S_i f_i \tag{γ}$$

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 81 Nechť jsou splněny předpoklady lemmatu 80 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 80 je vektor $(J^* S_{i-2l+2})^k f_{i-2l+2}$ lineární kombinací vektorů $(I - J^* S_{i-2l+1})(J^* S_{i-2l+1})$ f_{i-2l+1}, $0 \leq j \leq k$. Jelikož $j + 1$ 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 80 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 82 Nechť jsou splněny předpoklady lemmatu 80. 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 81 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 ($\overline{\text{S2}}$) nebo podmínku ($\overline{\text{S4}}$), 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 83 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 83 je speciálním případem věty 87.

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

6. 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

6.1. 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 $M = \{i \in N : \ i = (j-1)m + j, \ j \in N\}$, 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 ($\overline{\mathrm{S}}$)), 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 84 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

$$A_{k+1} = A_1 + (Y_k - A_1 D_k) R_k^{-1} V_k^T \tag{$\overline{\mathrm{AA}}$}$$

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 72 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

$$S_{k+1} = S_1 + (D_k - S_1 Y_k)(C_k - L_k + V_k^T S_1 Y_k)^{-1} V_k^T S_1 \tag{$\overline{\mathrm{SS}}$}$$

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 $A_{i+1} p = A_l p + U_k R_k^{-1} V_k^T p$.

6.2. 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 ($\delta = \sqrt{\varepsilon_M} / \parallel p \parallel$, 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 6.7).

6.3. 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, $1 \le j \le n$, 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 4.3). Rozdělme sloupce matice J do k disjunktních skupin $\mathcal{S}_i \subset \{1,\ldots,n\}$, $1 \le i \le k$, 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

$$\frac{f(x+\delta v_i) - f(x)}{\delta} \approx J v_i \qquad 1 \le i \le k$$

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 $J(\mathcal{S}_i)$). 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 $1 \le i \le n$, se počítají jen ty diference, které odpovídají nenulovým prvkům $(J(x))_{ij} \ne 0$. 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 5.3 (lemma 75).

6.4. 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 4.4 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 54 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

$$A_+ = A + \sum_{i=1}^{n} \frac{e_i^T (y - Ad) e_i (d^i)^T}{(d^i)^T d^i} \tag{$\overline{\text{AS}}$}$$

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 85 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 86 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 58 (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 85, 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 87 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

$$\begin{aligned}\| A_i d_i + f_i \| &\le \overline{\omega} \| f_i \| \\ x_{i+1} &= x_i + d_i \\ A_{i+1} &= \mathcal{P}_{QG} A_i\end{aligned}$$

$\forall i \in N$, kde $0 \le \overline{\omega} < 1$ (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) - (c) důkazu věty 70 můžeme přeformulovat (pomocí okolí) tak, že existují čísla $\delta > 0$, $\vartheta > 0$ taková, že pokud $\| x - x^* \| \le \delta$, $\| (A - J(x))d \| \le \vartheta \| d \|$ a $\| Ad + f \| \le \overline{\omega} \| f \|$, kde $0 \le \overline{\omega} < 1$, platí

$$\frac{1-\overline{\omega}}{\overline{J}} \| f \| \le \| d \| \le \frac{1+\overline{\omega}}{\underline{J}} \| f \|$$

kde $\| J^* \| < \overline{J}$ a $\| (J^*)^{-1} \| < 1/\underline{J}$ a

$$\| f(x+d) \| \le r \| f \|$$

(kde $\overline{\omega} < r < 1$). Zdůrazněme, že číslo $0 \le \overline{\omega} < 1$ může být libovolné zatímco čísla $\delta > 0$ a $\vartheta > 0$ mohou vycházet malá.
(a) Zvolme čísla $\overline{\delta} > 0$ a $\overline{\vartheta} > 0$ tak, aby platilo

$$\overline{\delta} \left(1 + \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \right) \le \delta$$

a

$$\overline{\vartheta} + \overline{L} \frac{\overline{J}}{\underline{J}} \frac{1+\overline{\omega}}{1-r} \overline{\delta} \le \vartheta / \sqrt{n}$$

Nechť $\| x_1 - x^* \| \le \overline{\delta}$ a $\| (A_1 - J(x_1) \| \le \overline{\vartheta}$. Dokážeme indukcí, že pro libovolný index $i \in N$ platí $\| x_i - x^* \| \le \delta$ a $\| A_i - J(x_i) \| \le \vartheta$. Pro $i = 1$ je toto tvrzení zřejmé. Předpokládejme platnost tohoto tvrzení pro $1 \le i \le k$. 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 85, 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 $\| f_i \| \to 0$, $\| d_i \| \to 0$ a tedy i $x_i \to x^*$.
(c) Podle lemmatu 85 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

$$
\begin{aligned}
\sum_{i=1}^{\infty} \frac{\| y_i - A_i d_i \|^2}{\| d_i \|^2} &\le \overline{M} \left(\| A_1 - J_1 \|_F - \lim_{i \to \infty} \| A_{i+1} - J_{i+1} \|_F \right) + \overline{ML}\sqrt{n} \sum_{i=1}^{\infty} \| d_i \| \le \\
&\le \overline{M} \| A_1 - J_1 \|_F + \overline{ML} \frac{\overline{J}}{\underline{J}} \frac{1 + \overline{\omega}}{1 - r} \| x_1 - x^* \| < \infty
\end{aligned}
$$

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 60) implikuje, že $x_i \to x^*$ Q-superlinárně.

6.5. 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 skupiny trojúhelníkových matic, L a U se nazývá symbolickou faktorizací. Na rozdíl od řídkého Choleského rozkladu (oddíl 4.3) 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 6.4. Výsledkem je aktualizace

$$U_+ = U + \sum_{i=1}^{n} \frac{e_i(v - Ud)e_i d^i}{(d^i)^T d^i} \tag{$\overline{\text{AD}}$}$$

kde d^i, $1 \le i \le n$, jsou vektory takové, že $d_j^i = d_j$, $U_{ij} \neq 0$ a $d_j^i = 0$, $U_{ij} = 0$ (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 sloupců. 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

$$e_i^T D_+ e_i e_i^T LUd = e_i^T Py$$

$1 \le i \le n$, neboli

$$e_i^T D_+ e_i = \frac{e_i^T Py}{e_i^T LUd} \tag{$\overline{\text{AR}}$}$$

Také metodu škálování sloupců je třeba po m iteračních krocích přerušovat s tím, že se provede nesymetrický trojúhelníkový rozklad $PJ = LU$.

6.6. 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 $\{1, \dots, n\} = \mathcal{S}_1 \cup \dots \cup \mathcal{S}_k$ sloupců matice J do k disjunktních skupin $\mathcal{S}_i$, $1 \le i \le k$, 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 6.3). Pak se v každém iteračním kroku určují sloupce matice J patřící pouze do jedné skupiny a ostatní sloupce se nemění. Konkrétněji, nechť $l = \text{mod}_k i$ ($\text{mod}_k i$ 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 skombinovat se Schubertovou metodou tak, že se v každém iteračním kroku po určení sloupců matice J, patřících do skupiny $\mathcal{S}_l$, provede navíc aktualizace $(\overline{\text{AS}})$.

6.7. 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 33 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 88 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 $\tilde{f}_i f_i \ne 0$ a $\tilde{p}_i q_i \ne 0$ $\forall 1 \le i \le n$. Pak platí $f_{n+1} = 0$ a vektor s_{n+1} je řešením soustavy rovnic $As + f = 0$.

Důkaz Předpokládejme, že $\tilde{f}_i^T f_i \ne 0$ a $\tilde{p}_i^T q_i \ne 0$ $\forall 1 \le i \le n$. 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, $1 \leq i \leq n$ (a také $\tilde{p}_i$, $1 \leq i \leq n$), jsou lineárně nezávislé. Jelikož $f_{n+1} = As_{n+1} + f$ (plyne to z rekurentních vztahů metody BCG), vektory $\tilde{f}_i$, $1 \leq i \leq n$, jsou lineárně nezávislé a

$$\tilde{f}_j^T f_{n+1} = 0 \qquad \forall 1 \leq j \leq n$$

musí platit $f_{n+1} = As_{n+1} + f = 0$.

Pro $i = 1$ $(\alpha) - (\gamma)$ platí, neboť není co dokazovat.

(a) Nechť $i \leq n$. 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

$$\tilde{f}_1 = -\tilde{p}_1$$

$$\tilde{f}_j = -\tilde{p}_j + \beta_{j-1} \tilde{p}_{j-1} \qquad \forall 1 < j \leq i$$

$$f_1 = -p_1$$

$$f_j = -p_j + \beta_{j-1} p_{j-1} \qquad \forall 1 < j \leq i$$

takže podle (a) platí

$$\tilde{f}_1^T f_{i+1} = -\tilde{p}_1^T f_{i+1} = 0$$

$$\tilde{f}_j^T f_{i+1} = -\tilde{p}_j^T f_{i+1} + \beta_{j-1} \tilde{p}_{j-1} f_{i+1} = 0 \qquad \forall 1 < j \leq i$$

$$f_1^T \tilde{f}_{i+1} = -p_1^T \tilde{f}_{i+1} = 0$$

$$f_j^T \tilde{f}_{i+1} = -p_j^T \tilde{f}_{i+1} + \beta_{j-1} p_{j-1} \tilde{f}_{i+1} = 0 \qquad \forall 1 < j \leq i$$

(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 73 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 89 Nechť jsou splněny předpoklady věty 88. 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

$$\mathcal{K}_i = \operatorname{span}\{f\ Af, \ldots, A^{i-1} f\}$$

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

$$\tilde{\mathcal{K}}_i = \operatorname{span}\{\tilde{f}\ (A^T)\tilde{f}, \ldots, (A^T)^{i-1} f\}$$

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 88), 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 88), tvoří tam bázi.

Poznámka 74 Nechť jsou splněny předpoklady věty 88. Pak platí

$$\begin{aligned}
f_i = \varphi_i(A) f \quad &, \quad \tilde{f}_i = \varphi_i^T(A) \tilde{f} \\
p_i = -\psi_i(A) f \quad &, \quad \tilde{p}_i = -\psi_i^T(A) \tilde{f}
\end{aligned}$$

$\forall 1 \leq i \leq n+1$, 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

$$\alpha_i = \frac{\tilde{f}_i^T f_i}{\tilde{p}_i^T A p_i} = \frac{\tilde{f}^T \varphi_i^2(A) f}{\tilde{f}^T A \psi_i^2(A) f}$$

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 90 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

$$\varphi_1^2 = I, \qquad \psi_1^2 = I, \qquad \varphi_1\psi_1 = I$$

a

$$\begin{aligned}\varphi_{i+1}\psi_i &= \varphi_i\psi_i - \alpha_i A \psi_i^2 \\ \varphi_{i+1}^2 &= \varphi_i^2 - \alpha_i A(\varphi_i\psi_i + \varphi_{i+1}\psi_i) \\ \varphi_{i+1}\psi_{i+1} &= \varphi_{i+1}^2 + \beta_i \varphi_{i+1}\psi_i \\ \psi_i^2 &= \varphi_{i+1}\psi_{i+1} + \beta_i(\varphi_{i+1}\psi_i + \beta_i \psi_i^2)\end{aligned}$$

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 34 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

$$\overline{s}_1 = 0, \qquad \overline{f}_1 = f, \qquad p_1 = f, \qquad u_1 = f$$

a

$$v_i = Ap_i, \qquad \alpha_i = \tilde{f}^T \overline{f}_i / \tilde{f}^T v_i$$

$$q_i = u_i - \alpha_i v_i$$

$$\overline{s}_{i+1} = \overline{s}_i - \alpha_i(u_i + q_i)$$

$$\overline{f}_{i+1} = \overline{f}_i - \alpha_i A(u_i + q_i), \qquad \beta_i = \tilde{f}^T \overline{f}_{i+1} / \tilde{f}^T \overline{f}_i$$

$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$

$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$

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 75 Jsou-li splněny předpoklady věty 88 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 88).

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 91 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 91) spolu s odpovídajícími rekurentními vztahy pro s_i jsou základem jednoduše zhlazené metody CGS.

Definice 35 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

$$v_i = A p_i, \quad \alpha_i = \tilde f^T f_i / \tilde f^T v_i$$
$$q_i = u_i - \alpha_i v_i$$
$$\overline{s}_{i+1} = \overline{s}_i - \alpha_i (u_i + q_i)$$
$$\overline{f}_{i+1} = \overline{f}_i - \alpha_i A (u_i + q_i), \quad \beta_i = \tilde f^T f_{i+1} / \tilde f^T \overline{f}_i$$
$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$
$$p_{i+1} = u_{i+1} + \beta_i (q_i + \beta_i p_i)$$
$$\lambda_i = -\frac{\overline{f}_{i+1}^T (f_i - \overline{f}_{i+1})}{\| f_i - \overline{f}_{i+1} \|^2}$$
$$s_{i+1} = \overline{s}_{i+1} + \lambda_i (s_i - \overline{s}_{i+1})$$

$$f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1})$$

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 36 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 76 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 91). 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 $f_i - f_{i+1}$, 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 předpodmiňovací 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

$$v_i = AC^{-1}p_i, \quad \alpha_i = \tilde{f}^T\overline{f}_i/\tilde{f}^T v_i$$
$$q_i = u_i - \alpha_i v_i$$
$$\overline{s}_{i+1} = \overline{s}_i + \alpha_i C^{-1}(u_i + q_i)$$
$$\overline{f}_{i+1} = \overline{f}_i + \alpha_i AC^{-1}(u_i + q_i), \quad \beta_i = \tilde{f}^T\overline{f}_{i+1}/\tilde{f}^T\overline{f}_i$$
$$u_{i+1} = \overline{f}_{i+1} + \beta_i q_i$$
$$p_{i+1} = u_{i+1} + \beta_i(q_i + \beta_i p_i)$$
$$[\lambda_i, \mu_i]^T = -(V_i^T V_i)^{-1} V_i^T \overline{f}_{i+1}$$

$$s_{i+1} = \overline{s}_{i+1} + \lambda_i(s_i - \overline{s}_{i+1}) + \mu_i C^{-1} p_i$$
$$f_{i+1} = \overline{f}_{i+1} + \lambda_i(f_i - \overline{f}_{i+1}) + \mu_i v_i$$

pro $1 \leq i \leq n$, (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 dobře podmíněná. Velmi účinné je předpodmiňování pomocí neúplného trojúhelníkového rozkladu.

$$P(A + E) = LU$$

kde L je dolní trojúhelníková matice s jednotkami na hlavní diagonále, U je horní trojúhelníková matice, P je permutační matice a E je matice zahrnující vliv potlačování nově vznikajících nenulových prvků. Permutační matice se volí tak, aby matice PA měla nenulové prvky (pivoty) na hlavní diagonále.

Nyní se budeme zabývat metodou GMRES, která patří mezi metody s dlouhými rekurentními vztahy. Princip metody GMRES spočívá v tom, že se generují ortogonálními vektory q_i $1 \leq i \leq n$, tak, že q_j $1 \leq j \leq i$, tvoří bázi v Krylovově podprostoru $\mathcal{K}_i$. Vektor $s_{i+1} \in R^n$ je 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 $1 \leq j \leq n$ se generují pomocí Grammova-Schmidtova ortogonalizačního procesu. Klasický Grammův-Schmidtův ortogonalizační proces používá rekurentní vztahy

$$\beta_1 q_1 = f$$

a

$$q_{i+1}^1 = Aq_i$$
$$\left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^1 \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j \end{aligned}\right\} 1 \leq j \leq i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1}$$

$1 \leq i \leq n-1$, kde koeficienty β_i, $1 \leq i \leq n$ se vybírají tak, aby vektory q_i, $1 \leq i \leq n$, měly jednotkovou normu. Stabilnější je modifikovaný Grammův-Schmidtův ortogonalizační proces

$$\beta_1 q_1 = f$$

a

$$q_{i+1}^1 = Aq_i$$
$$\left.\begin{aligned} \alpha_{ji} &= q_j^T q_{i+1}^j \\ q_{i+1}^{j+1} &= q_{i+1}^j - \alpha_{ji} q_j \end{aligned}\right\} 1 \leq j \leq i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1}$$

$1 \leq i \leq n-1$. Grammův-Schmidtův ortogonalizační proces generující báze Krylovových podprostorů $\mathcal{K}_i$, $1 \leq i \leq n$, se také nazývá Arnoldiovým procesem určeným maticí $A \in R^{n \times n}$ a vektorem $f \in R^n$.

Označíme-li $Q_i = [q_1, q_2, \ldots, q_i]$ a

$$H_i = \begin{bmatrix} \alpha_{11}, & \alpha_{12}, & \ldots, & \alpha_{1i} \\ \beta_2, & \alpha_{22}, & \ldots, & \alpha_{2i} \\ 0, & \beta_3, & \ldots, & \alpha_{3i} \\ \ldots & \ldots & \ldots & \ldots \\ 0, & 0, & \ldots, & \beta_{i+1} \end{bmatrix}$$

($H_i \in R^{(i+1) \times i}$ je horní Hessenbergova matice), můžeme Arnoldiův proces zapsat v maticovém tvaru

$$AQ_i = Q_{i+1}H_i$$

Položíme-li $s_{i+1} = Q_i z_i$, kde $z_i \in R^n$, platí

$$\| As_{i+1} + f \| = \| AQ_i z_i + f \| = \| Q_{i+1}H_i z_i + Q_{i+1}(\beta_1 e_1) \| = \| H_i z_i + \beta_1 e_1 \|$$

takže minimalizační podmínku můžeme zapsat ve tvaru

$$z_i = \arg\min_{z \in R^i} \| H_i z + \beta_1 e_1 \| \tag{$\overline{\mathrm{M}}$}$$

Věta 92 Nechť $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j \leq i$, $\mathcal{K}_i = \mathcal{K}_{i+1}$ 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 $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j \leq i$ a $\mathcal{K}_i = \mathcal{K}_{i+1}$, pak vektory q_i, $1 \leq j \leq i$, jsou lineárně nezávislé a $\beta_{i+1} = 0$. Platí tedy

$$AQ_i = Q_i \overline{H}_i$$

kde $\overline{H}_i \in R^{i\times i}$ je horní Hessenluyova matice, která vznikne z matice $H_i \in R^{(i+1)\times i}$ vyškrtnutím posledního řádku. Jelikož matice AQ_i má lineárně nezávislé sloupce a A je regulární), je matice $\overline{H}_i$ regulární a existuje řešení soustavy rovnic $\overline{H}_i z_i + \beta_1 e_1 = 0$. Položíme-li $s_{i+1} = Q_i z_i$ platí

$$\| As_{i+1} + f \| = \| \overline{H}_i z_i + \beta_1 e_1 \| = 0$$

Důsledek Metoda GMRES nalezne řešení soustavy rovnic $As + f = 0$ po nejvýše n krocích. Jestliže totiž $\mathcal{K}_j \neq \mathcal{K}_{j+1}$, $1 \leq j < n$, pak nutně $\mathcal{K}_n = \mathcal{K}_{n+1} = R^n$. Metoda GMRES nemůže selhat, neboť $\beta_{i+1} = 0$ implikuje $As_{i+1} + f = 0$.

Abychom mohli určit vektor z_i vyhovující podmínce ($\overline{\mathrm{M}}$), je třeba provést ortogonální rozklad

$$P_i(H_i z_i + \beta_1 e_1) = \begin{bmatrix} R_i \\ 0 \end{bmatrix} z_i + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix}$$

kde $P_i = \overline{P}_i \overline{P}_{i-1} \dots \overline{P}_1$ je součin Givensových matic elementárních rotací a

$$R_i = \begin{bmatrix} \rho_{11}, & \rho_{12}, & \dots, & \rho_{1i} \\ 0, & \rho_{22}, & \dots, & \rho_{2i} \\ 0, & 0, & \dots, & \rho_{ii} \end{bmatrix}, \qquad h_i = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \dots \\ \eta_i \end{bmatrix}$$

Je to postup, který byl již použit v metodě LSQR (oddíl 4.7), proto ho nebudeme znovu odvozovat. Uvedeme pouze výsledné rekurentní vztahy metody GMRES.

Definice 37 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

$$q^1_{i+1} = Aq_i$$

$$\overline{\alpha}_{1i} = q_1^T q^1_{i+1}, \qquad q^2_{i+1} = q^1_{i+1} - \overline{\alpha}_{1i} q_1$$

$$\left.\begin{array}{c} \alpha_{ji} = q_j^T q^j_{i+1}, \qquad q^{j+1}_{i+1} = q^j_{i+1} - \alpha_{ji} q_j \\ \rho_{j-1i} = \lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji} \\ \overline{\alpha}_{ji} = -\lambda_{j-1}\overline{\alpha}_{j-1i} + \tau_{j-1}\alpha_{ji} \end{array}\right\} 1 < j \leq i$$

$$\beta_{i+1} q_{i+1} = q^{i+1}_{i+1}$$

$$\rho_{ii} = \sqrt{\overline{\alpha}^2_{ii} + \beta^2_{i+1}}$$

$$\lambda_i = \frac{\overline{\alpha}_{ii}}{\rho_{ii}}, \qquad \tau_i = \frac{\beta_{i+1}}{\rho_{ii}}$$

$$\eta_i = \lambda_i \overline{\eta}_i, \qquad \overline{\eta}_{i+1} = -\tau_i \overline{\eta}_i$$

$1 \le i \le n$, 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

$$z_i = \arg \min_{z \in R^n} \| \begin{bmatrix} R_i \\ 0 \end{bmatrix} z + \begin{bmatrix} h_i \\ \overline{\eta}_{i+1} \end{bmatrix} \|$$

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

$$A\mathcal{K}_i = \operatorname{span}\{Af, \ldots, A^i f\}$$

odpadne použití ortogonálního rozkladu. Vektory q_j, $1 \le j \le i$ se opět určují pomocí Grammova-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 38 Nechť $A \in R^{n \times n}$ je regulární matice a $f \in R^n$. Pak iterační proces

$$f_1 = f, \qquad \beta_1 q_1 = Af$$

a

$$\gamma_i = q_i^T f_i$$
$$f_{i+1} = f_i - \gamma_i q_i$$
$$q_{i+1}^1 = A q_i$$
$$\left.\begin{array}{l} \alpha_{ji} = q_j^T q_{i+1}^j \\ q_{i+1}^{j+1} = q_{i+1}^j - \alpha_{ji} q_j \end{array}\right\} 1 \le j \le i$$
$$\beta_{i+1} q_{i+1} = q_{i+1}^{i+1}$$

$1 \le i \le n-1$, nazveme modifikovanou metodou GMRES určenou maticí $A \in R^{n \times n}$ a vektorem $f \in R^n$.

Jakmile modifikovaná metoda GMRES získá dostatečně malé rezidium ($\| f_{i+1} \| \le \overline{\omega} \| f \|$), můžeme proces ukončit a položit $s_{i+1} = [f, Q_{i-1}] z_i$, kde

$$\begin{bmatrix} \beta_1, & \alpha_{11}, & \dots, & \alpha_{1i-1} \\ 0, & \beta_2, & \dots, & \alpha_{2i-1} \\ \dots & \dots & \dots & \dots \\ 0, & 0, & \dots, & \beta_i \end{bmatrix} z_i = \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \dots \\ \gamma_i \end{bmatrix}$$

Poznámka 77 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 $C \approx A$.

6.8. Metody s lokálně omezeným krokem

Zhlazenou metodu CGS nebo metodu GMRES můžeme použít ke konstrukci metod s lokálně omezeným krokem. Nejprve se budeme zabývat nepřesnými metodami s lokálně omezeným krokem, kdy se zhlazenou metodou CGS nebo metodou GMRES generuje posloupnost vektorů $s_{i+1} \in R^n$, $1 \le i \le n$, které aproximují řešení soustavy rovnic $As + f = 0$, a pak se pokládá $s = s_{i+1}$, pokud $\| s_{i+1} \| \le \Delta$ a $\| f_{i+1} \| \le \overline{\omega} \| f \|$, $0 \le \overline{\omega} < 1$, nebo $s = s_i + \alpha_i (s_{i+1} - s_i)$ a $\| s \| = \Delta$, pokud $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \| f \|$, $1 \le j \le i$, a $\| s_{i+1} \| \ge \Delta$. Tato volba zřejmě splňuje podmínky (T1a), (T1b) metody s lokálně omezeným krokem (definice 32). Navíc je třeba zformulovat předpoklady, aby byla splněna i podmínka (T1c), neboli

$$\| f \| - \| As + f \| \ge 2 \underline{\sigma} \min(\| s \|, \| f \|)$$

kde $\underline{\sigma}$ je nějaká konstanta. Nejprve se zaměříme na metodu GMRES.

Lemma 93 Nechť $s_{i+1} \in R^n$, $1 \le i \le n$, jsou vektory generované metodou GMRES a $f_{i+1} = As_{i+1} + f$. Pak platí

$$\| s_{i+1} \|^2 \le \frac{1}{\underline{A}^2} \left(\| f \|^2 - \| f_{i+1} \|^2 \right)$$

kde $\underline{A}$ je číslo takové, že $\| A^{-1} \| \le 1/\underline{A}$.

Důkaz Nechť q_j, $1 \le j \le i$, je báze v $\mathcal{K}_i$ generovaná metodou GMRES. Označíme-li $Q_i = [q_1, \dots, q_i]$, platí $Q_i^T Q_i = I$ a $s_{i+1} = Q_i z_i$, takže

$$\| s_{i+1} \|^2 = z_i^T Q_i^T Q_i z_i = \| z_i \|^2$$

z druhé strany platí

$$\| f_{i+1} \| = \| A s_{i+1} + f \| = \| Q_{i+1}(H_i z_i + \beta_1 e_1) \| = \| H_i z_i + \beta_1 e_1 \|$$

Nechť $P_i = \overline{P}_i \overline{P}_{i-1} \dots \overline{P}_1$ je ortogonální matice (součin Givensových matic elementárních rotací) taková, že

$$P_i H_i = \begin{bmatrix} R_i \\ 0 \end{bmatrix}, \qquad P_i(\beta_1 e_1) = \begin{bmatrix} h_i \\ \| f_{i+1} \| \end{bmatrix}$$

Pak platí

$$\| h_i \|^2 + \| f_{i+1} \|^2 = \| P_i(\beta_1 e_1) \|^2 = \beta_1^2 = \| f \|^2$$

a z rovnice $R_i z_i + h_i = 0$ plyne

$$\| z_i \|^2 \le \| R_i^{-1} \|^2 \| h_i \|^2$$

spojením obou výsledků dostaneme

$$\| s_{i+1} \|^2 = \| z_i \|^2 \le \| R_i^{-1} \|^2 \left(\| f \|^2 - \| f_{i+1} \|^2 \right)$$

Protože platí

$$A Q_i = Q_{i+1} H_i = Q_{i+1} P_i^T R_i$$

můžeme psát

$$Q_i^T A^T A Q_i = R_i^T P_i Q_{i+1}^T Q_{i+1} P_i^T R_i = R_i^T R_i$$

takže matice R_i má stejná singulární čísla jako matice AQ_i. Ale

$$\frac{1}{\| R_i^{-1} \|^2} = \min_{\|u\|=1} u^T R_i^T R_i u = \min_{\|u\|=1} u^T Q_i^T A^T A Q_i u \ge \min_{\|v\|=1} v^T A^T A v = \frac{1}{\| A^{-1} \|^2} \ge \underline{A}^2$$

takže platí $\| z_i \|^2 \le \left(\| f \|^2 - \| f_{i+1} \|^2 \right) / \underline{A}^2$.

Lemma 94 Nechť jsou splněny předpoklady lemmatu 93 a nechť $s = s_{i+1}$, pokud $\| s_{i+1} \| \le \Delta$ a $\| f_{i+1} \| \le \overline{\omega} \, \| f \|$, $0 \le \overline{\omega} < 1$ nebo $s = s_i + \alpha_i(s_{i+1} - s_i)$ a $\| s \| = \Delta$, pokud $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \, \| f \|$, $1 \le j \le i$, a $\| s_{i+1} \| \ge \Delta$. Pak jestliže $\Delta \ge \underline{\gamma} \, \| f \|$, platí

$$\| f \| - \| As + f \| \ge 2\underline{\sigma} \min(\| s \|, \| f \|)$$

kde $2\underline{\sigma} = \min\left(1 - \overline{\omega}, \underline{A}^2 \underline{\gamma}/8\right)$

Důkaz Nechť $\| s_{i+1} \| \le \Delta$ a $\| f_{i+1} \| \le \overline{\omega} \, \| f \|$, $0 \le \overline{\omega} < 1$. Pak platí $s = s_{i+1}$ a tedy

$$\| f \| - \| As + f \| = \| f \| - \| f_{i+1} \| \ge (1 - \overline{\omega}) \, \| f \|$$

takže dostaneme tvrzení lemmatu s $2\underline{\sigma} = 1 - \overline{\omega}$. Nechť naopak $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \, \| f \|$, $1 \le j \le i$, a $\| s_{i+1} \| \ge \Delta$. Pak platí $s = \alpha_i s_{i+1} + (1 - \alpha_i) s_i$, kde $0 < \alpha_i \le 1$, takže $\| s \| \le \alpha_i \, \| s_{i+1} \| + (1 - \alpha_i) \, \| s_i \|$. Z konvexity normy dostaneme

$$\| As + f \| = \| \alpha_i (A s_{i+1} + f) + (1 - \alpha_i)(A s_i + f) \| \le \alpha_i \, \| f_{i+1} \| + (1 - \alpha_i) \, \| f_i \|$$

Předpokládejme nejprve, že $\|s_i\| \ge \Delta/2$. Potom podle Lemmatu 93 platí

$$\underline{A}^2 \|s_i\| \Delta/2 \le \|f\|^2 - \|f_i\|^2 \le 2\|f\| \left(\|f\| - \|f_i\| \right)$$

(totéž platí i pro $\|s_{i+1}\|$ and $\|f_{i+1}\|$), takže

$$\begin{aligned}\|f\| - \|As+f\| &\geq \alpha_i\left(\|f\| - \|f_{i+1}\|\right) + (1-\alpha_i)\left(\|f\| - \|f_i\|\right) \geq \\ &\geq \frac{1}{4}\underline{A}^2\underline{\gamma}\left(\alpha_i\|s_{i+1}\| + (1-\alpha_i)\|s_i\|\right) \geq \\ &\geq \frac{1}{4}\underline{A}^2\underline{\gamma}\|s\|\end{aligned}$$

takže dostaneme tvrzení lemmatu s $2\underline{\sigma} = \underline{A}^2\underline{\gamma}/4$. Předpokládejme nyní, že $\|s_i\| \leq \Delta/2$. Potom z konvexity normy plyne, že

$$\|f\| - \|As+f\| \geq \alpha_i\left(\|f\| - \|f_{i+1}\|\right) + (1-\alpha_i)\left(\|f\| - \|f_i\|\right) \geq \alpha_i\left(\|f\| - \|f_{i+1}\|\right)$$

Protože $s = s_i + \alpha_i(s_{i+1} - s_i)$ dostaneme

$$\Delta = \|s\| \leq \|s_i\| + \alpha_i(\|s_{i+1}\| + \|s_i\|) \leq \Delta/2 + 2\alpha_i\|s_{i+1}\|$$

takže

$$\alpha_i \geq \frac{1}{4}\frac{\Delta}{\|s_{i+1}\|}$$

a použijeme-li Lemma 93 můžeme psát

$$\begin{aligned}\|f\| - \|As+f\| &\geq \frac{1}{4}\frac{\Delta}{\|s_{i+1}\|}\left(\|f\| - \|f_{i+1}\|\right) \geq \frac{1}{8}\frac{\underline{\gamma}}{\|s_{i+1}\|}\left(\|f\|^2 - \|f_{i+1}\|^2\right) \geq \\ &\frac{1}{8}\underline{A}^2\underline{\gamma}\|s_{i+1}\| \geq \frac{1}{8}\underline{A}^2\underline{\gamma}\|s\|\end{aligned}$$

takže dostaneme tvrzení lemmatu s $2\underline{\sigma} = \underline{A}^2\underline{\gamma}/8$.

Věta 95 Uvažujme metodu s lokálně omezeným krokem, kde směrové vektory $s_i \in R^n$, $i \in N$, se určují pomocí metody GMRES způsobem uvedeným v lemmatu 94, přičemž matice A_i, $i \in N$, jsou stejnoměrně regulární ($\| A_i \|^{-1} \leq 1/\underline{A}\ \forall i \in N$). Pak jestliže $\Delta_i \geq \underline{\gamma} \| f_i \|\ \forall i \in N$, je splněna podmínka ($\overline{\text{T1}}$). Jsou-li navíc splněny předpoklady (J3)-(J5) a podmínka ($\overline{\text{T4}}$), platí

$$\liminf_{i\to\infty} \| f_i \| = 0$$

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 94 a věty 72.

Nyní nahradíme podmínku $\Delta_i \geq \underline{\gamma} \| f_i \|\ \forall i \in N$, podmínkou $|f_i^T A_i f_i| \geq \underline{\varepsilon} \| f_i \| \| A f_i \|\ \forall i \in N$, která zaručuje, že metoda CGS neselže již v prvním kroku.

Lemma 96 Nechť $s_{i+1} \in R^n$, $1 \leq i \leq n$, jsou vektory generované metodou GMRES nebo dvojnásobně zhlazenou metodou CGS. Pak platí

$$\| f \|^2 - \| f_{i+1} \|^2 \geq \left(\frac{f^T A f}{\| f \| \| A f \|}\right)^2 \| f \|^2$$

Důkaz Jelikož normy reziduí metody GMRES i zhlazené metody CGS monotonně klesají, stačí dokázat, že platí

$$\| f \|^2 - \| f_2 \|^2 \geq \left(\frac{f^T A f}{\| f \| \| A f \|}\right)^2 \| f \|^2$$

Jelikož $s_1 = 0$, a $\mathcal{K}_1 = \text{span}\{f\}$, dostaneme z (M) podmínku

$$\| f_2 \| = \| A(\mu_1 f) + f \| = \min_{\mu \in R} \| A(\mu f) + f \|$$

neboli

$$\mu_1 = \arg\min_{\mu \in R} \| A(\mu f) + f \|^2 = \mu^2 \| Af \|^2 + 2\mu f^T Af + \| f \|^2$$

což dává $\mu = -f^T Af / \| Af \|^2$, takže

$$\| f_2 \|^2 = \frac{(f^T Af)^2}{\| Af \|^4} \| Af \|^2 - 2\frac{(f^T Af)^2}{\| Af \|^2} + \| f \|^2 = \| f \|^2 - \frac{(f^T Af)^2}{\| Af \|^2 \| f \|^2} \| f \|^2$$

odkud plyne tvrzení lemmatu pro metodu GMRES. Použijeme-li dvojnásobně zhlazenou metodu CGS, platí

$$\| f_2 \| = \min_{[\lambda,\mu]^T \in R^2} \| \overline{f}_2 + \lambda(f_1 - \overline{f}_2) + \mu v_1 \| \le \min_{\mu \in R} \| f_1 + \mu v_1 \| = \min_{\mu \in R} \| f + \mu Af \|$$

(po dosazení $\lambda = 1$), takže dostaneme stejný výsledek jako v případě metody GMRES.

Lemma 97 Nechť jsou splněny předpoklady lemmatu 96 a nechť $s \in R^n$ je vektor určený stejně jako v lemmatu 94. Pak jestliže $|f^T Af| \ge \underline{\varepsilon} \| f \| \| Af \|$, platí

$$\| f \| - \| As + f \| \ge 2\underline{\sigma} \min(\| s \|, \| f \|)$$

kde $2\underline{\sigma} = \min(\underline{\varepsilon}^2/2, A\underline{\varepsilon}^2/8)$.

Důkaz Nechť $\| s_{i+1} \| \le \Delta$ a $\| f_{i+1} \| \le \overline{\omega} \| f \|$, $0 \le \overline{\omega} < 1$. Pak podle lemmatu 96 platí

$$\begin{aligned} 2 \| f \| (\| f \| - \| f_{i+1} \|) &\ge (\| f \| + \| f_{i+1} \|)(\| f \| - \| f_{i+1} \|) = \\ &= \| f \|^2 - \| f_{i+1} \|^2 \ge \left(\frac{f^T Af}{\| f \| \| Af \|} \right)^2 \| f \|^2 \end{aligned}$$

neboli

$$\| f \| - \| f_{i+1} \| \ge \frac{1}{2} \left(\frac{f^T Af}{\| f \| \| Af \|} \right)^2 \| f \| \ge \frac{1}{2} \underline{\varepsilon}^2 \| f \|$$

takže dostaneme tvrzení lemmatu s $2\underline{\sigma} = \underline{\varepsilon}^2/2$. Nechť naopak $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \| f \|$, $1 \le j \le i$, a $\| s_{i+1} \| \ge \Delta$. Jestliže $i > 1$, můžeme postupovat stejně jako v důkazu lemmatu 94, takže

$$\begin{aligned} \| f \| - \| As + f \| &\ge \alpha_i (\| f \| - \| f_{i+1} \|) + (1 - \alpha_i)(\| f \| - \| f_i \|) \ge \\ &\ge (\alpha_i + 1 - \alpha_i) \frac{1}{2} \underline{\varepsilon}^2 \| f \| = \frac{1}{2} \underline{\varepsilon}^2 \| f \| \end{aligned}$$

a opět dostaneme tvrzení lemmatu s $2\underline{\sigma} = \underline{\varepsilon}^2/2$. Jestliže $i = 1$, platí $s = \alpha s_2$, kde $0 < \alpha \le 1$. Můžeme tedy psát

$$\begin{aligned} \| f \|^2 - \| As + f \|^2 &= \| f \|^2 - \alpha^2 \| As_2 \|^2 - 2\alpha f^T As_2 - \| f \|^2 = \\ &= -\alpha^2 \| As_2 \|^2 - 2\alpha f^T As_2 \ge \alpha \left(- \| As_2 \| - 2 f^T As_2 \right) = \\ &= \alpha \left(\| f \|^2 - \| As_2 + f \|^2 \right) \end{aligned}$$

(jelikož $\alpha^2 \le \alpha$ pro $0 < \alpha \le 1$), neboli

$$\begin{aligned} 2 \| f \| (\| f \| - \| As + f \|) &\ge \| f \|^2 - \| As + f \|^2 \ge \alpha \left(\| f \|^2 - \| f_2 \| \right) \ge \\ &\ge \alpha \| f \| (\| f \| - \| f_2 \|) \end{aligned}$$

což dává

$$\| f \| - \| As + f \| \geq \frac{1}{2}\alpha(\| f \| - \| f_2 \|) \geq \frac{1}{4}\alpha\underline{\varepsilon}^2 \| f \|$$

Z rovnosti $f_2 = As_2 + f$ plyne $s_2 = A^{-1}(f_2 - f)$, takže

$$\| s_2 \| \leq \| A^{-1} \| (\| f_2 \| + \| f \|) \leq \frac{2}{\underline{A}} \| f \|$$

což po dosazení do předchozí nerovnosti dává

$$\| f \| - \| As + f \| \geq \frac{1}{8}\alpha\underline{A\varepsilon}^2 \| s_2 \| = \frac{1}{8}\underline{A\varepsilon}^2 \| s \|$$

takže dostaneme tvrzení lemmatu s $2\underline{\sigma} = \underline{A\varepsilon}^2/8$.

Věta 98 Uvažujme metodu s lokálně omezeným krokem, kde směrové vektory $s_i \in R^n$, $i \in N$, se určují pomocí metody GMRES nebo dvojnásobně zhlazené metody CGS způsobem uvedeným v lemmatu 94, přičemž matice A_i, $i \in N$, jsou stejnoměrně regulární ($\| A_i \|^{-1} \leq 1/\underline{\Delta}$ $\forall i \in N$. Pak jestliže $|f_i^T A_i f_i| \geq \underline{\varepsilon} \| f_i \| \| A_i f_i \|$ $\forall i \in N$, je splněna podmínka $(\overline{\text{T1}})$. Jsou-li navíc splněny předpoklady (J3)-(J5) a podmínka $(\overline{\text{T4}})$, platí

$$\liminf_{i \to \infty} \| f_i \| = 0$$

Důkaz Tvrzení věty je bezprostředním důsledkem lemmatu 97 a věty 72.

Metodu GMRES nebo dvojnásobně zhlazenou metodu CGS můžeme také použít ke konstrukci metod, které se nazývají metodami psí nohy. V tomto případě se generují vektory $s_{i+1} \in R^n$, $1 \leq i \leq m$, kde $m \ll n$ (obvykle $1 \leq m \leq 3$). Jestliže pro nějaký index $1 \leq i \leq m$ platí $\| s_{i+1} \| \leq \Delta$ a $\| f_{i+1} \| \leq \overline{\omega} \| f \|$, $0 \leq \overline{\omega} < 1$, pokládáme $s = s_{i+1}$. Jestliže pro nějaký index $1 \leq i \leq m$ platí $\| s_j \| < \Delta$, $\| f_j \| > \overline{\omega} \| f \|$, $1 \leq j \leq i$, a $\| s_{i+1} \| \geq \Delta$, pokládáme $s = s_i + \alpha_i(s_{i+1} - s_i)$ tak, že $\| s \| = \Delta$. Nenastane-li ani jeden z těchto případů určíme pomocí některé přímé eliminační metody řešení $s^* \in R^n$ soustavy rovnic $As + f = 0$ a pokládáme $s = s_{m+1} + \alpha_{m+1}(s^* - s_{m+1})$. Jednoduše se dá ukázat (podobně jako v důkazu lemmatu 94 nebo v důkazu lemmatu 97), že pokud platí $\Delta \geq \underline{\gamma} \| f \|$ nebo $|f^T A f| \geq \underline{\varepsilon} \| f \| \| Af \|$, 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 nekompletního 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

7. 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:

$$F(x) = F_A(x,t_1) + f_T(y(x,t_1)) \tag{$\overline{\mathrm{O}}$}$$

kde

$$\frac{dy(x,t)}{dt} = f_S(x,y,t), \qquad y(x,t_0) = f_I(x)$$

$$\frac{dF_A(x,t)}{dt} = f_A(y,t), \qquad F_A(x,t_0) = 0 \tag{$\overline{\mathrm{D}}$}$$

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.

7.1. 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

$$g^T(x) = g_A^T(x,t_1) + \frac{\partial f_T(y(x,t_1))}{\partial y} u(x,t_1) \tag{$\overline{\mathrm{O1}}$}$$

kde

$$\frac{du(x,t)}{dt} = \frac{\partial f_S(x,y,t)}{\partial y} u(x,t) + \frac{\partial f_S(x,y,t)}{\partial x}, \qquad u(x,t_0) = \frac{df_I(x)}{dx}$$

$$\frac{dg_A^T(x,t)}{dt} = \frac{\partial f_A(y,t)}{\partial y} u(x,t), \qquad g_A^T(x,t_0) = 0 \tag{$\overline{\mathrm{D1}}$}$$

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.

7.2. Zpětný výpočet gradientů

Nechť $p(t)$ je libovolná funkce taková, že $p : [t_0, t_1] \rightarrow R^{n_S}$ 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

$$\begin{aligned} F(x) \;&=\; \int_{t_0}^{t_1} \{f_A(y,t) + p^T(t) f_S(x,y,t) + \frac{dp^T(t)}{dt} y(x,t)\}\, dt \\ &\quad + p^T(t_0) y(x,t_0) - p^T(t_1) y(x,t_1) + f_T(y(x,t_1))\ . \end{aligned}$$

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

$$g(x) = \tilde{g}_A(x,t_0) + (\frac{df_I(x)}{dx})^T p(x,t_0) \tag{$\overline{\mathrm{O2}}$}$$

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

$$-\frac{d\tilde{g}_A(x,t)}{dt} = (\frac{\partial f_S(x,y,t)}{\partial x})^T p(t), \qquad \tilde{g}_A(x,t_1) = 0 \tag{$\overline{\mathrm{D2}}$}$$

Celkem se řeší $n_S + 1$ diferenciálních rovnic v přímém směru a $2n_S + n$ diferenciálních rovnic ve zpětném směru.

7.3. Přímý výpočet Hessovy matice

Označme $v(x,t) = du(x,t)/dx = d^2y(x,t)/dx^2$, takže $v : R^n \times [t_0,t_1] \rightarrow R^{n_S \times n \times n}$. Derivováním $(\overline{\mathrm{O1}})$ a $(\overline{\mathrm{D1}})$ dostaneme

$$G(x) = G_A(x,t_1) + u^T(x,t_1)\frac{\partial^2 f_T(y(x,t_1))}{\partial y^2}u(x,t_1) + \frac{\partial f_T(y(x,t_1))}{\partial y}v(x,t_1) \qquad (\overline{\mathrm{O3}})$$

kde

$$\begin{aligned}\frac{dv(x,t)}{dt} &= \frac{\partial f_S(x,y,t)}{\partial y}v(x,t) \\ &+ \left[\frac{\partial^2 f_S(x,y,t)}{\partial y^2}u(x,t) + \frac{\partial^2 f_S(x,y,t)}{\partial y \partial x}\right]u(x,t) \\ &+ \frac{\partial^2 f_S(x,y,t)}{\partial x \partial y}u(x,t) + \frac{\partial^2 f_S(x,y,t)}{\partial x^2},\end{aligned}$$

$$v(x,t_0) = \frac{d^2 f_I(x)}{dx^2}$$

$$\frac{dG_A(x,t)}{dt} = u^T(x,t)\frac{\partial^2 f_A(y,t)}{\partial y^2}u(x,t) + \frac{\partial f_A(y,t)}{\partial y}v(x,t), \qquad G_A(x,t_0) = 0 \qquad (\overline{\mathrm{D3}})$$

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.

7.4. 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 $z : [t_0,t_1] \rightarrow R^{n_S}$, $W : [t_0,t_1] \rightarrow R^{n_S \times n_S}$ (SPD) (obecně $W_1 \neq W(t_1)$). Jestliže $F(x) \rightarrow 0$, pak nutně $y(x,t) \rightarrow z(t)$ takže $\partial f_A(y(x,t),t)/\partial y \rightarrow 0$ a $\partial f_T(y(x,t_1))/\partial y \rightarrow 0$. Můžeme tedy zanedbat tyto členy v $(\overline{\mathrm{O3}})$ a $(\overline{\mathrm{D3}})$. Dostaneme tak

$$G(x) \approx B(x) = B_A(x,t_1) + u^T(x,t_1)W_1 u(x,t_1) \qquad (\overline{\mathrm{O4}})$$

kde

$$\frac{dB_A(x,t)}{dt} = u^T(x,t)W(t)u(x,t), \qquad B_A(x,t_0) = 0 \qquad (\overline{\mathrm{D4}})$$

Celkem se řeší $(n_S+1)(n+1)+n^2$ diferenciálních rovnic v přímém směru.

Seznam nejdůležitějších označení

Obsah