ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
SERIA III: MATEMATYKA STOSOWANA V (1975)

Ministerstwo Nauki i Szkolnictwa Wyższego

H. WOŹNIAKOWSKI (Warszawa)

# Metoda minimalnych $B$-błędów dla wielkich układów równań liniowych o dowolnej macierzy

**1. Wstęp.** Większość współczesnych metod iteracyjnych (np. metody T, mr, cg, SOR – por. [1], [2], [4], [5], [8]) rozwiązywania wielkich układów równań liniowych

$$A\vec{x} + \vec{b} = \vec{0}, \qquad A(n \times n), \quad \vec{b}(n \times 1), \tag{1}$$

wymaga, aby macierz $A$ była hermitowska i dodatnio określona. Szybkość zbieżności do rozwiązania $\vec{x}^*$ konstruowanego ciągu $\{\vec{x}_k\}$, jak również osiągana maksymalna graniczna dokładność, zależą przede wszystkim od wskaźnika uwarunkowania zadania (1) $\operatorname{cond}_2(A) = \|A\|_2 \|A^{-1}\|_2$. Większość zadań pojawiających się w praktyce obliczeniowej (np. numeryczne rozwiązywanie równań eliptycznych) spełnia powyższe założenia z $\operatorname{cond}_2(A) >> 1$.

Powstaje jednak problem, jak rozwiązywać wielkie układy (1) dla dowolnej nieosobliwej macierzy $A$. Często stosuje się w takiej sytuacji transformację Gaussa, polegającą na obustronnym przemnożeniu (1) przez macierz sprzężoną $A^H$ (gdy $A = (a_{ij})$, to $A^H = (\bar{a}_{ji})$). Otrzymujemy wówczas równoważny układ

$$M\vec{x} + \vec{g} = \vec{0}, \qquad \text{gdzie} \qquad M = A^H A, \quad \vec{g} = A^H \vec{b}. \tag{2}$$

Macierz $M$ jest hermitowska i dodatnio określona. Do układu (2) możemy zatem zastosować jedną z powyżej wspomnianych metod iteracyjnych. Zauważmy jednak, że

$$\operatorname{cond}_2(M) = [\operatorname{cond}_2(A)]^2,$$

czyli zadanie (2) może być znacznie gorzej uwarunkowane, niż zadanie (1). Powoduje to odpowiednie wydłużenie procesu iteracyjnego, jak również może spowodować znaczne pogorszenie osiąganej maksymalnej dokładności. O ile to możliwe, należy więc starać się tak sformułować zadanie wyjściowe, aby odpowiadająca mu macierz $A$ w (1) była hermitowska i dodatnio określona (por. [4], rozdział 1). I tak np., przy numerycznym rozwiązywaniu zadania samosprzężonego w liniowych równaniach różniczkowych, można zalecić przejście do równoważnego problemu wariacyjnego, a następnie bezpośrednio minimalizować odpowiedni funkcjonał kwadratowy. Niezależnie od sposobu dyskretyzacji zadania prowadzi to do układu liniowego o macierzy hermitowskiej i dodatnio określonej (por. [4]). Jeśli jednak nie potrafimy w ten sposób zmienić zadania (1) z zachowaniem rzędu wielkości uwarunkowania, to z konieczności musimy niekiedy rozwiązywać zadanie (2) dla macierzy $M$ szczególnej postaci $M = A^H A$.

Czy można wykorzystać ten znany rozkład macierzy $M$ na czynniki dla uzyskania bardziej efektywnej metody iteracyjnej?

Odpowiedzią pozytywną na to pytanie jest przedstawiona poniżej metoda minimalnych błędów, me (por. [3], str. 113, przypadek jednopunktowy, [2], str. 51), która jest szczególnym przypadkiem metody minimalnych $B$-błędów.

W niniejszym artykule przedstawiamy własności teoretyczne metody me($B$) oraz formułujemy jej szczegółowy algorytm. W końcowej części pracy rozważamy możliwość stosowania metod me($B$) dla wyznaczania uogólnionych rozwiązań układów liniowych (o prostokątnych lub osobliwych macierzach współczynników).

**2. Wielomiany jądrowe.** W dalszych rozważaniach będziemy korzystać z własności wielomianów jądrowych. Przypomnimy teraz ich definicję i niektóre własności (por. [5]). Niech $\{p_k\}$ będzie ciągiem rzeczywistych wielomianów ortogonalnych względem iloczynu skalarnego z wagą $\rho$, tzn. stopień $p_k(\lambda)$ jest równy $k$ oraz

$$(3) \qquad (p_i, p_j) \overset{\text{df}}{=} \int_a^b \rho(\lambda)\, p_i(\lambda) p_j(\lambda)\, d\lambda = 0 \qquad \text{dla} \qquad i \neq j.$$

Wielomiany $p_i$ spełniają formułę trójczłonową postaci

$$(4) \qquad \begin{aligned} &p_0(\lambda) = 1, \\ &p_{k+1}(\lambda) = \lambda\, p_k(\lambda) - \alpha_{k+1} p_k(\lambda) - \beta_k\, p_{k-1}(\lambda), \end{aligned}$$

gdzie

$$(4') \qquad \alpha_{k+1} = \frac{(\lambda p_k, p_k)}{(p_k, p_k)}, \qquad \beta_0 = 0, \qquad \beta_k = \frac{(\lambda p_k, p_{k-1})}{(p_{k-1}, p_{k-1})}, \qquad k = 1, 2, \ldots$$

Formuła (4) jest oczywiście słuszna tylko wtedy, gdy $\| p_k \| = \sqrt{(p_k, p_k)} \neq 0$. Niezerowość normy zależy od przyjętej w (3) wagi $\rho$. Jeśli $\rho$ jest w $[a, b]$ funkcją przedziałami ciągłą, nieujemną, co najmniej w jednym podprzedziale dodatnią, to dla każdego $k$, $\| p_k \| \neq 0$ i ciąg wielomianów ortogonalnych jest nieskończony. Gdy jednak przyjmiemy wagę postaci

$$(5) \qquad \rho(\lambda) = \sum_{p=1}^{m} c_p^2\, \delta(\lambda - \lambda_p),$$

gdzie $c_p \neq 0$, $\lambda_p \in [a, b]$ dla $p = 1, 2, \ldots, m$, a $\delta(\mu)$ oznacza uogólnioną funkcję Diraca, to iloczyn skalarny (3) sprowadza się do postaci sumy skończonej:

$$(6) \qquad (p_i, p_j) = \sum_{p=1}^{m} c_p^2 p_i(\lambda_p) p_j(\lambda_p).$$

Dla wagi $\rho$ typu (5) ciąg wielomianów $\{p_k\}$ jest skończony i kończy się na wielomianie $p_m$ o własności

(7) $$p_m(\lambda_p) = 0 \qquad \text{dla} \qquad p = 1, 2, \ldots, m, \cdot$$

(czyli $\| p_m \| = 0$).

Mając określone wielomiany ortogonalne $p_k$, definiujemy wielomiany jądrowe.

DEFINICJA. Niech $\lambda_0$ będzie liczbą rzeczywistą. Wielomian zmiennej $\lambda$,

$$K_i(\lambda_0, \lambda) = \sum_{k=0}^{i} \frac{p_k(\lambda_0)\, p_k(\lambda)}{(p_k, p_k)}$$

nazywamy *i-tym wielomianem jądrowym względem wagi* $\rho$. (W istocie $K_i$ jest wielomianem symetrycznym dwu zmiennych $\lambda_0, \lambda_j$; dla naszych celów będziemy jednak zawsze traktowali $\lambda_0$ jako ustalony parametr).

Przytaczamy własności wielomianów jądrowych, z których będziemy korzystali w dalszych rozważaniach.

(i) Jeśli $\lambda_0 \notin [a, b]$, to ciąg wielomianów jądrowych $\{K_i(\lambda_0, \cdot)\}_{i=0,1,\ldots}$ jest ciągiem ortogonalnym w sensie (3) z wagą $|\lambda - \lambda_0| \rho(\lambda)$.

(ii) Niech $P_i(\lambda_0, c)$ oznacza zbiór wszystkich wielomianów $\Pi$ stopnia $\leqslant i$, dla których $\Pi(\lambda_0) = c$. Wówczas jedynym w tym zbiorze rozwiązaniem warunku

(8) $$\inf_{\Pi \in P_i(\lambda_0, c)} \|\Pi\| = \|\Pi_i^*\|, \qquad \Pi_i^* \in P_i(\lambda_0, c),$$

jest wielomian

(9) $$\Pi_i^*(\lambda) = \frac{c}{K_i(\lambda_0, \lambda_0)} K_i(\lambda_0, \lambda).$$

Własność (i) pozwala konstruować wielomiany jądrowe $K_i$ za pomocą formuły trójczłonowej analogicznej do (4).

Własność (ii) oznacza, że spośród wszystkich wielomianów ograniczonego stopnia, przyjmujących w określonym punkcie określoną niezerową wartość, wielomian jądrowy najwyższego stopnia (pomnożony przez odpowiednią stałą) ma najmniejszą normę.

**3. Metody przybliżeń wielomianowych.** Rozważamy układ równań liniowych

$$M\vec{x} + \vec{g} = \vec{0}, \qquad M = M^H, \qquad M > 0$$

gdzie z (1) $M = A, \vec{g} = \vec{b}$, gdy $A = A^H > 0$, a $M = A^H A, \vec{g} = A^H \vec{b}$ w przeciwnym przypadku.

Analogicznie jak w [2] zakładamy, że informacja o macierzy układu jest dana poprzez procedurę, która dla danego elementu $\vec{x}$ przestrzeni wyznacza element $\vec{y} = M\vec{x}$.

Przy tym założeniu możemy konstruować jedynie wielomiany od macierzy $M$ na danym elemencie. Dlatego będziemy przybliżać rozwiązanie $\vec{x}^* = -M^{-1}\vec{b}$, konstruując ciąg $\{\vec{x}_k\}$ taki, że

(10) $$\vec{x}_k - \vec{x}^* = W_k(M)(\vec{x}_0 - \vec{x}^*),$$

gdzie $W_k$ jest wielomianem stopnia $\leqslant k$, $\vec{x}_0$ jest dowolnym przybliżeniem początkowym. Ponieważ nie znamy elementu $\vec{x}^*$, lecz tylko jego obraz $-\vec{g} = M\vec{x}^*$, więc dla zapewnienia efektywnej konstrukcji elementu $\vec{x}_k$ spełniającego warunek (10), musimy przyjąć (por. [2], str. 50) dodatkowy warunek, ograniczający wybór wielomianu $W_k$:

$$W_k(0) = 1. \tag{11}$$

Jeśli określimy ciąg wielomianów $\{W_k(\lambda)\}$ spełniających (11), to będziemy mówili, że metoda konstruująca ciąg elementów $\{\vec{x}_k\}$, spełniający równość (10), jest generowana przez ten ciąg wielomianów. Spróbujemy odpowiedzieć na pytanie: jak można racjonalnie wybrać wielomiany $W_k$?

Wydaje się celowe określić wielomiany $W_k$ tak, aby zminimalizować pewną normę błędu $\vec{x}_k - \vec{x}^*$.

Niech

$$\|\vec{x}\|_B \overset{\text{df}}{=} \sqrt{(B\vec{x}, \vec{x})}, \tag{12}$$

gdzie $B$ oznacza pewną, odpowiednio dobraną macierz hermitowską, dodatnio określoną $B = B^H > 0$.

Wielomiany $W_k \in P_k(0, 1)$, minimalizujące błąd $\vec{x}_k - \vec{x}^*$ w normie (12) spełniają zatem warunek

$$\|W_k(M)(\vec{x}_0 - \vec{x}^*)\|_B = \inf_{\mathcal{T} \in P_k(0,1)} \|\mathcal{T}(M)(\vec{x}_0 - \vec{x}^*)\|_B, \tag{13}$$

gdzie, jak poprzednio, $P_k(0, 1)$ oznacza klasę wielomianów stopnia $\leqslant k$, przyjmujących w zerze wartość jeden.

DEFINICJA. Konstrukcję ciągu $\{\vec{x}_k\}$ w myśl (10), gdzie wielomiany $W_k$ spełniają (11) i (13), nazywamy *metodą minimalnych B-błędów*, w skrócie metodą me($B$).

Zajmijmy się rozwiązaniem warunku (13). W tym celu początkowy wektor błędu, $\vec{x}_0 - \vec{x}^*$, przedstawmy w bazie wektorów własnych macierzy $M$.

Dokładniej, niech

$$\vec{x}_0 - \vec{x}^* = \sum_{j=1}^{m} \vec{\xi}_j c_j, \qquad m \leqslant n, \qquad c_j \neq 0, \tag{14}$$

gdzie

$$M\vec{\xi}_j = \vec{\xi}_j \lambda_j, \qquad \|\vec{\xi}_j\|_B = 1, \qquad 0 < \lambda_1 < \lambda_2 < \ldots < \lambda_m.$$

Zauważmy, że $\lambda_m \leqslant \|M\|_2$, $\lambda_1^{-1} \leqslant \|M^{-1}\|_2$.

Związek (10) możemy teraz zapisać w postaci:

$$\vec{x}_k - \vec{x}^* = \sum_{j=1}^{m} W_k(M)\vec{\xi}_j c_j = \sum_{j=1}^{m} \vec{\xi}_j W_k(\lambda_j) c_j. \tag{15}$$

Załóżmy dodatkowo, że macierz $B$, definiująca normę (12) jest tak dobrana, że

$$B\vec{\xi}_j = \beta_j \vec{\xi}_j, \qquad \beta_j > 0, \qquad j = 1, 2, \ldots, m. \tag{16}$$

Założenie (16) jest spełnione np. wtedy, gdy $B = M^p$, $\beta_j = \lambda_j^p$, z dowolnym rzeczywistym $p$ (lub ogólnie, dla dowolnej macierzy $B$, przemiennej z $M$).

Z (14), (15) i (16) wynika wówczas

$$\|\vec{x}_k - \vec{x}^*\|_B = \sqrt{\sum_{j=1}^{m} |c_j|^2 \beta_j |W_k(\lambda_j)|^2}. \tag{17}$$

Ostatnia zależność wskazuje na ścisły związek poszukiwanych wielomianów $W_k$ z wielomianami jądrowymi. Rzeczywiście przyjmując (por. (5)):

$$\rho(\lambda) = \sum_{j=1}^{m} |c_j|^2 \beta_j \delta(\lambda - \lambda_j) \tag{18}$$

i wykorzystując zależności (9) i (11), dochodzimy do wniosku, że rozwiązanie problemu (13) wyraża się wzorem

$$W_k(\lambda) = \frac{K_k(0, \lambda)}{K_k(0, 0)}, \tag{19}$$

gdzie $K_k(0, \lambda)$ oznacza $k$-ty wielomian jądrowy względem wagi (18) przy $\lambda_0 = 0$ (por. [5], str. 8).

Otrzymaliśmy więc

TWIERDZENIE 1. *Metoda minimalnych B-błędów dla macierzy B, przemiennej z macierzą układu, jest generowana przez unormowane wielomiany jądrowe* (19), *względem wagi* (18). ■

Powyższe twierdzenie jest wariantem twierdzenia 6 zawartego w [5], str. 8.

Dla szczególnych macierzy $B$ metoda me($B$) sprowadza się do różnych znanych metod. I tak przyjmując szczególne macierze $B$ otrzymujemy:

$1^o$ dla $B = M^0 = I$ metoda minimalnych błędów (me) (por. [2], str. 51, [3], str. 113 – przypadek jednopunktowy, tzn. $k = 1$). Norma (12) jest teraz równa normie euklidesowej. Wybór wielomianów jądrowych (19) zapewnia minimalizację normy euklidesowej błędu. W punkcie 6 pokażemy, że praktyczne stosowanie metody me wymaga szczególnej postaci macierzy $M$, np. w postaci iloczynu sprzężonych czynników $M = A^H A$.

$2^o$ dla $B = M$ – metoda sprzężonych gradientów (cg) (por. [4], [5]).

$3^o$ dla $B = M^2$ – metoda minimalnych residuów (mr) (por. [2] oraz [5], gdzie E. Stiefel nazywa metodę mr($M^2$) zmodyfikowaną metodą sprzężonych gradientów).

Warunek (13) oznacza teraz minimalizację wektorów residualnych

$$\vec{r}_k \overset{\mathrm{df}}{=} M\vec{x}_k + \vec{g}$$

w normie euklidesowej.

**4. Charakter zbieżności metody minimalnych $B$-błędów.** Rozpatrywana metoda konstruuje ciąg $\{\vec{x}_k\}$ w myśl reguły

$$\vec{x}_k - \vec{x}^* = W_k(M)(\vec{x}_0 - \vec{x}^*), \qquad W_k(M) = K_k(0, M)/K_k(0, 0) \tag{20}$$

dla ciągu wielomianów jądrowych $\{K_k\}$ względem wagi (18).

TWIERDZENIE 2. *Niech zachodzi równość* (14) *oraz niech*

$$\sigma = \frac{\sqrt{\lambda_m} - \sqrt{\lambda_1}}{\sqrt{\lambda_m} + \sqrt{\lambda_1}} \qquad (\lambda_m \leqslant \|M\|_2, \lambda_1^{-1} \leqslant \|M^{-1}\|_2).$$

*Ciągi błędów* $\{\vec{e}_k\}$, $\vec{e}_k = \vec{x}_k - \vec{x}^*$ *oraz residuów* $\{\vec{r}_k\}$, $\vec{r}_k = M\vec{x}_k + \vec{g}$ *odpowiadające ciągowi* $\{\vec{x}_k\}$, *konstruowanemu w metodzie minimalnych B-błędów, spełniają zależności*:

1. *dla* $k < m$

$$\|\vec{e}_k\|_B \leqslant 2\sigma^k \|\vec{e}_0\|_B \leqslant 2\sigma^k \|B^{1/2}M^{-1}\|_2 \|\vec{r}_0\|_2, \tag{21}$$

$$\|\vec{e}_k\|_2 \leqslant 2\sigma^k \|B^{-1/2}\|_2 \|\vec{e}_0\|_B \leqslant 2\sigma^k \|B^{-1/2}\|_2 \|B^{1/2}M^{-1}\|_2 \|\vec{r}_0\|_2, \tag{22}$$

$$\|\vec{r}_k\|_B \leqslant 2\sigma^k \lambda_m \|\vec{e}_0\|_B \leqslant 2\sigma^k \lambda_m \|B^{1/2}M^{-1}\|_2 \|\vec{r}_0\|_2, \tag{23}$$

$$\|\vec{r}_k\|_2 \leqslant 2\sigma^k \|B^{-1/2}M\|_2 \|\vec{e}_0\|_B \leqslant 2\sigma^k \operatorname{cond}_2(B^{-1/2}M) \|\vec{r}_0\|_2; \tag{24}$$

2. *dla* $k \geqslant m$

$$\vec{e}_k = \vec{r}_k = \vec{0} \qquad (tzn.\ \vec{x}_k = \vec{x}^*). \tag{25}$$

U w a g i. Nierówności (21)–(24) wyrażają oszacowanie błędów i residuów w normach $\|\cdot\|_B$ oraz $\|\cdot\|_2$ dla początkowych kroków iteracyjnych, $k \leqslant m$. Zależność (25) oznacza, że metoda me($B$) jest skończona i dokładnie po $m$ krokach, $m \leqslant n$, otrzymujemy rozwiązanie. Własność ta jest niestety tylko własnością teoretyczną. Przy numerycznej realizacji metody me($B$) nie możemy na ogół liczyć nawet na przybliżone spełnienie (25) (por. punkt 7), a zależności (21)–(24) spełnione są w przybliżeniu, i to tylko dla dostatecznie małych $k$.

D o w ó d. Niech $k < m$. Z (13) wynika, że dla każdego $\pi \in P_k(0, 1)$ mamy

$$\|\vec{e}_k\|_B \leqslant \|\pi(M)\vec{e}_0\|_B = \|\pi(M)B^{1/2}\vec{e}_0\|_2 \leqslant \max_{\lambda_1 \leqslant \lambda \leqslant \lambda_m} |\pi(\lambda)| \, \|\vec{e}_0\|_B.$$

Wybierzmy wielomian $\pi_k^* \in P_k(0, 1)$ tak, aby

$$\max_{\lambda_1 \leqslant \lambda \leqslant \lambda_m} |\pi_k^*(\lambda)| = \inf_{\pi \in P_k(0,1)} \max_{\lambda_1 \leqslant \lambda \leqslant \lambda_m} |\pi(\lambda)|. \tag{26}$$

Rozwiązaniem warunku (26) jest (por. np. [1]) wielomian

$$\pi_k^*(\lambda) = \frac{T_k(f(\lambda))}{T_k(f(0))}, \tag{27}$$

gdzie $T_k$ oznacza $k$-ty wielomian Czebyszewa pierwszego rodzaju, zaś

$$f(\lambda) = \frac{\lambda_m + \lambda_1}{\lambda_m - \lambda_1} - \frac{2}{\lambda_m - \lambda_1}\,\lambda.$$

U w a g a. Metoda, w której wielomiany $W_k$ w (10) są określone przez (27), jest nazywana metodą Czebyszewa (T) (por. [1], [2], [4] i [5]). Z własności wielomianów Czebyszewa wynika

$$\max_{\lambda_1 \leqslant \lambda \leqslant \lambda_m} |\pi_k^*(\lambda)| \leqslant 2\,\sigma^k,$$

dla wartości $\sigma$ zdefiniowanej w twierdzeniu 2. Stąd i z faktu, że $\pi^*$ minimalizuje $B$-normę w $P_k(0, 1)$, otrzymujemy oszacowanie:

$$\|\vec{e}_k\|_B \leqslant 2\sigma^k \|\vec{e}_0\|_B = 2\sigma^k \|B^{1/2}M^{-1}\vec{r}_0\|_2 \leqslant 2\sigma^k \|B^{1/2}M^{-1}\|_2 \|\vec{r}_0\|_2,$$

co stanowi dowód (21).

Zależności (22)–(24) wynikają z (21) oraz z faktu równoważności norm ($B$-normy i normy euklidesowej)

$$\|B^{-1/2}\|_2^{-1} \|\vec{x}\|_2 \leqslant \|\vec{x}\|_B \leqslant \|B^{1/2}\|_2 \|\vec{x}\|_2.$$

Załóżmy teraz, że $k \geqslant m$. Ponieważ $0 \notin [\lambda_1, \lambda_m]$, z własności (i) wielomianów jądrowych wynika, że ciąg $\{W_k\}$ z (20) jest ciągiem wielomianów ortogonalnych względem wagi

$$\rho_1(\lambda) = \sum_{j=1}^{m} |c_j|^2 \beta_j \lambda_j \delta(\lambda - \lambda_j). \tag{28}$$

Waga (28) jest wagą typu (5). Na mocy (7) mamy

$$W_k(\lambda_p) = 0 \tag{29}$$

dla $p = 1, 2, \ldots, m$. Jest to równoważne zależności

$$W_k(M)\,\vec{e}_0 = \vec{0},$$

czyli z (20) otrzymujemy

$$\vec{x}_k = \vec{x}^*, \qquad \vec{e}_k = \vec{r}_k = \vec{0},$$

co kończy dowód. ■

Omówimy twierdzenie 2 dla szczególnych macierzy $B$.

$1^\circ$ W metodzie me, $B = I$, mamy więc

$$\|\vec{e}_k\|_2 \leqslant 2\sigma^k \|e_0\|_2 \leqslant 2\sigma^k \|M^{-1}\|_2 \|\vec{r}_0\|_2,$$

$$\|\vec{r}_k\|_2 \leqslant 2\sigma^k \operatorname{cond}_2(M) \|\vec{r}_0\|_2.$$

$2^\circ$ W metodzie cg, $B = M$,

$$\|\vec{e}_k\|_2 \leqslant 2\sigma^k \|M^{-1/2}\|_2 \|M^{1/2}\vec{e}_0\|_2 \leqslant 2\sigma^k \|M^{-1}\|_2 \|\vec{r}_0\|_2,$$

$$\|\vec{r}_k\|_2 \leqslant 2\sigma^k \operatorname{cond}_2(M^{1/2}) \|\vec{r}_0\|_2$$

(por. [2], str. 53).

$3^\sigma$ W metodzie mr, $B = M^2$,

$$\|\vec{e}_k\|_2 \leqslant 2\sigma^k \|M^{-1}\|_2 \|\vec{r}_0\|_2,$$

$$\|\vec{r}_k\|_2 \leqslant 2\sigma^k \|\vec{r}_0\|_2,$$

(por. [2], str. 53).

**5. Charakter zbieżności** w przypadku $M = A^H A$. Z twierdzenia 2 wynika, że wielkością charakteryzującą szybkość zbieżności jest $\sigma = \sigma(M, \vec{e}_0)$,

$$\sigma = \frac{\sqrt{\lambda_m/\lambda_1} - 1}{\sqrt{\lambda_m/\lambda_1} + 1},$$

przy czym $\lambda_m/\lambda_1 \leqslant \operatorname{cond}_2(M) = [\operatorname{cond}_2(A)]^2$. Należy podkreślić, że wartości własne $\lambda_1, \lambda_m$ macierzy $M$, występujące w rozłożeniu spektralnym (14) początkowego wektora błędu, są zazwyczaj najmniejszą i największą wartościami własnymi tej macierzy. A więc

$$\sigma = \frac{\sqrt{\operatorname{cond}_2(M)} - 1}{\sqrt{\operatorname{cond}_2(M)} + 1} = \frac{\operatorname{cond}_2(A) - 1}{\operatorname{cond}_2(A) + 1}. \tag{30}$$

Gdyby było $A = A^H > 0$, to przyjmując $M = A$ mielibyśmy na ogół proces znacznie szybciej zbieżny.

Widzimy zatem, że transformacja Gaussa jak gdyby znacznie pogarsza szybkość zbieżności procesu przybliżonego.

Pokażemy teraz, że pogorszenie zbieżności wiąże się przede wszystkim z nieokreślonością macierzy $A$, a nie jak wydawałoby się, z transformacją Gaussa.

Rozpatrzmy układ (1),

$$A\vec{x} + \vec{b} = \vec{0},$$

z dodatkowym założeniem o macierzy $A$:

$$A = A^H, \qquad \operatorname{spectr}(A) \subset [-b, -a] \cup [a, b], \tag{31}$$

gdzie $b = \|A\|_2$, $a = 1/\|A^{-1}\|_2$.

Do powyższego układu nie stosujemy teraz transformacji Gaussa, lecz konstruujemy ciąg $\{\vec{x}_k\}$ w myśl reguły:

$$\vec{x}_k - \vec{x}^* = W_k(A)(\vec{x}_0 - \vec{x}^*), \qquad W_k \in P_k(0, 1). \tag{32}$$

Z (32) wynika

$$\|\vec{x}_k - \vec{x}^*\|_2 \leqslant \|W_k(A)\|_2 \|\vec{x}_0 - \vec{x}^*\|_2,$$

gdzie, na mocy (31),

$$\| W_k(A) \|_2 \leqslant \| W_k \| \overset{\mathrm{df}}{=} \max_{a \leqslant |\lambda| \leqslant b} |W_k(\lambda)| \qquad (\lambda \text{ rzeczywiste}).$$

Wielomiany $W_k$ wybierzmy tak, aby zminimalizować normę $\| W_k \|$, tzn.

$$\| W_k \| = \inf_{\mathcal{T} \in P_k(0,1)} \| \mathcal{T} \|. \tag{33}$$

Powyższa metoda jest wariantem metody Czebyszewa dla macierzy hermitowskich, nieosobliwych i nieokreślonych.

Łatwo sprawdzić, że rozwiązaniem (33) dla parzystych wskaźników jest wielomian,

$$W_{2k}(\lambda) = \frac{T_k(g(\lambda))}{T_k(g(0))}, \tag{34}$$

gdzie, tak jak poprzednio, $T_k$ jest $k$-tym wielomianem Czebyszewa, a

$$g(\lambda) = \frac{b^2 + a^2}{b^2 - a^2} - \frac{2}{b^2 - a^2} \lambda^2.$$

Z (34) oraz z faktu, że $b/a = \mathrm{cond}_2(A)$, otrzymujemy

$$\| W_{2k} \| = \frac{1}{|T_k(g(0))|} \cong 2\left(\frac{\mathrm{cond}_2(A) - 1}{\mathrm{cond}_2(A) + 1}\right)^k.$$

Ostatecznie konstruowany ciąg $\{\vec{x}_k\}$ spełnia zależność

$$\| \vec{x}_{2k} - \vec{x}^* \|_2 \leqslant 2\sigma^k \| \vec{x}_0 - \vec{x}^* \|_2.$$

Proces przybliżający (32), (33), optymalny w klasie macierzy $A$ spełniających (31), ma więc ten sam charakter zbieżności, co metoda me($B$), zastosowana do układu o macierzy $A^H A$.

Nieokreśloność macierzy $A$ może więc wiązać się ze znacznym pogorszeniem szybkości zbieżności optymalnych procesów przybliżonych, bez względu na to, czy stosujemy transformację Gaussa, czy też nie.

Podkreślamy zatem raz jeszcze: Rozwiązywanie wielkich układów równań liniowych o niedodatnio określonej macierzy jest na ogół zadaniem znacznie bardziej pracochłonnym, niż zadanie z dodatnio określoną macierzą. Należy zatem zawsze, o ile to tylko możliwe, dążyć do formułowania zadania wyjściowego tak, aby odpowiadająca mu macierz $A$ była hermitowska, dodatnio określona.

**6. Algorytm metody minimalnych $B$-błędów.** W dowodzie twierdzenia 2 stwierdziliśmy, że wielomiany $\{W_k\}$ z wzoru (20) są wielomianami ortogonalnymi względem wagi (28). Jako wielomiany ortogonalne, spełniają formułę trójczłonową typu (4), którą dla wygody zapiszemy teraz w postaci (por. [5], str. 10)

$$W_{k+1}(\lambda) = W_k(\lambda) + \frac{1}{q_k}[e_{k-1}(W_k(\lambda) - W_{k-1}(\lambda)) - \lambda W_k(\lambda)]. \tag{35}$$

Przy warunkach początkowych $W_0(0) = 1$ oraz $e_{-1} = 0$, postać (35) zapewnia spełnienie warunku $W_k(0) = 1$ dla $k = 0, 1, \ldots$ Z ortogonalności $\{W_k\}$ wynikają następujące równości na $e_{k-1}$ i $q_k$,

$$q_k = \frac{(\lambda W_k, W_k)}{(W_k, W_k)} - e_{k-1}, \qquad e_k = \frac{(W_{k+1}, W_{k+1})}{(W_k, W_k)} q_k.$$

Iloczyny skalarne $(W_k, W_k)$ i $(\lambda W_k, W_k)$, na mocy (6), (15) i (28), są równe

$$(W_k, W_k) = \sum_{j=1}^{m} |c_j|^2 \beta_j \lambda_j W_k^2(\lambda_j) = (\vec{r}_k, B(\vec{x}_k - \vec{x}^*)),$$

$$(\lambda W_k, W_k) = \sum_{j=1}^{m} |c_j|^2 \beta_j \lambda_j^2 W_k^2(\lambda_j) = (\vec{r}_k, B\vec{r}_k),$$

gdzie $\vec{r}_k = M\vec{x}_k + \vec{g}$.

Podstawiając w (35) na miejsce $\lambda$ macierz $M$ i działając otrzymanym operatorem na wektor $\vec{x}_0 - \vec{x}^*$, uwzględniając wreszcie (10), otrzymujemy:

Algorytm metody me$B$

$$\vec{x}_{k+1} = \vec{x}_k + \frac{1}{q_k}\{e_{k-1}(\vec{x}_k - \vec{x}_{k-1}) - \vec{r}_k\}$$

$$\vec{r}_k = M\vec{x}_k + \vec{g},$$

$$q_k = \frac{(\vec{r}_k, B\vec{r}_k)}{(\vec{r}_k, B(\vec{x}_k - \vec{x}^*))} - e_{k-1},$$

$$e_{-1} = 0, \quad e_k = \frac{(\vec{r}_{k+1}, B(\vec{x}_{k+1} - \vec{x}^*))}{(\vec{r}_k, B(\vec{x}_k - \vec{x}^*))} q_k. \quad \blacksquare$$

Zauważmy, że warunkiem koniecznym stosowalności metody me$(B)$ jest umiejętność obliczenia współczynników

$$c_k \equiv (\vec{r}_k, B(\vec{x}_k - \vec{x}^*)) = (\vec{x}_k - \vec{x}^*, MB(\vec{x}_k - \vec{x}^*)). \tag{36}$$

Współczynniki te potrafimy zawsze obliczyć jeśli $B = M^p$ dla $p$ naturalnego, gdyż wówczas

$$c_k = (\vec{r}_k, M^{p-1}\vec{r}_k).$$

Tak więc metody cg $(p = 1)$ oraz mr $(p = 2)$ prowadzą do algorytmów realizowalnych bez dodatkowych założeń o macierzy $M$, i o jej widmie.

Przejdźmy do metody me, $B = I$. Na ogół nie potrafimy obliczać współczynników $c_k$, a tym samym nie możemy stosować metody me do dowolnego układu o macierzy $M = M^H > 0$. Jeśli jednak układ pochodzi z transformacji Gaussa, $M = A^H A$, to na mocy (1) i (2) wielkość

$$c_k = (\vec{x}_k - \vec{x}^*, A^H A(\vec{x}_k - \vec{x}^*)) = \|A\vec{x}_k + \vec{b}\|_2^2$$

może być efektywnie obliczana. W tym przypadku jest więc możliwa realizacja metody me (por. [7]).

**7. Stabilność metody minimalnych *B*-błędów.** Niezbędnym warunkiem stosowalności dowolnej metody w praktyce obliczeniowej jest jej numeryczna stabilność, tzn. własność zapewniająca, że błąd wytworzony przy numerycznej realizacji algorytmu jest porównywalny z przeniesionym błędem reprezentacji danych początkowych. (Błąd ten w głównej mierze zależy od uwarunkowania zadania). Kwestia numerycznej stabilności metody minimalnych *B*-błędów nie jest w pełni wyjaśniona. Wiemy z przeprowadzonych testów, że metody te ($B = M^p$, $p = 0, 1, 2$) nie są stabilne przynajmniej w klasycznym sensie tego pojęcia. Często się zdarza, szczególnie dla zadań źle uwarunkowanych, że po wykonaniu $m$ kroków, kiedy teoretycznie powinniśmy osiągnąć prawdziwe rozwiązanie, otrzymujemy przybliżenie $\vec{x}_m$, dla którego $\| \vec{x}_m - \vec{x}^* \|$ jest tylko nieznacznie mniejsze od $\| \vec{x}_0 - \vec{x}^* \|$. Z drugiej jednak strony, dla szczególnie wybranych wektorów $\vec{x}_0$ i dla pewnych rozkładów wartości własnych macierzy $M$, metody me($B$) dają zaskakująco dobre przybliżenia.

Pytanie, co należy założyć o macierzy $M$ i wektorze $\vec{x}_0$, aby zagwarantować numeryczną stabilność postępowania jest nadal problemem otwartym. Z uwagi na numeryczną niestabilność metody me($B$) stosowanie jej jako metody samodzielnej nie może być generalnie zalecane.

Istnieje jednak możliwość połączenia metody me($B$) ze stabilną metodą Czebyszewa (T) tak, aby wykorzystać pozytywne własności obu metod i zapewnić numeryczną stabilność całego procesu obliczeniowego. Problem łączenia tych metod był poruszany w pracach [4] i [5]. Prace [2] i [7] zawierają próbę pełnego zalgorytmizowania metod mr-T, me-T odpowiednio.

**8. Uogólnienie na przypadek macierzy osobliwych i prostokątnych.** Rozpatrzmy problem znalezienia uogólnionego rozwiązania układu

$$A\vec{x} + \vec{b} = \vec{0}, \tag{37}$$

gdzie dane są: macierz $A$ $(s \times n)$ i wektor $\vec{b}$ $(s \times 1)$. Rozwiązanie uogólnione ([0], str. 37) definiujemy jako wektor $\vec{x}^*$ $(n \times 1)$, o najmniejszej normie euklidesowej, spośród wektorów $\vec{x}$ minimalizujących $\| A\vec{x} + \vec{b} \|_2$.

Rozwiązanie uogólnione $\vec{x}^*$, spełnia równanie

$$A^H (A\vec{x} + \vec{b}) = \vec{0}. \tag{38}$$

Jeśli macierz $M = A^H A$, $M$ $(n \times n)$, jest nieosobliwa, to (37) sprowadza się do nieosobliwego układu

$$M\vec{x} + \vec{g} = \vec{0}, \qquad \vec{g} = A^H \vec{b}. \tag{39}$$

Układ ten możemy rozwiązać metodą iteracyjną startując z dowolnego przybliżenia początkowego $\vec{x}_0$.

Załóżmy teraz, że macierz $M$ jest osobliwa. Niech $\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_n$ będą wektorami własnymi macierzy $M$, tworzącymi bazę ortonormalną przestrzeni zespolonej $C^n$. Niech

$$M\vec{\xi}_i = \vec{\xi}_i \lambda_i, \qquad 0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_p, \quad \lambda_{p+1} = \lambda_{p+2} = \ldots = \lambda_n = 0.$$

Rozłóżmy $C^n$ na sumę prostą,

$$C^n = C_1 \oplus C_2,$$

gdzie

$$C_1 = \operatorname{lin}(\vec{\xi}_1, \vec{\xi}_2, \ldots, \vec{\xi}_p), \qquad C_2 = \operatorname{lin}(\vec{\xi}_{p+1}, \vec{\xi}_{p+2}, \ldots, \vec{\xi}_n).$$

Wówczas każdy wektor $\vec{x}$ możemy przedstawić w postaci $\vec{x} = \vec{x}_1 + \vec{x}_2$, $\vec{x}_i \in C_i$, $i = 1, 2$.

Zauważmy, że „najkrótsze" rozwiązanie i wyraz wolny z (39) spełniają warunki,

$$\vec{x}^* = \vec{x}_1^*, \qquad \vec{g} = \vec{g}_1.$$

Powróćmy do ciągu $\{\vec{x}_k\}$ z (10); uwzględniając powyższe relacje otrzymujemy:

$$\vec{x}_k - \vec{x}^* = W_k(M)(\vec{x}_0 - \vec{x}^*) = W_k(M)(\vec{x}_{0,1} - \vec{x}_1^*) + W_k(M)\vec{x}_{0,2}.$$

Z warunku $W_k(0) = 1$ wynika,

$$W_k(M)\vec{x}_{0,2} = \vec{x}_{0,2},$$

a stąd

$$\|\vec{x}_k - \vec{x}^*\|_B^2 = \|\vec{x}_{0,2}\|_B^2 + \|W_k(M)(\vec{x}_{0,1} - \vec{x}^*)\|_B^2 \tag{40}$$

dla dowolnej hermitowskiej macierzy $B$, spełniającej warunki (16). Z (40) wynika, że warunkiem koniecznym zbieżności ciągu $\vec{x}_k$ do $x^*$ jest $\vec{x}_{0,2} = \vec{0}$, co możemy zawsze zapewnić, przyjmując $\vec{x}_0 = \vec{0}$.

Niech

$$\vec{x}^* = \sum_{j=1}^{m} \vec{\xi}_j c_j \qquad \text{dla} \qquad c_j \neq 0,\ M\vec{\xi}_j = \vec{\xi}_j \lambda_j,\ 0 < \lambda_1 < \lambda_2 < \ldots < \lambda_m,\ m \leqslant p \leqslant n.$$

Dla $\vec{x}_0 = \vec{0}$ otrzymujemy z (40)

$$\vec{x}_k \in C_1, \qquad \|\vec{x}_k - \vec{x}^*\|_B^2 = \sum_{j=1}^{m} |c_j|^2 \beta_j W_k^2(\lambda_j). \tag{41}$$

Widzimy zatem, że metoda me($B$) ma sens nawet w przypadku osobliwej macierzy $M$ o ile tylko $\vec{x}_0 = \vec{0}$.

Jako macierz $B$ możemy przyjąć

$$B = M^p + P_2,$$

gdzie $p \geqslant 0$, a $P_2$ oznacza macierz taką, że

$$P_2 \vec{x} = \vec{x}_2.$$

Charakter zbieżności jest analogiczny jak w twierdzeniu 2.

Zauważmy w końcu, że dla metody me współczynniki $c_k$ z (36) są równe

$$c_k = \|A\vec{x}_k - A\vec{x}^*\|_2^2.$$

Potrafimy je obliczać tylko wtedy, gdy $A\vec{x}^* = -\vec{b}$, co oznacza, że układ (37) nie może być sprzeczny.

Udowodniliśmy zatem

TWIERDZENIE 3. *Niech $\vec{x}^*$ będzie uogólnionym rozwiązaniem układu*

$$M\vec{x} + \vec{g} = \vec{0}, \qquad M = A^H A, \qquad \vec{g} = A^H \vec{b}.$$

$1^\sigma$ *Metoda* me$(B)$ *dla* $B = M^p + P_2$, $p \geqslant 0$, *konstruuje ciąg* $\{x_k\}$ *zbieżny do* $\vec{x}^*$ *dla przybliżenia początkowego* $\vec{x}_0$ *takiego, że*

$$\vec{x}_0 = \begin{cases} \text{dowolne}, & \text{gdy } M \text{ nieosobliwa}, \\ \vec{0}, & \text{gdy } M \text{ osobliwa}. \end{cases}$$

*Niech*

$$\vec{x}_0 - \vec{x}^* = \sum_{j=1}^{m} \vec{\xi}_j c_j \in \boldsymbol{C}_1, \qquad c_j \neq 0, \qquad M\vec{\xi}_j = \vec{\xi}_j \lambda_j, \qquad 0 < \lambda_1 < \lambda_2 < \ldots < \lambda_m.$$

$2^\sigma$ *Konstruowany ciąg* $\{\vec{x}_k\}$ *spełnia zależności:*

(a) $\vec{x}_k - \vec{x}^* \in \boldsymbol{C}_1$, $\quad \|\vec{x}_k - \vec{x}^*\|_2 \leqslant 2 d_1 \sigma^k \|\vec{r}_0\|_2$, $\quad \|\vec{r}_k\|_2 \leqslant 2 d_2 \sigma^k \|\vec{r}_0\|_2$,

*gdzie*

$$d_2 = (\lambda_m/\lambda_1)^{|1-p/2|}, \qquad d_1 = \begin{cases} 1/\lambda_1, & p \leqslant 2, \\ d_2/\lambda_1, & p \geqslant 2, \end{cases}$$

$$\sigma = \frac{\sqrt{\lambda_m/\lambda_1} - 1}{\sqrt{\lambda_m/\lambda_1} + 1}, \qquad k = 1, 2, \ldots$$

(b) $\vec{x}_k = \vec{x}^*$, $\quad \vec{r}_k = \vec{0}$ $\quad$ *dla* $\quad k \geqslant m$.

$3^\sigma$ *Algorytm metody* me$(B)$ *jest następujący*

$$\vec{x}_{k+1} = \vec{x}_k + \frac{1}{q_k}\{e_{k-1}(\vec{x}_k - \vec{x}_{k-1}) - \vec{r}_k\},$$

*gdzie*

$$q_k = \frac{(\vec{r}_k, M^p \vec{r}_k)}{(\vec{r}_k, M^p(\vec{x}_k - \vec{x}^*))} - e_{k-1}, \qquad \vec{r}_k = M\vec{x}_k + \vec{g},$$

$$e_{-1} = 0, \qquad e_k = \frac{c_{k+1}}{c_k} q_k, \qquad \text{dla} \qquad c_k = (\vec{r}_k, M^{p-1}\vec{r}_k).$$

*Algorytm ten jest efektywny dla naturalnego p oraz dla p = 0, jeśli układ* (37) *nie jest sprzeczny* ($c_k = \|A\vec{x}_k + \vec{b}\|_2^2$). ■

**Bibliografia**

[0] G a n t m a c h e r, *Teoria macierzy* (po rosyjsku), II wyd., Moskwa 1966.
[1] G. G o l u b, R. S. V a r g a, *Chebyshev semi iterative methods, succesive overrelaxation iterative methods*, Part I, II, Num. Math. 3 (1961), str. 147–169.
[2] A. K i e ł b a s i ń s k i, G. W o ź n i a k o w s k a, H. W o ź n i a k o w s k i, *Algorytmizacja metod najlepszej strategii dla wielkich układów równań o symetrycznej dodatnio określonej macierzy*, Matematyka Stosowana 1 (1973), str. 47–68.
[3] M. A. K r a s n o s i e l s k i i inni, *Przybliżone rozwiązywanie równań operatorowych* (po rosyjsku), Moskwa 1969.
[4] H. R u t i s h a u s e r, E. S t i e f e l i inni, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problem*, 1959, ETH.
[5] E. S t i e f e l, *Kernel Polynomials in Linear Algebra and their Numerical Applications*, NBS. Appl. Math. Series 49 (1958), str 1–22.
[6] J. H. W i l k i n s o n, *Błędy zaokrągleń w procesach algebraicznych*, Warszawa 1967.
[7] G. W o ź n i a k o w s k a i H. W o ź n i a k o w s k i, *Algorytmizacja metody* me-T, Matematyka Stosowana, ten tom, str. 51–60.
[8] D. Y o u n g, *Iterative Solution of Large Linear Systems*, Academic Press, 1971.