Efektivní předpodmiňovače pro diferenční verzi Newtonovy metody

Ladislav Lukšan, Ctirad Matonoha, Jan Vlček
Ústav informatiky AVČR, Pod vodárenskou věží 2, 182 07 Praha 8
a Technická Univerzita Liberec, Hálkova 6, 461 17 Liberec

Minimalizace bez omezení

$$x^* = \arg\min_{i \in R^n} F(x), \quad F \in \mathcal{C}^2 : R^n \to R, \quad n - \text{velké}.$$

Označení

$$g(x) = \nabla F(x), \quad G(x) = \nabla^2 F(x),$$

$$\|G(x)\| \leq \overline{G}, \quad \forall x \in R^n.$$

Iterační krok

$$x_{k+1} = x_k + \alpha_k s_k, \quad k \in N,$$

kde s_k je směrový vektor a α_k je délka kroku. V tomto příspěvku se budu zabývat určením směrového vektoru s_k.

Newtonova metoda

$$F(x_k + s) \approx Q(x_k + s) = F(x_k) + g^T(x_k)s + \frac{1}{2}s^T G(x_k)s.$$

Výběr délky kroku

$$s_k = \arg \min_{s \in \mathcal{M}_k} Q(x_k + s).$$

Metoda spádových směrů

$$\mathcal{M}_k = R^n.$$

Metoda s lokálně omezeným krokem

$$\mathcal{M}_k = \{s \in R^n : \|s\| \leq \Delta_k\}.$$

V tomto příspěvku budu předpokládat, že matice G ani její struktura nejsou explicitně známy. Směrový vektor (minimum kvadratické funkce) se v tomto případě určuje iteračně metodou sdružených gradientů. Vnější index k budu většinou vynechávat.

Určení směrového vektoru metodou sdružených gradientů
(metoda spádových směrů)

$s_1 = 0, \quad g_1 = g, \quad h_1 = C^{-1} g_1, \quad \rho_1 = g_1^T h_1, \quad p_1 = -h_1.$
Do $i = 1$ **to** m
$q_i = G p_i, \quad \sigma_i = p_i^T q_i.$
If $\sigma_i < \varepsilon \|p_i\|$ **then** $s = s_i,$ stop.
$\alpha_i = \rho_i / \sigma_i, \quad s_{i+1} = s_i + \alpha_i p_i, \quad g_{i+1} = g_i + \alpha_i q_i,$
$h_{i+1} = C^{-1} g_{i+1}, \quad \rho_{i+1} = g_{i+1}^T h_{i+1}.$
If $\|g_{i+1}\| \leq \omega \|g_1\|$ **or** $i = m$ **then** $s = s_i,$ stop.
$\beta_i = \rho_{i+1} / \rho_i, \quad p_{i+1} = -h_{i+1} + \beta_i p_i.$
End do

Určení směrového vektoru metodou sdružených gradientů
(metoda s lokálně omezeným krokem)

$s_1 = 0, \quad g_1 = g, \quad h_1 = C^{-1} g_1, \quad \rho_1 = g_1^T h_1, \quad p_1 = -h_1.$

Do $i = 1$ **to** m

$q_i = G p_i, \quad \sigma_i = p_i^T q_i.$

If $\sigma_i \leq 0$ **then** $s = s_i + \lambda_i p_i, \ \|s\| = \Delta_i,$ stop.

$\alpha_i = \rho_i / \sigma_i.$

If $\|s_i + \alpha_i p_i\| \geq \Delta_i$ **then** $s = s_i + \lambda_i p_i, \ \|s\| = \Delta_i,$ stop.

$s_{i+1} = s_i + \alpha_i p_i, \quad g_{i+1} = g_i + \alpha_i q_i,$

$h_{i+1} = C^{-1} g_{i+1}, \quad \rho_{i+1} = g_{i+1}^T h_{i+1}.$

If $\|g_{i+1}\| \leq \omega \|g_1\|$ **or** $i = m$ **then** $s = s_i,$ stop.

$\beta_i = \rho_{i+1} / \rho_i, \quad p_{i+1} = -h_{i+1} + \beta_i p_i.$

End do

Diferenční verze nepřesné Newtonovy metody

Násobení $q = Gp$ se nahražuje numerickým derivováním

$$G(x)p \approx \frac{g(x + \delta p) - g(x)}{\delta},$$

kde $\delta = \varepsilon / \|p\|$ (obvykle $\varepsilon = \sqrt{\varepsilon_M}$, kde ε_M je strojová přesnost).

Věta 1 *Nechť funkce $F \in \mathcal{C}^2 : R^n \to R$ má lipschitzovsky spojité druhé derivace (s konstantou $\overline{L}$). Nechť $q = G(x)p$ a*

$$\tilde{q} = \frac{g(x + \delta p) - g(x)}{\delta}, \qquad \delta = \frac{\varepsilon}{\|p\|},$$

Pak

$$\|\tilde{q} - q\| \leq \frac{1}{2} \varepsilon \overline{L} \|p\|.$$

Věta 2 *Uvažujme metodu sdružených gradientů aplikovanou na soustavu lineárních rovnic* $G(x)s + g = 0$, *kde vektory* $q_i = G(x)p_i$ *jsou nahraženy vektory* $\tilde{q}_i = (g(x+\delta_i p_i) - g(x))/\delta_i$, $\delta_i = \varepsilon/\|p_i\|$. *Předpokládejme, že jsou splněny předpoklady věty 1 a označme*

$$s_{m+1} = \sum_{i=1}^{m} \alpha_i p_i, \quad g_{m+1} = \sum_{i=1}^{m} \alpha_i q_i, \quad \tilde{g}_{m+1} = \sum_{i=1}^{m} \alpha_i \tilde{q}_i$$

(takže $g_{m+1} = G(x)s_{m+1} + g$, *počítáme-li přesně). Pak platí*

$$\|\tilde{g}_{m+1} - g_{m+1}\| \leq \overline{\vartheta}\|s_{m+1}\|, \quad \overline{\vartheta} = \frac{m}{2}\varepsilon\overline{L}.$$

Poznámka 1 Předpokládejme, že v m-tém kroku metody sdružených gradientů platí $\|\tilde{g}_{m+1}\| \leq \overline{\omega}\|g\|$, $0 < \overline{\omega} < 1$. Pak, položíme-li $s = s_{m+1}$ a $\tilde{g} = \tilde{g}_{m+1}$, můžeme podle předpokladu a podle věty 2 psát

$$\frac{\|\tilde{G}s + g\|}{\|g\|} \leq \overline{\omega}, \qquad \frac{\|(\tilde{G} - G)s\|}{\|s\|} \leq \overline{\vartheta},$$

kde $\tilde{G}$ je nějaká symetrická matice, pro kterou platí $\tilde{G}s + g = \tilde{g}$, a kde $\overline{\vartheta} = m\varepsilon\overline{L}/2$. Tyto vztahy dovolují odhadnout asymptotickou rychlost konvergence.

Nevýhodou diferenční verze nepřesné Newtonovy metody je velký počet vnitřních iterací $\Rightarrow$ velký počet vyčíslení gradientů, je-li matice $G = G(x)$ špatně podmíněná. Proto je třeba metodu sdružených gradientů vhodně předpodmínit. Protože neznáme matici G, nelze použít standardní postupy. Budeme studovat tyto možnosti:

(1) Použití metody BFGS s omezenou pamětí.

(2) Použití pásových matic určených standardní metodou BFGS, která je ekvivalentní předpodmíněné metodě sdružených gradientů.

(3) Použití pásových matic určených numerickým derivováním.

(4) Použití tridiagonálních matic určených Lanczosovou metodou, která je ekvivalentní nepředpodmíněné metodě sdružených gradientů.

(1) Použití metody BFGS s omezenou pamětí

V k-tém kroku Newtonovy metody se jako předpodmiňovač použije matice $C_k^{-1} = H_k$. Matice $H_k = H_k^k$ se určuje rekurentně tak, že $H_{k-l}^k = \gamma_{k-l} I$, kde l je počet aktualizací (obvykle $l = 3$), a

$$\begin{aligned} H_{j+1}^k &= H_j^k + \left(\frac{y_j^T H_j^k y_j}{y_j^T d_j} + 1 \right) \frac{d_j d_j^T}{y_j^T d_j} - \frac{H_j^k y_j d_j^T + d_j (H_j^k y_j)^T}{y_j^T d_j} \\ &= V_j^T H_j^k V_j + \frac{d_j d_j^T}{y_j^T d_j} \end{aligned}$$

pro $k - l \leq j \leq k - 1$, kde

$$V_j = I - \frac{y_j d_j^T}{y_j^T d_j}, \quad d_j = x_{j+1} - x_j, \quad y_j = g_{j+1} - g_j.$$

Matice H_k se nekonstruuje, používají se Strangovy rekurence.

Strangovy rekurence

Máme vypočítat vektor $h_i = C_k^{-1} g_i = H_k\, g_i$ v i-tém vnitřním kroku metody sdružených gradientů, použité v k-tém vnějším kroku Newtonovy metody. Položíme $u_k = g_i$ a zpětnou rekurencí

$$\sigma_j = \frac{d_j^T u_{j+1}}{y_j^T d_j}, \qquad u_j = u_{j+1} - \sigma_j y_j$$

spočteme čísla σ_j a vektory u_j, $k-1 \geq j \geq k-l$. Pak položíme $v_{k-l} = \gamma_{k-l} u_{k-l}$ a přímou rekurencí

$$v_{j+1} = v_j + \left(\sigma_j - \frac{y_j^T v_j}{y_j^T d_j} \right) d_j$$

spočteme vektory v_{j+1}, $k-l \leq j \leq k-1$. Nakonec položíme $h_i = v_k$.

(2) Použití pásových matic určených standardní metodou BFGS

Metoda BFGS s přesným výběrem délky kroku, aplikovaná na ryze konvexní kvadratickou funkci, je ekvivalentní metodě sdružených gradientů. Metoda BFGS generuje posloupnost matic B_i, $1 \leq i \leq m$, tak, že $B_1 = C$ a

$$B_{i+1} = B_i + \frac{y_i y_i^T}{d_i^T y_i} - \frac{B_i d_i (B_i d_i)^T}{d_i^T B_i d_i} = B_i + \frac{Gp_i (Gp_i)^T}{p_i^T G p_i} + \frac{g_i g_i^T}{p_i^T g_i}$$

pro $1 \leq i \leq m$, kde $d_i = s_{i+1} - s_i = \alpha_i p_i$ a $y_i = g_{i+1} - g_i = G d_i$. Přitom p_i a g_i jsou vektory určené metodou sdružených gradientů. Použijeme-li místo vektorů $q_i = Gp_i$ a g_i vektory $\tilde{q}_i$ (určené numerickým derivováním) a $\tilde{g}_i$, můžeme psát $B_1 = C$ a

$$B_{i+1} = B_i + \frac{\tilde{q}_i \tilde{q}_i^T}{p_i^T \tilde{q}_i} + \frac{\tilde{g}_i \tilde{g}_i^T}{p_i^T \tilde{g}_i}, \quad 1 \leq i \leq m.$$

Z předchozího vyjádření je patrné, že k určení matic B_i, $1 \leq i \leq m$, se používají pouze vektory generované předpodmíněnou metodou sdružených gradientů (s maticovým násobením nahraženým numerickým derivováním). Matice B_i, $1 \leq i \leq m$, se v korekčních členech nevyskytují, takže můžeme ukládat pouze jejich vybrané pásy. Jsou-li vektory $\tilde{q_i}$ a $\tilde{g_i}$ dobrou aproximací vektorů q_i a g_i, jsou matice B_i, $1 \leq i \leq m$, pozitivně definitní a je-li počet kroků metody sdružených gradientů dostatečně velký, je matice $B = B_{m+1}$ dobrou aproximací matice G a můžeme ji (nebo její část) použít jako předpodmiňovač v dalším iteračním kroku Newtonovy metody. Vyšetříme tři speciální případy.

Diagonální předpodmínění

V případě, že $C = D$, kde D je diagonální matice obsahující diagonální prvky matice B, nenastávají žádné potíže, neboť pozitivně definitní matice B má kladné prvky na hlavní diagonále. Diagonální předpodmínění pro úlohy s řídkými Hessovými maticemi zdůvodňuje tato věta.

Věta 3 *Nechť $\mathcal{D}_n$ je množina všech diagonálních matic řádu n a D je diagonální matice obsahující diagonální prvky matice G. Pak platí*

$$\kappa(GD^{-1}) \leq l \min_{M \in \mathcal{D}_n} \kappa(GM^{-1}),$$

kde κ je spektrální číslo podmíněnosti a l je maximální počet nenulových prvků v řádcích matice G (pro pentadiagonální matici G je $l = 5$).

Tridiagonální předpodmínění

Nechť nyní $C = T$, kde T je tridiagonální matice obsahující prvky tří hlavních diagonál matice B. V tomto případě nemusí být matice C pozitivně definitní (i když B je pozitivně definitní). Jako příklad uvažujme matice

$$B = \begin{bmatrix} 2 & -2 & 2 \\ -2 & 3 & -3 \\ 2 & -3 & 4 \end{bmatrix}, \qquad T = \begin{bmatrix} 2 & -2 & 0 \\ -2 & 3 & -3 \\ 0 & -3 & 4 \end{bmatrix}.$$

Obě tyto matice mají kladné prvky na hlavní diagonále a kladné hlavní subdeterminanty druhého řádu. Platí ale $\det B = 2$ a $\det T = -10$, takže T není pozitivně definitní, i když B je pozitivně definitní. Abychom tento nedostatek odstranili, je třeba matici T upravit tak, aby byla pozitivně definitní.

Lemma 1 *Uvažujme tridiagonální matici*

$$T = \begin{bmatrix} \alpha_1 & \beta_1 & \dots & 0 & 0 \\ \beta_1 & \alpha_2 & \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & \dots & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

a označme Δ_i hlavní subdeterminant i-tého řádu matice T (obsahující řádky a sloupce s indexy $1, 2, \dots, i$). Pak platí $\Delta_1 = \alpha_1$ a

$$\Delta_i = \alpha_i \Delta_{i-1} - \beta_{i-1}^2 \Delta_{i-2}, \qquad 2 \leq i \leq n,$$

kde pokládáme $\Delta_0 = 1$.

Toto lemma lze použít k důkazu následující věty.

Věta 4 *Tridiagonální matice T je pozitivně definitní právě tehdy, když $\gamma_i > 0$ pro $1 \leq i \leq n$, kde $\gamma_1 = \alpha_1$ a*

$$\gamma_i = \alpha_i - \frac{\beta_{i-1}^2}{\gamma_{i-1}}, \qquad 2 \leq i \leq n.$$

Větu 4 můžeme použít tak, že počítáme čísla γ_i, $1 < i \leq n$, a pokud pro nějaký index platí $\gamma_i \leq 0$, zmenšíme mimodiagonální prvek β_{i-1} tak, aby platilo $\beta_{i-1}^2 < \gamma_{i-1}\alpha_i$ (například položíme $\beta_{i-1}^2 = \lambda_{i-1}\gamma_{i-1}\alpha_i$, kde $0 < \lambda_{i-1} < 1$). Potíž je v tom, že zvolíme-li λ_{i-1} nevhodně, může být výsledná tridiagonální matice špatně podmíněná. Pro praktické účely je výhodnější použít následující větu a její důsledek.

Věta 5 *Uvažujme tridiagonální matici T s kladnými prvky na hlavní diagonále. Pak jsou-li matice*

$$\begin{bmatrix} 2\alpha_1 & 2\beta_1 \\ 2\beta_1 & \alpha_2 \end{bmatrix}, \qquad \begin{bmatrix} \alpha_i & 2\beta_i \\ 2\beta_i & \alpha_{i+1} \end{bmatrix}, \qquad \begin{bmatrix} \alpha_{n-1} & 2\beta_{n-1} \\ 2\beta_{n-1} & 2\alpha_n \end{bmatrix},$$

kde $2 \leq i < n-2$, pozitivně semidefinitní a alespoň jedna z nich je pozitivně definitní, je matice T pozitivně definitní.

Důsledek 1 *Nechť tridiagonální matice T obsahuje hlavní diagonálu a poloviny vedlejších diagonál pozitivně definitní matice B (takže $\alpha_i = b_{i,i}$, $1 \leq i \leq n-1$ a $\beta_i = b_{i,i+1}/2$, $1 \leq i \leq n-1$). Pak T je pozitivně definitní.*

Důsledek 1 můžeme použít tak, že zmenšíme prvky vedlejších diagonál matice B na polovinu. Pak je výsledná tridiagonální matice pozitivně definitní. Větu 5 můžeme použít tak, že počítáme determinanty $\alpha_i\alpha_{i+1} - 4\beta_i^2$, $1 \leq i \leq n-1$, a pokud pro nějaký index platí $\alpha_i\alpha_{i+1} - 4\beta_i^2 < 0$, zmenšíme mimodiagonální prvek β_i tak, aby platilo $\beta_i^2 = \alpha_i\alpha_{i+1}/4$.

Pentadiagonální předpodmínění

Větu 5 a důsledek 1 lze zobecnit tak, že platí i pro obecnou pásovou matici. Ukážeme, jak to vypadá v případě pentadiagonální matice

$$P = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_1 & \dots & 0 & 0 & 0 \\ \beta_1 & \alpha_2 & \beta_2 & \dots & 0 & 0 & 0 \\ \gamma_1 & \beta_2 & \alpha_3 & \dots & 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 & \dots & \alpha_{n-2} & \beta_{n-2} & \gamma_{n-2} \\ 0 & 0 & 0 & \dots & \beta_{n-2} & \alpha_{n-1} & \beta_{n-1} \\ 0 & 0 & 0 & \dots & \gamma_{n-2} & \beta_{n-1} & \alpha_n \end{bmatrix}.$$

Na takto definovanou matici P se budeme často odvolávat.

Věta 6 *Uvažujme pentadiagonální matici P s kladnými prvky na hlavní diagonále. Pak, jsou-li matice*

$$\begin{bmatrix} \alpha_i & (3/2)\beta_i & 3\gamma_i \\ (3/2)\beta_i & \alpha_{i+1} & (3/2)\beta_{i+1} \\ 3\gamma_i & (3/2)\beta_{i+1} & \alpha_{i+2} \end{bmatrix}, \qquad 1 \leq i < n-2,$$

pozitivně semidefinitní, je matice P pozitivně definitní.

Důsledek 2 *Nechť pentadiagonální matice P obsahuje hlavní diagonálu, dvě třetiny prvních vedlejších diagonál a třetiny druhých vedlejších diagonál pozitivně definitní matice B (takže $\alpha_i = b_{i,i}$, $1 \leq i \leq n$, $\beta_i = 2b_{i,i+1}/3$, $1 \leq i \leq n-1$ a $\gamma_i = b_{i,i+2}/3$, $1 \leq i \leq n-2$). Pak P je pozitivně definitní.*

Důsledek 2 můžeme použít tak, že v matici B zmenšíme prvky prvních vedlejších diagonál na dvě třetiny a prvky druhých vedlejších diagonál na třetinu. Pak je výsledná tridiagonální matice pozitivně definitní. Větu 6 můžeme použít tak, že nejprve počítáme subdeterminanty $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2$, $1 \leq i \leq n-1$, a pokud je některý z nich záporný, zmenšíme mimodiagonální prvek β_i tak, aby platilo $\beta_i^2 = (4/9)\alpha_i\alpha_{i+1}$. Pak počítáme determinanty matic uvedených ve větě 6 a je-li některý z nich záporný, upravíme odpovídající prvek γ_i užitím následující věty.

Věta 7 *Determinanty Δ_i matic uvedených ve větě 6 spočteme podle vzorce*

$$\Delta_i = \alpha_{i+1}\left(\alpha_i\alpha_{i+2} - 9\gamma_i^2\right) - \frac{9}{4}\left(\alpha_i\beta_{i+1}^2 + \alpha_{i+2}\beta_i^2 - 6\beta_i\beta_{i+1}\gamma_i\right).$$

Determinant Δ_i je nezáporný právě tehdy, když $\underline{\gamma}_i \le \gamma_i \le \overline{\gamma}_i$, kde

$$\begin{aligned}\underline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} - \sqrt{D_i}\right),\\ \overline{\gamma}_i &= \frac{1}{3\alpha_{i+1}}\left(\frac{9}{4}\beta_i\beta_{i+1} + \sqrt{D_i}\right)\end{aligned}$$

jsou kořeny kvadratické rovnice $\Delta_i = 0$. Přitom

$$D_i = \left(\alpha_i\alpha_{i+1} - \frac{9}{4}\beta_i^2\right)\left(\alpha_{i+1}\alpha_{i+2} - \frac{9}{4}\beta_{i+1}^2\right)$$

je diskriminant této rovnice (vydělený číslem 36), který je nezáporný, pokud $\alpha_i\alpha_{i+1} - (9/4)\beta_i^2 \ge 0$ a $\alpha_{i+1}\alpha_{i+2} - (9/4)\beta_{i+1}^2 \ge 0$.

Poznámka 2 Věta 7 nabízí dvě možnosti, jak volit nový prvek γ_i v případě, že $\Delta_i < 0$. V prvním případě pokládáme $\gamma_i := \underline{\gamma}_i$, pokud $\gamma_i < \underline{\gamma}_i$, nebo $\gamma_i := \overline{\gamma}_i$, pokud $\gamma_i > \overline{\gamma}_i$. Tento způsob je náročnější na výpočet a dává horší praktické výsledky. Výhodnější je pokládat

$$\gamma_i = \frac{1}{2}(\underline{\gamma}_i + \overline{\gamma}_i) = \frac{3}{4}\frac{\beta_i \beta_{i+1}}{\alpha_{i+1}}.$$

(3) Použití pásových matic určených numerickým derivováním

Předpokládejme, že Hessova matice má pásovou strukturu (i když ve skutečnosti tomu tak není). Prvky této fiktivní pásové matice, kterou použijeme jako předpodmiňovač, lze určit numerickým derivováním. Provádí se to pouze jednou na začátku vnějšího kroku Newtonovy metody.

K určení všech prvků pásové matice, která má $k-1$ párů vedlejších diagonál (takže $k = (l+1)/2$, kde l je šířka pásu), stačí použít k diferencí gradientů, tedy spočítat v každém vnějším kroku Newtonovy metody k gradientů navíc. Vyšetříme opět tři speciální případy.

Diagonální předpodmínění

Poznámka 3 Předpokládejme, že Hessova matice je diagonální. Pak lze všechny její prvky aproximovat pomocí jedné diference gradientů

$$G(x)v \approx g(x+v) - g(x), \qquad v = [\delta_1, \ldots, \delta_n]^T,$$

kde $\delta_1, \ldots, \delta_n$ jsou vhodné diference. K předpodmínění pak použijeme diagonální matici $C = D = \text{diag}(\alpha_1, \ldots, \alpha_n)$, kde $Dv = g(x+v) - g(x)$. Po dosazení dostaneme $\alpha_i \delta_i = g_i(x+v) - g_i(x)$, neboli

$$\alpha_i = \frac{g_i(x+v) - g_i(x)}{\delta_i}, \qquad 1 \leq i \leq n.$$

Poznámka 4 Diference lze volit dvojím způsobem.

(1) Pokládáme $\delta_i = \delta$, $1 \leq i \leq n$, takže $v = \delta e$, kde e je vektor, jehož všechny prvky jsou jednotkové. Pak lze (tak jako ve větě 1) volit $\delta = \sqrt{\varepsilon_M}/\|e\| = \sqrt{\varepsilon_M/n}$.

(2) Pokládáme $\delta_i = \sqrt{\varepsilon_M}\max(|x_i|, 1)$, $1 \leq i \leq n$. Tento způsob je méně náchylný k vlivu zaokrouhlovacích chyb.

V obou případech lze psát $\delta_i = \varepsilon\overline{\delta}_i$, $1 \leq i \leq n$, kde $\varepsilon = \sqrt{\varepsilon_M}$ a buď $\overline{\delta}_i = 1/\sqrt{n}$ nebo $\overline{\delta}_i = \max(|x_i|, 1)$ pro $1 \leq i \leq n$.

Nevýhodou předpodmiňovačů založených na numerickém derivování je skutečnost, že nemusí být pozitivně definitní. Uvažujme ryze konvexní kvadratickou funkci $F : R^2 \to R$, kde

$$F(x) = \frac{1}{2} x^T \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x, \qquad g(x) = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} x.$$

Pak platí

$$\frac{g(x + \delta e) - g(x)}{\delta} = \begin{bmatrix} 1 & -2 \\ -2 & 6 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

takže

$$De = \begin{bmatrix} \alpha_1 & 0 \\ 0 & \alpha_2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 4 \end{bmatrix},$$

což dává $\alpha_1 = -1$, $\alpha_2 = 4$, takže matice D není pozitivně definitní.

Tuto nevýhodu můžeme odstranit tak, že pokládáme

$$\alpha_i = \frac{|g_i(x+v) - g_i(x)|}{\delta_i}, \qquad 1 \le i \le n.$$

Zdůvodnění této úpravy udává následující tvrzení.

Věta 8 *Nechť $\mathcal{D}_n$ je množina všech diagonálních matic řádu n a $D = \mathrm{diag}(\alpha_1, \ldots, \alpha_n)$ je diagonální matice taková, že*

$$\alpha_i = \sum_{j=1}^{n} |G_{ij}|, \qquad 1 \le j \le n,$$

kde G_{ij}, $1 \le j \le n$, jsou prvky i-tého řádku matice G. Pak platí

$$\kappa_1(GD^{-1}) = \min_{M \in \mathcal{D}_n} \kappa_1(GM^{-1}),$$

kde κ_1 je l_1 číslo podmíněnosti (součin l_1 norem matice a její inverze).

Má-li matice G pouze kladné prvky a položíme-li $v = \delta e$, můžeme psát $De = (g(x + \delta e) - g(x))/\delta \approx Ge$, takže

$$\alpha_i \approx \sum_{j=1}^{n} G_{ij} = \sum_{j=1}^{n} |G_{ij}|$$

a matice D je podle věty 8 ideálním diagonálním předpodmiňovačem (v l_1 normě) pro soustavu rovnic $Gs + g = 0$. Nemá-li matice G pouze kladné prvky platí

$$|\alpha_i| \approx \left| \sum_{j=1}^{n} G_{ij} \right| \leq \sum_{j=1}^{n} |G_{ij}|,$$

takže prvky upravené matice D jsou dolním odhadem prvků ideálního diagonálního předpodmiňovače.

Tridiagonální předpodmínění

Věta 9 *Nechť Hessova matice funkce F je tridiagonální (jako matice T). Položme $v_1 = [\delta_1, 0, \delta_3, 0, \delta_5, 0, \ldots]$, $v_2 = [0, \delta_2, 0, \delta_4, 0, \delta_6, \ldots]$, kde $\delta_i = \varepsilon\overline{\delta}_i$, $1 \leq i \leq n$. Pak pro $1 < i < n$ platí*

$$\alpha_1 = \lim_{\varepsilon\to 0} \frac{g_1(x+v_1) - g_1(x)}{\delta_1}, \quad \beta_1 = \lim_{\varepsilon\to 0} \frac{g_1(x+v_2) - g_1(x)}{\delta_2},$$

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, \quad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 1,$$

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, \quad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+1}}, \quad \mathrm{mod}(i,2) = 0,$$

$$\alpha_n = \lim_{\varepsilon\to 0} \frac{g_n(x+v_1) - g_n(x)}{\delta_n}, \quad \mathrm{mod}(n,2) = 1,$$

$$\alpha_n = \lim_{\varepsilon\to 0} \frac{g_n(x+v_2) - g_n(x)}{\delta_n}, \quad \mathrm{mod}(n,2) = 0.$$

Poznámka 5 Věta 9 udává způsob určení tridiagonálního předpodmiňovače. Zvolí se pevně číslo ε (například $\varepsilon = \sqrt{\varepsilon_M}$) a prvky matice $C = T$ se vypočtou podle vzorců uvedených ve větě 9 (ve kterých je vynechán limitní přechod).

Matice $C = T$ získaná podle poznámky 5 nemusí být pozitivně definitní, i když Hessova matice je pozitivně definitní (jako příklad lze uvést ryze konvexní kvadratickou funkci tří proměnných s pozitivně definitní Hessovou maticí

$$G = \begin{bmatrix} 1 & -1 & -2 \\ -1 & 4 & -1 \\ -2 & -1 & 8 \end{bmatrix}.$$

Uvedeme dvě věty, které podporují volbu tridiagonálního předpodmínění v případech, kdy skutečná Hessova matice je pentadiagonální.

Věta 10 *Nechť Hessova matice $G(x)$ je pentadiagonální, pozitivně definitní a diagonálně dominantní. Pak, platí-li $\delta_i = \varepsilon\overline{\delta}$, $1 \leq i \leq n$, a je-li číslo ε dostatečně malé, je matice $C = T$ získaná podle poznámky 5 pozitivně definitní a diagonálně dominantní.*

Poznámka 6 Věta 10 vyžaduje, aby všechny diference byly stejné, což je splněno například tehdy, když $\delta_i = \sqrt{2\varepsilon_M/n}$, $1 \leq i \leq n$. Numerické testy však ukazují, že volba $\delta_i = \sqrt{\varepsilon}\max(|x_i|, 1)$, $1 \leq i \leq n$, je výhodnější.

Pro mnohé praktické úlohy je matice T pozitivně definitní. Uvažujme okrajovou úlohu pro obyčejnou diferenciální rovnici druhého řádu

$$y''(t) = \varphi(y(t)), \quad 0 \le t \le 1, \quad y(0) = y_0, \quad y(1) = y_1,$$

kde funkce $\varphi : R \to R$ je dvakrát spojitě diferencovatelná na R. Rozdělíme-li interval $[0, 1]$ na $n + 1$ částí pomocí uzlových bodů $t_i = ih$, $0 \le i \le n + 1$, kde $h = 1/(n + 1)$ je krok sítě, a nahradíme-li druhé derivace v uzlových bodech diferencemi

$$y''(t_i) = \frac{y(t_{i-1}) - 2y(t_i) + y(t_{i+1})}{h^2},$$

kde $1 \le 1 \le n$, dostaneme soustavu n nelineárních rovnic

$$h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1} = 0,$$

kde $x_i = y(t_i)$, $0 \le 1 \le n + 1$, takže $x_0 = y_0$ a $x_{n+1} = y_1$.

Řešíme-li tuto soustavu metodou nejmenších čtverců, má minimalizovaná funkce tvar

$$F(x) = \frac{1}{2}\sum_{i=1}^{n} f_i^2(x) = \frac{1}{2}\sum_{i=1}^{n}\left(h^2\varphi(x_i) + 2x_i - x_{i-1} - x_{i+1}\right)^2,$$

kde $x = [x_1, \ldots, x_n]^T$.

Věta 11 *Aplikujme diferenční verzi Newtonovy metody na uvedený součet čtverců, kde funkce $\varphi : R \to R$ je lineární. Pak, platí-li $\delta_i = \varepsilon\overline{\delta}$, $1 \leq i \leq n$, a je-li číslo ε dostatečně malé, je matice $C = T$ získaná podle poznámky 5 pozitivně definitní.*

Pentadiagonální předpodmínění

Věta 12 *Nechť Hessova matice funkce F je pentadiagonální (jako matice P). Položme $v_1 = [\delta_1, 0, 0, \delta_4, 0, 0, \ldots]$, $v_2 = [0, \delta_2, 0, 0, \delta_5, 0, \ldots]$, $v_3 = [0, 0, \delta_3, 0, 0, \delta_6, \ldots]$, kde $\delta_i = \varepsilon\bar{\delta}_i$, $1 \leq i \leq n$. Pak platí*

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$

$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 1,$$

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$

$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 2,$$

$$\alpha_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_3) - g_i(x)}{\delta_i}, \qquad \beta_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_1) - g_i(x) - \delta_{i-2}\gamma_{i-2}}{\delta_{i+1}},$$

$$\gamma_i = \lim_{\varepsilon\to 0} \frac{g_i(x+v_2) - g_i(x) - \delta_{i-1}\beta_{i-1}}{\delta_{i+2}}, \qquad \mathrm{mod}(i,3) = 0,$$

(4) Použití tridiagonálních matic určených Lanczosovou metodou

Prvky tridiagonální matice T získané Lanczosovou metodou lze určit z koeficientů metody sdružených gradientů (které označíme vlnkou) podle převodních vztahů $\alpha_1 = 1/\tilde{\alpha}_1$ a

$$\beta_i^2 = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i^2}, \quad \alpha_{i+1} = \frac{\tilde{\beta}_i}{\tilde{\alpha}_i} + \frac{1}{\tilde{\alpha}_{i+1}}, \quad 1 \leq i \leq m,$$

kde m je číslo takové, že $\tilde{\alpha}_i > 0$ pro $1 \leq i \leq m$.

Věta 13 *Uvažujme metodu sdružených gradientů (aplikovanou na kvadratickou funkci s Hessovou maticí G) takovou, že $\tilde{\alpha}_i > 0$ pro $1 \leq i \leq m$. Pak tridiagonální matice T_m řádu m s prvky určenými podle převodních vztahů je pozitivně definitní.*

Poznámka 7 Tridiagonální matice T_m má dimenzi $m \leq n$. Abychom dostali předpodmiňovač dimenze n, položíme

$$\begin{aligned} C &= [Q_m, Q_{n-m}] \begin{bmatrix} T_m & 0 \\ 0 & I_{n-m} \end{bmatrix} [Q_m, Q_{n-m}]^T \\ &= (I - Q_m Q_m^T) + Q_m T_m Q_m^T \end{aligned}$$

kde Q_m je matice s m ortonormálními sloupci získaná symetrickým Lanczosovým procesem a Q_{n-m} je matice s $n-m$ ortonormálními sloupci taková, že matice $[Q_m, Q_{n-m}]$ je čtvercová a ortogonální.

Věta 14 *Nechť jsou splněny předpoklady věty 13. Pak předpodmiňovač uvedený v poznámce 7 je pozitivně definitní a platí*

$$C^{-1} = (I - Q_m Q_m^T) + Q_m T_m^{-1} Q_m^T.$$

Odvrhování předpodmiňovačů

Je důležité umět rozhodnout, zda předpodmiňovač použít nebo odvrhnout (týká se to hlavně pásových předpodmiňovačů určených metodou BFGS nebo numerickým derivováním). Indefinitní předpodmiňovač je nevhodný i v případě, že Hessova matice není pozitivně definitní.

K testování pozitivní definitnosti a špatné podmíněnosti matice se hodí Gillův–Murrayův rozklad. Pokud v některém eliminačním kroku je pivot menší než $\delta \max(1, \max_{1 \leq i \leq n}(|\alpha_i|))$, kde δ je předepsaná mez, rozklad předpodmiňovače ukončíme a předpodpodmiňovač odvrhneme. Provést Gillův-Murrayův rozklad až do konce a použít získanou pozitivně definitní matici jako předpodmiňovač se nevyplácí (dokládají to numerické experimenty). Číslo δ se obvykle volí tak, že $\delta = 10^{-12}$. Někdy je však třeba zvolit větší hodnotu (například $\delta = 10^{-2}$).

Závěrečné poznámky

- Předpodmiňovače založené na použití metody BFGS s omezenou pamětí nevyžadují žádné korekce. Jsou poměrně robustní, ale nejsou nejefektivnější.
- Pásové předpodmiňovače založené na použití standardní metody BFGS je třeba předem upravit, jinak jsou při provádění rozkladu většinou odvrhnuty. Velmi se osvědčily úpravy založené na větě 5, kdy se mimodiagonální prvky zmenšují tak, aby se záporné subdeterminanty vynulovaly. Ukazuje se, že takto získané předpodmiňovače je třeba častěji odvrhovat (například volbou $\delta = 10^{-2}$).

- Pásové předpodmiňovače určené numerickým derivováním stačí upravit tak, že diagonální prvky nahradíme jejich absolutními hodnotami. Pro odvrhování stačí volit $\delta = 10^{-12}$ (kromě diagonálních předpodmiňovačů, které jsou citlivější na odvrhování).
- Tridiagonální předpodmiňovače získané Lanczosovou metodou není třeba korigovat (jsou podle věty 14 pozitivně definitní). Lze je však určovat pouze v nepředpodmíněném kroku Newtonovy metody. To vyvolává řadu technických potíží (musí se upravovat iterační proces metody sdružených gradientů).

Numerické porovnání

Diferenční verze nepřesné Newtonovy metody používající různé předpodmiňovače byly testovány pomocí souboru 71 testovacích úloh s 1000 proměnnými. Výsledky testů jsou uvedeny v tabulce, která obsahuje tyto údaje:

- `NIT` – celkový počet iterací.
- `NFV` – celkový počet použitých funkčních hodnot.
- `NFG` – celkový počet použitých gradientů.
- `NCG` – celkový počet vnitřních iterací.
- `NCN` – celkový počet předpodmíněných vnějších iterací.
- `NCP` – počet problémů se zvýšenou mezí pro odvrhování.
- `Čas` – celkový čas výpočtu.

Testované metody:

- `TN` – nepředpodmíněná Newtonova metoda.
- `TNLM` – předpodmínění pomocí metody BFGS s omezenou pamětí.
- `TNVM` – pásové předpodmínění pomocí standardní metody BFGS (1–diagonální, 2–tridiagonální, 3–pentadiagonální).
- `TNND` – pásové předpodmínění pomocí numerického derivování (1–diagonální, 2–tridiagonální, 3–pentadiagonální).
- `TNLT` – Tridiagonální předpodmínění pomocí Lanczosovy metody.
- `LMVM` – Metoda BFGS s omezenou pamětí.
- `CG` – Nelineární metoda sdružených gradientů.

Metody `LMVM` a `CG` jsou uvedeny pro srovnání (nemají nic společného s Newtonovou metodou).

Metoda	NIT	NFV	NFG	NCG	NCN	NCP	Čas
TN	7425	11827	372789	359505	-	-	66.08
TNLM	7270	12521	233269	219347	7270	-	42.55
TNVM-1	7095	10303	274344	262855	4335	37	50.43
TNVM-2	6751	9252	139989	129933	4260	37	27.47
TNVM-3	6803	8857	229501	219820	4027	36	51.67
TNND-1	6522	8491	347384	331709	3857	40	59.51
TNND-2	7573	11245	147391	119434	4409	3	25.45
TNND-3	7107	10726	125262	91665	4943	4	24.57
TNLT	7398	11672	352199	339081	6808	1	55.61
LMVM	121314	127189	127189	-	-	-	39.59
CG	109166	325994	325994	-	-	-	75.72

Z údajů uvedených v této tabulce lze vyvodit několik závěrů:

- Diferenční verze Newtonovy metody konvergují velmi rychle, vyžadují však větší počet gradientů.
- Nepředpodmíněná Newtonova metoda nemůže konkurovat metodě BFGS s omezenou pamětí.
- Diagonální předpodmiňovače a předpodmiňovače získané Lanczosovou metodou nejsou příliš účinné.
- Pásové předpodmiňovače určené metodou BFGS je třeba upravovat. Často je také třeba zvýšit mez pro odvrhování.
- Pásové předpodmiňovače určené numerickým derivováním vyžadují pouze minimální korekce. Takto upravená Newtonova metoda je účinnější, než metoda BFGS s omezenou pamětí.

Literatura

[1] N.J.Higham: Accuracy and Stability of Numerical Algorithms. SIAM, Philadelphia 2002.

[2] L.Lukšan: Numerické optimalizační metody. Nepodmíněná minimalizace. Výzkumná zpráva V-1058, Ústav informatiky AV ČR, Praha 2009 (`www.cs.cas.cz/luksan/lekce4.pdf`).

[3] L.Lukšan, J.Vlček: Sparse test problems for unconstrained optimization. Výzkumná zpráva V-1064, Ústav informatiky AV ČR, Praha 2010 (`ftp.cs.cas.cz/pub/reports/v1064-10.ps`).