ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
SERIA III: MATEMATYKA STOSOWANA I (1973)

A. KIEŁBASIŃSKI, GRAŻYNA WOŹNIAKOWSKA, H. WOŹNIAKOWSKI (Warszawa)

Algorytmizacja metod najlepszej strategii dla wielkich układów równań o symetrycznej, dodatnio określonej macierzy

Wstęp. Rozważamy zadanie znalezienia rozwiązania układu liniowego równań algebraicznych wysokiego stopnia postaci:

$$A\vec{x} + \vec{b} = \vec{0}, \tag{1}$$

gdzie $\vec{b}$ jest znanym wektorem, A macierzą kwadratową $n \times n$, symetryczną, dodatnio określoną.

Wymiar n jest tak duży, że ogólną ilość elementów macierzy $A \Rightarrow n^2$ znacznie przekracza pamięć operacyjną maszyny. Ponadto macierz A jest zazwyczaj szczególnej postaci, co często oznacza, że tylko nieliczne elementy w każdym wierszu są różne od zera. Z zadaniami postaci (1) spotykamy się często w praktyce np. przy problemach brzegowych dla równań różniczkowych cząstkowych typu eliptycznego, przy projektowaniu sieci, kratownic itp. Zadanie powyższe było przedmiotem rozważań licznych badaczy w ostatnich 20-tu latach. Wiadomo, że dla pewnych specjalnych typów macierzy A opracowano stabilne, wydajne metody rozwiązywania układu (1) (por. np. [1], [5]). Wskazano również metody optymalne dla pewnych klas zadań (por. [1], [7]).

Najlepsze z tych metod, np. metoda Czebyszewa, wymagają pewnej dodatkowej informacji o widmie macierzy A. Minimalną informacją jest na ogół znajomość najmniejszego przedziału zawierającego widmo macierzy A. Informacja taka nie zawsze jest łatwo dostępna. Z drugiej strony, istnieją metody, jak na przykład metoda sprzężonych gradientów lub metoda minimalnych residuów, nie wymagające żadnej informacji o widmie macierzy A. Metody te są jednak niestabilne.

W [5] i [6] znajdujemy propozycje połączenia dwóch metod: metody Czebyszewa z jedną z metod gradientowych, tak aby algorytm nie wymagał dodatkowych informacji o widmie macierzy A. O ile nam wiadomo, pomysł ten nie doczekał się w literaturze pełnego zalgorytmizowania, a doświadczenia ze stosowania tych połączeń w niektórych ośrodkach były negatywne. W poniższej pracy przedstawiamy próbę pełnego zalgorytmizowania metody Czebyszewa i metody minimalnych residuów (por. [2]).

Z przeprowadzonych rozważań teoretycznych i eksperymentów wynika stabilność i dobra wydajność tej metody.

1. Stabilność numeryczna i maksymalna graniczna dokładność w procesach iteracyjnych. W przypadku małych układów równań znamy algorytmy skończone (np. metoda eliminacji Gaussa, Householdera), które realizowane w arytmetyce zmiennopozycyjnej z t cyfrową mantysą binarną (por. [8]), pozwalają wyznaczyć wektor $\vec{x}$ będący rozwiązaniem układu zaburzonego

$$(A+E)\vec{x}+\vec{b}=\vec{0}, \qquad \text{gdzie } \|E\|_2 \leqslant 2^{-t} K \|A\|_2,$$

dla pewnej stałej K zależnej tylko od wymiaru n macierzy A oraz przyjętej metody (por. [8]).

Będziemy zakładali, że układ (1) jest tak dobrze uwarunkowany, że zaburzone układy, z zaburzeniami ograniczonymi jak powyżej, są nieosobliwe.

Stąd błąd rozwiązania

$$\vec{x}^* - \vec{x}, \qquad \text{gdzie } \vec{x}^* = -A^{-1}\vec{b}$$

spełnia zależność

$$\|\vec{x}^* - \vec{x}\|_2 \leqslant 2^{-t} \|A^{-1}\|_2 \|A\|_2 * K * \max(\|\vec{x}\|_2, \|\vec{x}^*\|_2). \tag{2}$$

Maksymalna dokładność osiągana przez stabilne skończone algorytmy zależy więc głównie od wskaźnika uwarunkowania macierzy A : $\|A^{-1}\|_2 \|A\|$. Fakt, że otrzymane rozwiązanie jest zawsze ścisłym rozwiązaniem układu zaburzonego o zaburzeniach co najwyżej K razy większych niż błąd reprezentacji danych, uznajemy za dowód wysokiej jakości algorytmu. Algorytmy takie za J. H. Wilkinsonem nazywamy *numerycznie stabilnymi.*

Przechodząc do analizy metod iteracyjnych spróbujmy odpowiedzieć na pytanie, czy możemy oczekiwać, że i w tym przypadku maksymalna osiągana dokładność również zależy w podobny sposób jak w (2) od wskaźnika uwarunkowania zadania?

Rozważmy w tym celu proces iteracyjny o zbieżności liniowej z wykładnikiem q.

Niech każdemu konstruowanemu elementowi $\vec{x}_i$ odpowiada zaburzone równanie

$$(A+E_i)\vec{y}_i+\vec{b}=\vec{0}, \qquad \text{gdzie}$$

$$\|E_i\|_2 \leqslant 2^{-t} * K * \|A\|_2,$$

takie przy tym, że dla jego rozwiązania $\vec{y}_i$ jest spełniona nierówność

$$\|\vec{x}_i - \vec{y}_i\|_2 \leqslant q \|\vec{x}_{i-1} - \vec{y}_i\|_2.$$

Można pokazać, że tę własność ma metoda iteracji prostych, różne metody nadrelaksacji, metoda Czebyszewa i inne. Stąd łatwo otrzymujemy nierówność

$$\|\vec{x}_i - \vec{x}^*\|_2 \leqslant \left(\frac{1+q}{1-q}\right) * (1-q^i)\Delta_i + q^i \|\vec{x}_0 - \vec{x}^*\|_2$$

gdzie $\Delta_i = \max\limits_{j \leqslant i} \|\vec{x}^* - \vec{y}_j\|_2$ (por. [4], str. 42).

Ponieważ normy $\|\vec{x}^* - \vec{y}_j\|_2$ posiadają oszacowanie typu (2), więc przy przyjętych założeniach o zbieżności procesu możemy jedynie gwarantować osiągnięcie takiej dokładności jak zapisana w (2) z wielkością

$$\mathcal{M} = \frac{\|A^{-1}\|_2 \|A\|_2}{1-q}(1+q)$$

na miejsce wskaźnika uwarunkowania $\|A\|_2 \|A^{-1}\|_2$.

Ponieważ, jak się okaże, realizowane procesy mają w najlepszym przypadku iloraz zbieżności

$$q = \frac{\sqrt{\|A^{-1}\|_2 \, \|A\|_2} - 1}{\sqrt{\|A^{-1}\|_2 \, \|A\|_2} + 1},$$

więc dla zadań źle uwarunkowanych zachodziłaby równość przybliżona

$$\mathcal{H} \cong (\|A^{-1}\|_2 \, \|A\|)^{3/2}$$

Jeśli ta pesymistyczna prognoza nie jest przesadzona (obecnie potrafimy tylko pokazać, że w przypadku metody iteracji prostych oraz metody Czebyszewa osiągana jest dokładność typu (2)) – to możemy powiedzieć, że osiągana dokładność jest $(\|A^{-1}\|_2 \, \|A\|_2)^{1/2}$ razy gorsza przy stosowaniu procesów iteracyjnych w porównaniu ze stabilnymi algorytmami skończonymi.

W przypadku takich procesów iteracyjnych, jedynie dla klas zadań dostatecznie dobrze uwarunkowanych możemy mówić o stabilności w sensie Wilkinsona. Konsekwencją tej pesymistycznej prognozy mógłby być również wniosek, że metodami iteracyjnymi jest sens rozwiązywać jedynie zadania, których wskaźnik uwarunkowania spełnia nierówność

$$(*) \qquad \|A^{-1}\|_2 \, \|A\|_2 < (2^t/K)^{2/3},$$

podczas gdy błędy reprezentacji numerycznej pozwalają nam uważać za wystarczająco dobrze uwarunkowane (dla danej arytmetyki) te zadania, dla których

$$\|A^{-1}\|_2 \, \|A\|_2 < 2^t/10.$$

Z drugiej strony, istnieje praktyczne, jeszcze silniejsze ograniczenie uwarunkowania zadania. Mianowicie, z wyrażenia dla q wynika, że dziesięciokrotna redukcja błędów wymaga około $\sqrt{\|A^{-1}\|_2 \, \|A\|_2}$ kroków iteracji. Trudno jest sobie wyobrazić sytuację, w której dla uzyskania jednej cyfry skłonni bylibyśmy wykonać dużo więcej niż 1000 kroków. Stąd

$$\|A^{-1}\|_2 \, \|A\|_2 \leqslant 10^6,$$

co automatycznie zapewnia warunek (*), gdyż dla współczesnych maszyn cyfrowych $t \geqslant 30$.

2. Metody iteracyjne. W dalszym ciągu założymy, że informacja o macierzy A jest nam dana poprzez procedurę, która dla danego wektora $\vec{x}$ oblicza wektor $\vec{y} = A\,\vec{x}$. Przy tym założeniu potrafimy jedynie konstruować dowolny wielomian od macierzy A na danym elemencie $\vec{x}$.

Niech $\vec{x}^*$ będzie rozwiązaniem (1):

$$\vec{x}^* = -A^{-1}\,\vec{b}.$$

Rozwiązania $\vec{x}$ poszukujemy metodą iteracyjną konstruując ciąg $\{\vec{x}_k\}$ taki, aby zachodziła zależność

$$(3) \qquad \vec{x}_k - \vec{x}^* = W_k(A)\,(\vec{x}_0 - \vec{x}^*)$$

gdzie $W_k(u)$ jest wielomianem zmiennej u stopnia $\leqslant k$, zaś $\vec{x}_0$ dowolnym przybliżeniem początkowym $\vec{x}^*$.

Zauważmy przede wszystkim, że wielomian $W_k(A)$ musimy wybrać tak, aby wyredukować nieznane $\vec{x}^*$ ze związku (3). Przepisując ten związek w postaci

$$\{W_k(A) - I\}\vec{x}^* = W_k(A)\vec{x}_0 - \vec{x}_k$$

widzimy, że zmuszeni jesteśmy narzucić warunek

$$W_k(0) = 1,$$

gdyż tylko wtedy

$$W_k(u) - 1 = a_k u^k + \ldots + a_1 u,$$

a więc wektor

$$\{W_k(A) - I\}\vec{x}^* = (a_k A^{k-1} + a_{k-1}A^{k-2} + \ldots + a_1 I)\, A\, \vec{x}^* =$$

$$= -(a_k A^{k-1} + a_{k-1}A^{k-2} + \ldots + a_1 I)\,\vec{b}$$

jest konstruowalnym wektorem.

a. *Metoda Czebyszewa* (T). Ze związku (3) otrzymujemy oszacowanie

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

Oczywiście jesteśmy zainteresowani takim wyborem wielomianu $W_k(A)$, aby przy warunku $W_k(0) = 1$, norma spektralna $\|W_k(A)\|_2$ była możliwie mała.

Niech $[\lambda_{\min}, \lambda_{\max}]$ oznacza przedział wartości własnych macierzy A, $\lambda_{\min} > 0$. Wówczas z symetrii macierzy A mamy:

$$\|W_k(A)\|_2 = \max_{\lambda \in \operatorname{spect}(A)} |W_k(\lambda)| \leqslant \|W_k\| = \max_{\lambda_{\min} \leqslant \lambda \leqslant \lambda_{\max}} |W_k(\lambda)|.$$

Zatem poszukujemy wielomianu spełniającego warunek

$$\|W_k\| = \inf_{P \in P_k(0,1)} \|P\|, \tag{4}$$

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

Rozwiązaniem problemu (4) jest, jak wiadomo (por. np. [1]), wielomian

$$W_k(u) = \frac{T_k\left(\dfrac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}} - \dfrac{2}{\lambda_{\max} - \lambda_{\min}}\, u\right)}{T_k\left(\dfrac{\lambda_{\max} + \lambda_{\min}}{\lambda_{\max} - \lambda_{\min}}\right)}, \tag{5}$$

gdzie $T_k(u)$ oznacza k-ty wielomian Czebyszewa pierwszego rodzaju. Ponadto wiemy (por. [1]), że

$$\|W_k\|_2 \leqslant 2 \cdot \left(\frac{\sqrt{\lambda_{max}} - \sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}} + \sqrt{\lambda_{min}}}\right)^k \cancel{\|\vec{x}_0 - \vec{x}^*\|_2}.$$

Metodę, w której wielomiany W_k są definiowane przez (5), nazywamy metodą Czebyszewa.

Tak więc w metodzie Czebyszewa konstruujemy ciąg $\{\vec{x}_k\}$ dla którego

$$\|\vec{x}_k - \vec{x}^*\|_2 \leqslant 2 \cdot \left(\frac{\sqrt{\lambda_{max}} - \sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}} + \sqrt{\lambda_{min}}}\right)^k \|\vec{x}_0 - \vec{x}^*\|_2. \tag{6a}$$

Mnożąc związek (3) przez macierz A i oznaczając $\vec{r}_k = A\,\vec{x}_k + \vec{b}$ otrzymujemy dla metody T

$$\|\vec{r}_k\|_2 \leqslant 2 \cdot \left(\frac{\sqrt{\lambda_{max}} - \sqrt{\lambda_{min}}}{\sqrt{\lambda_{max}} + \sqrt{\lambda_{min}}}\right)^k * \|\vec{r}_0\|_2. \tag{6b}$$

Zauważmy, że metoda ta wymaga znajomości najmniejszej i największej wartości własnej macierzy A, lub ogólniej przedziału zawierającego widmo A.

Ponadto wielomiany W_k konstruowane w tej metodzie nie zależą od przybliżenia początkowego $\vec{x}_0$. Dodatkowa ewentualna informacja zawarta w wektorze $\vec{x}_i$ nie jest wykorzystywana w tej metodzie dla przyspieszenia zbieżności.

b. *Metoda sprzężonych gradientów* (cg). Niech

$$\vec{x}_0 - \vec{x}^* = \sum_{j=1}^{m} c_j\,\vec{\xi}_j \qquad (m \leqslant n,\ c_j \neq 0), \tag{7}$$

gdzie $\vec{\xi}_j$ są unormowanymi wektorami własnymi przynależnymi do różnych wartości własnych λ_j $(\lambda_1 < \lambda_2 < \ldots < \lambda_m)$ macierzy A: $A\,\vec{\xi}_j = \lambda_j\,\vec{\xi}_j$, $\|\vec{\xi}_j\|_2 = 1$. Zastanówmy sie jak najracjonalniej wybrać wielomiany W_k? Chcemy zminimalizować błąd $\vec{x}_k - \vec{x}^*$, a więc z (3) na wielomian W_k narzućmy warunek

$$\|W_0(A)(\vec{x}_0 - \vec{x}^*)\|_2 = \inf_{P \in P_k(0,1)} \|P(A)(\vec{x}_0 - \vec{x}^*)\|_2. \tag{8}$$

Zatem $W_k(A)$ minimalizuje normę $\|W_k(A)(\vec{x}_0 - \vec{x}^*)\|_2$ w klasie wielomianów $P_k(0, 1)$. (Zastąpienie $\|P(A)(\vec{x}_0 - \vec{x}^*)\|_2$ oszacowaniem $\|P(A)\|_2\ \|\vec{x}_0 - \vec{x}^*\|_2$ prowadzi do metody Czebyszewa).

Z ogólnej teorii (por. [6]) wynika, że wielomiany $W_k(x)$, $k = 0, 1, \ldots$ będące rozwiązaniem problemu (8) są wielomianami ortogonalnymi względem dyskretnego iloczynu skalarnego

$$(9) \qquad (f, g) = \sum_{j=1}^{m} \lambda_j c_j^2 f(\lambda_j) g(\lambda_j)$$

(stałe c_j pochodzą z rozłożenia spektralnego wektora $\vec{x}_0 - \vec{x}^*$ (por. (7))).

Niestety, aby wyznaczać współczynniki formuły trójczłonowej dla wielomianów W_k ortogonalnych względem (9) musielibyśmy obliczać między innymi wielkości

$$(\vec{r}_k, \vec{x}_k - \vec{x}^*), \qquad \text{dla } k = 0, 1, \ldots$$

gdzie $\vec{r}_k$ jest wektorem residualnym ($\vec{r}_k = A\,\vec{x}_k + \vec{b}$).

Ponieważ, na ogół nie potrafimy obliczyć tego iloczynu skalarnego, zmuszeni jesteśmy więc zrezygnować z metody minimalizującej normę błędu $\|\vec{x}_k - \vec{x}^*\|_2$.

Nie mogąc skonstruować efektywnie metody minimalizującej $\|\vec{x}_k - \vec{x}^*\|_2$ zastanówmy się, czy minimalizacja innych wektorów pochodnych od $\vec{x}_k - \vec{x}^*$ nie doprowadzi nas do metod efektywnie konstruowalnych. I tak przechodzimy do metody sprzężonych gradientów (cg), która minimalizuje wielkość

$$(\vec{r}_k, \vec{x}_k - \vec{x}^*) = (A(\vec{x}_k - \vec{x}^*), \vec{x}_k - \vec{x}^*) = \|A^{1/2}(\vec{x}_k - \vec{x}^*)\|_2^2.$$

W metodzie cg korzystamy więc z wielomianów $W_k(A)$, które spełniają relację

$$(10) \qquad \|W_k(A)\,A^{1/2}(\vec{x}_0 - \vec{x}^*)\|_2 = \inf_{P \in P_k(0,1)} \|P(A)\,A^{1/2}(\vec{x}_0 - \vec{x}^*)\|_2.$$

Wielomany $W_k(\lambda)$ są więc ortogonalne względem iloczynu skalarnego

$$(11) \qquad (f, g) = \sum_{j=1}^{m} \lambda_j^2 c_j^2 f(\lambda_j) g(\lambda_j).$$

Dla wyznaczania współczynników formuły trójczłonowej wystarczy teraz (por. [5]) znajomość łatwo obliczalnych wielkości

$$(\vec{r}_k, \vec{r}_k), \qquad (A\,\vec{r}_k, \vec{r}_k) \qquad \text{dla } k = 0, 1, \ldots$$

Metoda cg jest metodą skończoną,tzn. teoretycznie, po m-krokach otrzymujemy wielomian W_m taki, że

$$W_m(\lambda_j) = 0 \qquad \text{dla } j = 1, \ldots, m, \qquad \text{czyli } \vec{x}_m = \vec{x}^*.$$

Z uwagi jednak na błędy zaokrągleń występujące przy numerycznej realizacji metody cg nie otrzymamy po m krokach rozwiązania, co więcej metoda ta, jako metoda ortogo-

nalizacyjna bez poprawiania, nie jest stabilna, a więc wektor $\vec{x}_m$ może znacznie różnić się od rozwiązania $\vec{x}^*$.

Ze związku (10) możemy otrzymać ponadto oszacowanie k-tego przybliżenia rozwiązania. Kładąc za wielomian $P(A)$ wielomian definiowany związkiem (5), otrzymujemy

$$\|\vec{x}_k - \vec{x}^*\|_2 \leqslant \frac{2}{\lambda_{\min}}\left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^k * \|\vec{r}_0\|_2. \tag{12a}$$

Analogicznie otrzymujemy oszacowanie dla wektora residualnego r:

$$\|\vec{r}_k\|_2 \leqslant 2 * \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}} * \left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^k * \|\vec{r}_0\|_2. \tag{12b}$$

c. *Metoda minimalnych residuów* (mr). Jak wiemy wektor residualny $\vec{r}_k$ spełnia związek

$$\vec{r}_k = W_k(A)\vec{r}_0.$$

Metoda minimalnych residuów polega na wyborze wielomianów $W_k(A)$ tak, aby zminimalizować normę $\vec{r}_k$, czyli

$$\|W_k(A)\vec{r}_0\|_2 = \inf_{P \in P_k(0,1)} \|P(A)\vec{r}_0\|_2. \tag{13}$$

Wielomiany $W_k(A)$ są teraz wielomianami ortogonalnymi względem iloczynu skalarnego

$$(f, g) = \sum_{j=1}^{m} \lambda_j^3 c_j^2 f(\lambda_j) g(\lambda_j)$$

(E. Stiefel w [6] nazywa tę metodę *zmodyfikowaną metodą sprzężonych gradientów*).

Metoda ta, podobnie jak metoda cg, jest teoretycznie metodą skończoną. Z uwagi jednak na niestabilność tej metody nie możemy liczyć na osiągnięcie rozwiązania po m krokach. Ze związku (13) kładąc za $P(A)$ wielomian (5) otrzymujemy oszacowanie

$$\|\vec{r}_k\|_2 \leqslant 2 * \left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^k * \|\vec{r}_0\|_2 \tag{14a}$$

(por. (6b), (12b)), stąd też

$$\|\vec{x}_k - \vec{x}^*\|_2 \leqslant \frac{2}{\lambda_{\min}} * \left(\frac{\sqrt{\lambda_{\max}} - \sqrt{\lambda_{\min}}}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^k * \|\vec{r}_0\|_2 \tag{14b}$$

(por. (6a), (12a)).

Ze związków (12) i (14) widzimy, że redukcja błędu $\vec{x}_k - \vec{x}^*$ w obu metodach cg, mr jest, w sensie oszacowania, jednakowa. Ale w metodzie minimalnych residuów minimalizacja residuum jest na ogół $\sqrt{\frac{\lambda_{max}}{\lambda_{min}}}$ razy lepsza niż w metodzie cg.

Należy zwrócić uwagę, że malenie norm residuów w metodzie mr pozwala nam na racjonalne sterowanie procesem obliczeniowym (por. punkt 4) natomiast w metodzie cg często się zdarza, że początkowe normy $\vec{r}_k$ rosną, nieraz dość znacznie, w porównaniu do normy $\vec{r}_0$. Fakt ten bardzo utrudnia racjonalne korzystanie z tej metody. Dlatego w dalszym ciągu skoncentrujemy swoją uwagę przede wszystkim na metodzie mr. Oszacowania (14) nie obejmują oczywiście błędu wytworzonego zaokrągleniami, ale obrazują w pewnym stopniu minimalne oczekiwane szybkości malenia błędów. W związku z tym stanowią ważną przesłankę dla konstrukcji algorytmu łączącego metodę mr z metodą Czebyszewa.

3. Algorytmy metod T, mr. Korzystając z formuł rekurencyjnych dla wielomianów Czebyszewa, jak również z formuł trójczłonowych dla wielomianów ortogonalnych w metodzie mr, możemy podać związki rekurencyjne na wyznaczanie kolejnych przybliżeń $\vec{x}_k$ oraz residuów $\vec{r}_k$ (por. [5]).

gdzie współczynniki e_k, q_k są równe:

1° dla metody T na przedziale $< a, b >$

$$q_0 = \frac{a+b}{2},$$

$$q_1 = \frac{b+a}{2} - \frac{(b-a)^2}{4(b+a)},$$

$$q_k = \frac{b+a}{2} - \frac{b-a}{16\, q_{k-1}} \qquad (k = 2, 3, \ldots),$$

$$e_{-1} = 0$$

$$e_0 = \frac{(b-a)^2}{4(a+b)},$$

$$e_k = \frac{(b-a)^2}{8\,(b+a) - 16\, e_{k-1}} \qquad (k = 1, 2, \ldots);$$

2° dla metody mr

$$e_{-1} = 0,$$

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

$$e_k = q_k \frac{(A\vec{r}_{k+1}, \vec{r}_k)}{(A\vec{r}_k, \vec{r}_k)} \qquad (k = 0, 1, \ldots).$$

Numeryczna realizacja metody T wymaga 3 wektorów, a więc ilość miejsc pamięci jest w przybliżeniu równa $3n$. Koszt jednego kroku metody T jest w przybliżeniu równy kosztowi mnożenia danego wektora przez macierz A. Realizację metody mr można również przeprowadzić dysponując tylko trzema wektorami $3n$ miejsc w pamięci, lecz wówczas koszt każdego kroku jest zbliżony do kosztu dwóch mnożeń wektora przez macierz A.

Stosując rekurencyjne wyznaczanie kolejnych residuów i dysponując pięcioma wektorami ($5n$ miejsc w pamięci) możemy realizować jeden krok w metodzie mr kosztem jednego tylko mnożenia wektora przez macierz A.

4. Metoda mr - T. Załóżmy, że nasze zadanie polega na znalezieniu takiego wektora $\vec{x}_k$, aby odpowiadający mu wektor residualny $\vec{r}_k$ spełniał nierówność:

$$\|\vec{r}_k\|_2 \leqslant \text{eps} * \|\vec{r}_0\|_2, \tag{15}$$

gdzie eps jest dodatnią, z góry zadaną liczbą, wyrażającą żądaną redukcję normy początkowego wektora residualnego.

Jak wiemy, z punktu 3, metoda Czebyszewa (T) wymaga znajomości λ_{min}, λ_{max} lub przedziału $< a,b >$ zawierającego widmo macierzy A $(0 < a)$.

Największą wartość λ_{max} jest zazwyczaj łatwo oszacować, chociażby z twierdzenia Gerszgorina, względnie, kładąc za b dowolną normę macierzy A zgodną z normą wektora. Dużo trudniej jest podać dodatnie oszacowanie z dołu wartości własnych (a). Załóżmy zatem, że stosujemy metodę T na fałszywym przedziale $<a, b>$, gdzie $\lambda_{min} < a$, $\lambda_{max} \leqslant b$. Metodę Czebyszewa na przedziale $< a, b >$ oznaczamy przez $T < a, b >$.

Zbadajmy, jak zachowuje się wektor residualny $\vec{r}_k$. Podobnie jak w (7) niech

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

Wówczas k-ty wektor residualny jest równy

$$\vec{r}_k = \sum_{j=1}^{m} c_j \lambda_j W_k(\lambda_j) \vec{\xi}_j, \tag{16}$$

gdzie wielomian $W_k(u)$ unormowany warunkiem $W_k(0) = 1$, najlepiej przybliża zero w przedziale $< a, b >$ w klasie wielomianów $P_k(0,1)$, czyli

$$W_k(x) = \frac{T_k\left(\frac{b+a}{b-a} - \frac{2}{b-a}x\right)}{T_k\left(\frac{b+a}{b-a}\right)}.$$

Ze związku (16) widzimy, że składowe wektora residualnego $\vec{r}_k$ w kierunkach przynależnych do wartości własnych $\lambda_j \geqslant a$ są mnożone przez $W_k(\lambda_j)$, przy czym

$$|W_k(\lambda_j)| \leqslant 2 \cdot \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^k.$$

Natomiast składowe odpowiadające wartościom własnym $\lambda_j < a$ są redukowane tak jak:

$$W_k(\lambda_j) \cong \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \cdot \frac{\sqrt{b - \lambda_j} + \sqrt{a - \lambda_j}}{\sqrt{b - \lambda_j} - \sqrt{a - \lambda_j}}\right)^k.$$

Dla dostatecznie dużych k mamy zatem prawo napisać „równość asymptotyczną"

$$\vec{r}_k \cong \sum_{\lambda_j < a} \lambda_j c_j W_k(\lambda_j) \vec{\xi}_j \cong \left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}} \cdot \frac{\sqrt{b - \lambda_1} + \sqrt{a - \lambda_1}}{\sqrt{b - \lambda_1} - \sqrt{a - \lambda_1}}\right)^k c_1 \lambda_1 \vec{\xi}_1, \tag{17}$$

gdzie λ_1 jest najmniejszą wartością własną ($\lambda_{min} \leqslant \lambda_1$) w rozłożeniu spektralnym wektora $\vec{r}_0$.

Obliczając iloraz Rayleigh dla wektora $\vec{r}_k$, otrzymujemy

$$\lambda_1 \leqslant \frac{(A\vec{r}_k, \vec{r}_k)}{(\vec{r}_k, \vec{r}_k)} \cong \lambda_1.$$

Tak więc, konstruując ciąg $\{\vec{x}_k\}$ metodą $T < a, b >$, gdzie $\lambda_{min} < a$, $\lambda_{max} \leqslant b$ i obliczaczając ilorazy Rayleigh dla wektorów residualnych otrzymujemy przybliżenie λ_1 — najmniejszej wartości własnej w rozłożeniu spektralnym $\vec{r}_0$.

Co więcej można pokazać, że

$$\frac{(A\vec{r}_k, \vec{r}_k)}{(\vec{r}_k, \vec{r}_k)} = \lambda_1 \left(1 + \left(\frac{\lambda_2}{\lambda_1} - 1\right)\left(\frac{\lambda_2 c_2}{\lambda_1 c_1}\right)^2 * \omega_1^{2k} + O(\omega_2^{2k})\right),$$

gdzie

$$\omega_i = \frac{\sqrt{b-\lambda_i} + \sqrt{a-\lambda_i}}{\sqrt{b-\lambda_1} + \sqrt{a-\lambda_1}} * \frac{\sqrt{b-\lambda_1} - \sqrt{a-\lambda_1}}{\sqrt{b-\lambda_i} - \sqrt{a-\lambda_i}} \qquad \text{dla } i = 1, 2.$$

Tak więc szybkość zbieżności ilorazów Rayleigh do λ_1 jest liniowe z ilorazem ω_1^2 zależnym od a, b oraz λ_1, λ_2. Należy podkreślić, że im zadanie jest gorzej uwarunkowane (tzn. $\frac{\lambda_{max}}{\lambda_{min}} \gg 1$), tym więcej należy wykonać kroków metodą $T < a, b >$, aby uzyskać dobre przybliżenie λ_1 (por. punkt 6).

Znajomość λ_1 wystarczy już, aby zastosować metodę $T < \lambda_1, b >$. Jako punkt startowy należałoby oczywiście przyjąć ostatnie przybliżenie z metody $T < a, b >$. Dla nowo konstruowanego ciągu błąd dąży do zera tak, jak

$$\left(\frac{\sqrt{b} - \sqrt{\lambda_1}}{\sqrt{b} + \sqrt{\lambda_1}}\right)^k.$$

Oczywiście, zazwyczaj $\lambda_1 = \lambda_{min}$, ale gdy wektor $\vec{r}_0$ jest ortogonalny do kilku wektorów własnych przynależnych do najmniejszych wartości własnych macierzy A, to możemy wykorzystać to szczególne położenie wektora $\vec{r}_0$ i przyśpieszyć proces obliczeniowy pracując na przedziale $< \lambda_1, b >$.

Podkreślamy, że otrzymane kolejno przybliżenia λ_1 są większe od λ_1, a więc nie grozi nam pogorszenie szybkości na skutek zbytniego rozszerzenia przedziału. Znamy zatem przybliżenie λ_1, tzn.

$$a_1 := \frac{(A\vec{r}_{k_1}, \vec{r}_{k_1})}{(\vec{r}_{k_1}, \vec{r}_{k_1})},$$

gdzie $\vec{r}_{k_1}$ jest ostatnim obliczonym wektorem z metody $T < a, b >$. Wskaźnik k_1 dobieramy tak, aby

$$\left(\frac{\sqrt{b} - \sqrt{a}}{\sqrt{b} + \sqrt{a}}\right)^{k_1} \leqslant \frac{\text{eps}}{2}.$$

Powyższy warunek oznacza, że jeśli nieznane λ_{min} byłoby większe lub równe a, to wektor $\vec{r}_{k_1}$ spełniałby warunek (15) i zadanie byłoby już wykonane.

Powtórne zastosowanie metody T ma jednak tę wadę, że składowe błędu przynależne do kierunków o wartościach własnych $\lambda_j \geqslant a$, które są już zredukowane w stosunku eps

byłyby ponownie redukowane w stosunku $2 \cdot \left(\frac{\sqrt{b}-\sqrt{a_1}}{\sqrt{b}+\sqrt{a_1}}\right)^k$. Nie wykorzystalibyśmy zatem szczególnej postaci wektora $\vec{r}_{k_1}$ (por. (17)), który w przybliżeniu leży w podprzestrzeni rozpiętej na wektorach własnych przynależnych do $\lambda_j \leqslant a$. Jesteśmy teraz głównie zainteresowani redukcją składowych z tej podprzestrzeni.

Chcąc wykorzystać szczególną postać wektora $\vec{r}_{k_1}$ zastosujemy teraz kilka kroków metodą mr. Metoda minimalnych residuów – mr konstruuje ciąg $\vec{r}_{k+k_1}$ w myśl reguły

$$\vec{r}_{k+k_1} = W_k(A)\vec{r}_{k_1},$$

gdzie wielomian $W_k(A)$ jest tak wybrany z klasy $P_k(0,1)$, aby zminimalizować normę $\vec{r}_{k+k_1}$. Można pokazać, że przy k_1 dostatecznie dużym dla kilku początkowych kroków ($k \ll k_1$) metody mr

$$W_k(x) \cong \left(1-\frac{x}{\lambda_1}\right)\left(1-\frac{x}{\lambda_2}\right)\cdot \ldots \cdot \left(1-\frac{x}{\lambda_k}\right),$$

tak więc kilka kroków redukuje głównie składowe z podprzestrzeni rozpiętej na wektorach własnych przynależnych do $\lambda_1, \lambda_2, \ldots, \lambda_k$, a więc te składowe, które były najmniej redukowane w metodzie $T < a, b >$.

Natomiast składowe odpowiadające większym wartościom własnym są odpowiednio przemnażane przez $W_k(\lambda_j)$, a więc często ulegają zwiększeniu. W metodzie mr powiększanie odpowiednich składowych nie jest groźne, gdyż wielomian $W_k(A)$ minimalizuje normę residuum $\vec{r}_k$.

Gdybyśmy zamiast metody mr chcieli używać metody cg, to wzrost normy residuum, szczególnie przy zadaniach źle uwarunkowanych byłby bardzo istotny.

Metoda mr jako metoda ortogonalizacyjna, nie jest metodą stabilną tak więc wykonywanie zbyt dużej ilości kroków tą metodą może być niecelowe.

Powstaje problem, jak długo należy pracować metodą mr? W oparciu o zależności (14) proponujemy następujące kryterium: *tak długo wykonujemy kolejne kroki, jak długo spełnione są warunki:*

$$1^\circ \quad \|\vec{r}_{k+k_1}\|_2 \leqslant 2 \cdot \left(\frac{\sqrt{b}-\sqrt{a_1}}{\sqrt{b}+\sqrt{a_1}}\right)^k \cdot \|\vec{r}_k\|_2,$$

$$2^\circ \quad \frac{(A\vec{r}_{k+k_1}, \vec{r}_{k+k_1})}{(\vec{r}_{k+k_1}, \vec{r}_{k+k_1})} < a.$$

Warunek 1° oznacza, że metoda mr jest co najmniej tak szybko zbieżna jak metoda $T < a_1, b >$.

Jeśli a_1 jest dostatecznie dobrym przybliżeniem λ_1, to warunek 1° powinien teoretycznie zawsze być spełniony. Badamy zatem, czy błędy zaokrągleń nie przekroczyły już oszacowania teoretycznego.

Warunek 2° bada zależność między kolejnymi ilorazami Rayleigh, a dolnym krańcem przedziału $< a, b >$ z metody $T < a, b >$. Iloraz Rayleigh jest średnią ważoną wartości własnych $\lambda_1, \lambda_2, \ldots, \lambda_m$. Z uwagi na to, że składowe odpowiadające $\lambda_j \geqslant a$ były już zredukowane przez metodę $T < a, b >$, a kilka kolejnych ilorazów powinno przybliżać najmniejsze wartości własne, to nie spełnienie warunku 2° oznacza że składowe odpowiada-

jące $\lambda_j \geqslant a$ uległy znacznemu zwiększeniu i są porównywalne ze składowymi odpowiadającymi $\lambda_j < a$. Po wykonaniu kilku kroków metodą mr, powiedzmy k kroków, przechodzimy ponownie do metody Czebyszewa.

Z uwagi na to, że wielomian $W_k(A)$ zredukował głównie składowe z podprzestrzeni rozpiętej na wektorach własnych przynależnych do najmniejszych wartości własnych, metodę T chcielibyśmy zastosować teraz do przedziału $< \tilde{a}, b >$, gdzie

$$\tilde{a} = \min\left(a, \max_{0 \leqslant i \leqslant k}\left(\frac{(A\vec{r}_{k+k_1}, \vec{r}_{k+k_1})}{(\vec{r}_{k+k_1}, \vec{r}_{k+k_1})}\right)\right).$$

Stosując metodę T $< \tilde{a}, b >$ z wektorem $\vec{x}_{k+k_1}$ jako wektorem początkowym chcemy zredukować normę wektora $\vec{r}_{k+k_1}$ w stosunku $\mathcal{H} = \frac{\text{eps}}{\|\vec{r}_{k+k_1}\|_2}\|\vec{r}_0\|_2$. Gdy jednak liczba $\mathcal{H}$ jest bliska jedynki, to wykonamy niewielką ilość kroków metodą T $< \tilde{a}, b >$, a więc nie zredukujemy wiele składowych odpowiadających $\lambda_j \geqslant \tilde{a}$ w porównaniu do składowych odpowiadających wartościom własnym $\lambda_j < \tilde{a}$. A więc metoda mr nie będzie mogła redukować błędu w podprzestrzeni mało redukowanej przez T $< \tilde{a}, b >$. Dlatego, chcąc umożliwić redukcję składowych o małych wartościach własnych, będziemy redukować normę $\vec{r}_{k+k_1}$ w stosunku

$$\min\left(\frac{\text{eps}}{\|\vec{r}_{k+k_1}\|}\|\vec{r}_0\|, \frac{\alpha}{b}\right),$$

gdzie

$$\alpha = \min_{0 \leqslant i \leqslant k} \frac{(A\vec{r}_{k+k_1}, \vec{r}_{k+k_1})}{(\vec{r}_{k+k_1}, \vec{r}_{k+k_1})}$$

jest najlepszym znanym przybliżeniem najmniejszej wartości własnej. Stosowanie naprzemian, jak opisaliśmy powyżej, metod T i mr, nazywamy metodą mr-T. Podkreślamy, że metoda mr-T nie wymaga znajomości najmniejszej wartości własnej, zaś przy pierwszym zastosowaniu metody T $< a, b >$ możemy położyć na przykład $a = b/2$ (por. punkt 6).

5. Uwagi o algorytmie mr-T. Przypominamy, że naszym celem jest konstrukcja takiego wektora $\vec{x}_k$, aby

$$\|\vec{r}_k\|_2 \leqslant \text{eps} * \|\vec{r}_0\|_2, \tag{18}$$

gdzie eps jest z góry zadaną liczbą.

Oczywiście, jeśli nierówność (18) zachodzi dla danego k, to kończymy proces obliczeniowy. Przerwanie procesu obliczeniowego może mieć również miejsce, gdy

1° macierz układu A nie jest dodatnio określona lub jest numerycznie osobliwa,

2° osiagnęliśmy graniczną maksymalną dokładność w danej arytmetyce (por. [8]).

Omówimy pokrótce te dwa przypadki.

1° W metodzie minimalnych residuów — mr — obliczamy wielkości $(A\vec{r}_k, \vec{r}_k)$, a więc bez zwiększania kosztu procesu możemy obliczyć iloraz Rayleigh

$$\lambda = \overline{\mathrm{fl}}\left(\frac{(A\vec{r}_k, \vec{r}_k)}{(\vec{r}_k, \vec{r}_k)}\right) = \frac{((A - E_k)\vec{r}_k, \vec{r}_k)}{(\vec{r}_k, \vec{r}_k)},$$

gdzie $\overline{\mathrm{fl}}$ oznacza arytmetykę zmiennopozycyjną z zastosowaniem algorytmu poprawiania przy obliczaniu iloczynów skalarnych (por. [3]), zaś E_k jest macierzą pozornych równoważnych zaburzeń (por. [8]) o normie euklidesowej

$$\|E_k\|_E \leqslant s * 2^{-t} * \|A\|_E,$$

gdzie s maksymalna ilość niezerowych elementów w wierszach macierzy A, t ilość cyfr mantysy w arytmetyce zmiennopozycyjnej.

Obliczony iloraz Rayleigh

$$\lambda \geqslant \lambda_1 - \|E_k\|_E,$$

gdzie λ_1 jest najmniejszą wartością własną w rozłożeniu spektralnym wektora $\vec{r}_0$. Zatem A nie jest dodatnio określona, gdy

$$\lambda < -\|E_k\|_E,$$

natomiast A jest numerycznie osobliwa, gdy

$$|\lambda| \leqslant \|E_k\|_E.$$

2° Osiągnięcie granicznej maksymalnej dokładności będzie miało miejsce wtedy, gdy obliczony wektor residualny jest na poziomie błędu wytworzonego, czyli gdy

$$\|\vec{r}_k\|_2 \leqslant s * 2^{-t} * \|A\|_E * \|\vec{x}_k\|_2.$$

Załóżmy teraz, że wykonaliśmy proces obliczeniowy metodą mr-T otrzymując jako przybliżenie rozwiązania wektor $\vec{x}_k$ i odpowiadający mu wektor residualny $\vec{r}_k$. Czy po wykonaniu metody mr-T możemy otrzymać oszacowanie błędu: $\vec{x}^* - \vec{x}_k$?

Wiemy, że

$$\vec{r}_k = (A - E_k)\,\vec{x}_k + \vec{b},$$

zatem

$$\vec{x}^* - \vec{x}_k = A^{-1} E_k \vec{x}_k - A^{-1} \vec{r}_k,$$

stąd też

$$\|\vec{x}^* - \vec{x}_k\|_2 \leqslant 2^{-t} (s\,\|A\|_E * \|\vec{x}_k\|_2 + \|\vec{r}_k\|_2)/\lambda_{\min}(A). \tag{19}$$

Jedyną wielkością, której nie znamy jest $\lambda_{\min}(A)$. Pamiętamy jednak że w metodzie mr-T możemy często dostać dostatecznie dobre przybliżenie najmniejszej wartości włas-

nej λ_1 w rozłożeniu spektralnym wektora $\vec{r}_0$. Jeśli $\lambda_1 \cong \lambda_{min}$, to możemy mieć nadzieję, że po wykonaniu metody mr-T otrzymamy przybliżenie λ_{min}, a zatem obliczymy oszacowanie błędu a posteriori w myśl (19). Podkreślamy, że aby obliczyć λ_1, to odpowiednie składowe w kierunkach odpowiadającym najmniejszym wartościom własnym w wektorze $\vec{r}_0$ nie mogą być zbyt małe.

Często się zdarza, że dla zadań bardzo źle uwarunkowanych w danej arytmetyce powyższe założenie nie jest spełnione. Mianowicie, załóżmy, że

$$\vec{x}_0 - \vec{x}^* = \sum_{j=1}^{n} c_j \vec{\xi}_j$$

ma składowe c_j mniej więcej równe sobie, powiedzmy $c_j = 1$. Wówczas wektor $\vec{r}_0$ równy

$$\vec{r}_0 = \sum_{j=1}^{n} \lambda_j \vec{\xi}_j$$

posiada składowe patologicznie małe w kierunkach odpowiadających małym wartościom własnym.

Dla takich zadań, zazwyczaj, nie otrzymamy dobrego przybliżenia λ_{min}, a tym samym nie potrafimy oszacować błędu $\|\vec{x}^* - \vec{x}_k\|_2$. Dlatego też, z oszacowania błędu a posteriori danego przez metodę mr-T można korzystać tylko wtedy, gdy jesteśmy pewni znajomości najmniejszej wartości własnej λ_{min} (por. punkt 6).

6. Przykłady numeryczne. Metodę mr-T testowaliśmy dla układów

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

o macierzy A postaci

$$A = \left(I - \frac{2}{(\vec{w}, \vec{w})}\vec{w} \cdot \vec{w}^T\right) \operatorname{diag}(\lambda_i) \left(I - \frac{2}{(\vec{w}, \vec{w})}\vec{w} \cdot \vec{w}^T\right),$$

gdzie $\vec{w}$ – wektor losowy, zaś wartości własne λ_i: $0 < \lambda_i \leqslant 1$.

Za wektor prawych stron $\vec{b}$ kładliśmy

$$\vec{b} = -\overline{\mathrm{fl}}\,(A\vec{x}^*),$$

gdzie $\vec{x}^* = [1, 2, \ldots, n]^T$, oznacza arytmetykę zmiennopozycyjną z zastosowaniem algorytmu poprawiania przy obliczaniu iloczynów skalarnych (por. [3]). Obliczenia przeprowadzaliśmy dla różnych rozkładów wartości własnych. Poniżej zamieszczamy tabelkę dla następujących rozkładów:

1. Rozkład jednostajny

$$\lambda_i = \frac{i-k}{n}, \qquad i = 1, 2, \ldots, n, \tag{20}$$

dla pewnej wartości parametru k. Dla $k = 1$, otrzymywaliśmy macierz osobliwą, dla $k > 1$ macierz nie dodatnio określoną.

2. Wysunięta jedna wartość własna

$$\lambda_1 = \frac{1}{k^2}, \quad \lambda_i = \frac{1}{k} + \left(1 - \frac{1}{k}\right) \cdot \frac{i-2}{n-2}, \quad i = 2, \ldots, n.$$

3. Wysunięte grupy wartości własnych

$$\lambda_i = \begin{cases} \frac{1}{k^2} + \frac{w_i}{k^3}, & 1 \leqslant i \leqslant 10, \\ \frac{1}{10} + \frac{w_i}{k^3}, & 11 \leqslant i \leqslant 20, \\ \frac{2}{10} + \frac{w_i}{k^3}, & 21 \leqslant i \leqslant 30, \\ 0.7 + \frac{w_i}{8}, & i \geqslant 31, \end{cases}$$

gdzie w_i oznacza wylosowaną liczbę z przedziału $< -1, +1 >$.

4. Rozkład geometryczny

$$\lambda_i = \begin{cases} \left(\frac{1}{2}\right)^{k+1-i}, & i \leqslant k, \\ 0.7 + \frac{w_i}{8}, & i > k, \end{cases}$$

gdzie w_i wylosowana liczba z przedziału $< -1, 1 >$.

Obliczenia przeprowadziliśmy dla metody mr-T oraz dla metody będącej połączeniem metody Czebyszewa z metodą sprzężonych gradientów (cg-T).

Przypominamy, że $\lambda_{max} = \|A\|_2 \leqslant 1$, a obliczenie zaczynaliśmy z losowo wybranego wektora $\vec{r}_0$.

Powyższe metody testowaliśmy dla różnych wartości n, $30 \leqslant n \leqslant 200$. Wyniki były jednak przede wszystkim zależne od rozkładu wartości własnych, a nie od n.

Wszystkie obliczenia przeprowadziliśmy w Zakładzie Obliczeń Numerycznych Uniwersytetu Warszawskiego na maszynie GIER ($t = 28$).

W kolumnie (3) podajemy numer rozkładu wartości własnych w myśl (20). Kolumna (4) zawiera najmniejszą wartość własną, kolumna (5) jej przybliżenie uzyskane w danej metodzie. W kolumnach (6) i (7) podajemy normę wektora residualnego oraz błąd względny otrzymanego wyniku. Kolumna (8) zawiera ilość kroków iteracyjnych wykonanych w metodzie, natomiast w kolumnie (9) podajemy ilość kroków w metodzie Czebyszewa realizowanej na prawdziwym przedziale $< \lambda_{min}, 1 >$, która dałaby nam wektor $\vec{r}_k$ o tej samej normie co metoda mr-T.

W przypadku macierzy osobliwej lub nie dodatnio określonej podajemy ilość kroków po wykonaniu których otrzymaliśmy informację o numerycznej osobliwości lub nie dodatniej określoności układu.

Test metody mr-T

Porównanie z metodą cg-T

1	2	3	4	5	6	7	8	9
Lp.	Metoda	Rozkład wartości własnych	λ_{min}	Obliczone λ_{min}	Norma $\lVert \vec{r}_k \rVert_2$	Błąd $\frac{\lVert \vec{x}_k - \vec{x}^* \rVert_2}{\lVert \vec{x}^* \rVert_2}$	Ilość kroków w mr-T	Ilość kroków w T
1	mr-T cg-T	1	$-2_{10}-2$				15 28	
2	mr-T cg-T	1	0				60 830	
3	mr-T cg-T	1	$2_{10}-2$	$2.002_{10}-2$ $2.07_{10}-2$	$3_{10}-6$ $3_{10}-6$	$5_{10}-7$ $_{10}-7$	50 52	65
4	mr-T cg-T	4	$9.76_{10}-4$	$2_{10}-3$ $3_{10}-3$	$3_{10}-6$ $5_{10}-6$	$5_{10}-7$ $_{10}-6$	92 200	292
5	mr-T cg-T	2	$6.25_{10}-4$	$6.25_{10}-4$ $1.0_{10}-3$	$4_{10}-6$ $3_{10}-6$	$2_{10}-5$ $_{10}-5$	83 247	367
6	mr-T cg-T	3	$6_{10}-4$	$6.2_{10}-4$ $6.5_{10}-4$	$2_{10}-6$ $2_{10}-6$	$2_{10}-5$ $_{10}-5$	32 51	379
7	mr-T cg-T	3	$2.5_{10}-5$	$2.51_{10}-5$ $9_{10}-5$	$2_{10}-6$ $3_{10}-5$	$5_{10}-5$ $_{10}-3$	47 47	1851
8	mr-T cg-T	2	$4_{10}-6$	$2_{10}-3$ $4_{10}-3$	$_{10}-5$ $2_{10}-5$	$_{10}-2$ $_{10}-2$	144 156	4211
9	mr-T cg-T	4	$9.5_{10}-7$	$3_{10}-5$ $1_{10}-3$	$7_{10}-6$ $_{10}-4$	$4_{10}-3$ $5_{10}-2$	155 155	6660

Porównanie kolumn (8) i (9) uwidacznia ogromną przewagę metod połączonych nad metodą T, szczególnie dla układów źle uwarunkowanych. Ponadto z przeprowadzonych eksperymentów wynika przewaga metody mr-T nad metodą cg-T.

Podkreślamy jeszcze raz, że dla zadań źle uwarunkowanych (por. przykłady), nie dostajemy dobrego przybliżenia wartości własnej λ_{min}, a tym samym oszacowanie błędu a posteriori jest zaniżone w stosunku do błędu rzeczywistego.

procedure *mr T (n,x,b,a1,a2,eps,s,macheps)*;
value *n,a2*;
real *a1,a2,eps,macheps*;
array *x,b*;
comment *Procedura mr-T rozwiązuje układ równań liniowych wysokiego stopnia o symetrycznej, dodatnio określonej macierzy A postaci:*

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

Procedura ta ponadto podaje przybliżenie najmniejszej wartości własnej macierzy A.
Parametry wejścia i wyjścia procedury mr-T:

n – *zmienna całkowita, wymiar macierzy A.*

x – *tablica jednowymiarowa o zakresie wskaźników od 1 do n. Na wejściu na x podstawiamy przybliżenie początkowe, na wyjściu x zawiera przybliżenie rozwiązania.*

b – *tablica jednowymiarowa o zakresie wskaźników od 1 do n, zawierająca wektor wyrazów wolnych.*

a1 – *zmienna rzeczywista, na wejściu wartość a1 dowolna, na wyjściu a1 jest równe przybliżeniu najmniejszej wartości własnej macierzy A.*

a2 – *zmienna rzeczywista równa największej wartości własnej macierzy A lub jej ograniczenie z góry.*

eps – *zmienna rzeczywista. Na wejściu eps oznacza ile razy mamy zredukować normę początkowego wektora residualnego* $\|\vec{r}_k\|_2 \leqslant eps \times \|\vec{r}_0\|_2$.
Na wyjściu eps jest równe normie ostatniego wektora residualnego: eps := $\|\vec{r}_k\|_2$.

s – *zmienna całkowita. Na wejściu s równa się maksymalnej ilości niezerowych elementów w wierszach macierzy A* $(1 \leqslant s \leqslant n)$.
Na wyjściu na zmiennej s zapamiętujemy:
$s = 0$ *redukcję błędu o eps:* $\|\vec{r}_k\|_2 \leqslant eps \times \|\vec{r}_0\|_2$,
$s = -1$ *macierz A nie dodatnio określona,*
$s = -2$ *macierz A numerycznie osobliwa,*
$s = -3$ *osiągnięcie granicznej maksymalnej dokładności,*

macheps – *zmienna rzeczywista, na wejściu równa* $2\uparrow(-t)$, *gdzie t – ilość cyfr mantysy w arytmetyce zmiennopozycyjnej dla danej maszyny. Na wyjściu oszacowanie błędu bezwzględnego* $\|\vec{x}^* - \vec{x}_k\|_2$ *(jeśli* $s = 0$ *lub* $s = -3$*).*
Ponadto w treści procedury mr-T występuje deklaracja procedury A, która wykonuje między innymi mnożenie wektora przez macierz układu.
Nagłówek procedury musi mieć postać: A(x,y,z,e,q,f,G), gdzie x,y,z – tablice jednowymiarowe o zakresie wskaźników od 1 do n,
e,q,f – zmienne rzeczywiste,
G – zmienna boolowska.
Po wykonaniu procedury A(x,y,z,e,q,f,G) powinniśmy otrzymać:

$$z := (e1 \times y - Ax)/q1,$$

gdzie

$$e1 := \textbf{if } G \textbf{ then } (Ax,x)/f \times q \textbf{ else } e,$$
$$q1 := \textbf{if } G \textbf{ then } (Ax,Ax)/(Ax,x) - e1 \textbf{ else } q.$$

Następnie

```
                f := (Ax,x),
                e := e1,
                q := q1;

begin integer i, k, k1;
    real a11, c, c1, c2, c3, c4, e, f, q, r0, r1, r2, t, t1, u, v, xn;
    array r, dr, dx [1:n];
    boolean G, H;
    comment włączenie deklaracji procedury A;
    a11 := a1 := a2/2;
    t := macheps × s × a2;
    t1 := t × t;
    H := false;
    comment metoda T na przedziale [a1, a2];
E1 :G := false;
    u := v := −1;
    A(x, b, r, u, v, f, G);
    r0 := 0;
    for i := 1 step 1 until n do
    begin
        r0 := r0 + r [i]↑2;
        dx [i] := 0
    end;
    c := sqrt (a2/a1);
    c := (c−1)/(c+1);
    c1 := (a2−a1)/2;
    c2 := c1×c1/4;
    c4 := (a1+a2)/2;
    c3 := eps×eps×r0;
    u := if eps < a11/a2 then eps else a11/a2;
    k1 := ln (u/2)/ln (c) −1;
    if k1 < 0 then k1 := 0;
    e := 0;
    for k := 0, k + 1 while k ⩽ k1 ∨ H do
    begin
        q := if k=0 then c4 else if k=1 then q−2× c2/q else c4−c2/q;
        xn := 0;
        for i := 1 step 1 until n do
        begin
            u := dx [i] := (e × dx [i] −r [i])/q;
            u := x [i] := x [i] + u;
            xn := xn + u × u
        end;
        u := v := −1;
        A (x, b, r, u, v, f, G);
        e := if k=0 then 2× c2/c4 else c2/(c4−e);
        r1 := 0;
        for i := 1 step 1 until n do
```

```
        r1 := r1 + r [i]↑2;
        if r1 ≤ c3 then
        begin
            s := 0;
            go to E3
        end redukcja błędu o eps;
        if H then
        begin
            if r1 ≥ r2 ∨ r1 ≤ .5 × t1 × xn then go to E3;
            r2 := r1
        end;
        if –, H ∧ r1 > r0 then
        begin
            s := if r1 ≤ n × t1 × xn then –3 else –1;
            go to E3
        end badania nie dodatniej określoności;
        if –, H ∧ r1 ≤ t1 × xn then
        begin
            s := –3;
            H := true;
            r2 := r1;
            a1 := a11;
            go to E1
        end granicznej dokładności;
    end metody T;
    comment metoda mr;
    q := 0;
    for i := 1 step 1 until n do
    dx [i] := dr [i] := 0;
    k := 0;
    r0 := r2 := r1;
    G := true;
E2: A (r, dr, dr, e, q, f, G);
    k := k+1;
    c1 := f/r1;
    if a11 > c1 then a11 := c1;
    if abs (c1) < t then
    begin
        s := –2;
        go to E3
    end badania numerycznej osobliwości;
    if c1 < –t then
    begin
        s := –1;
        go to E3
    end;
    if k=1 then
```

```
begin
    c2 := c1;
    c := sqrt (a2/(if c1 > a1 then a1 else c1));
    c4 := c := ((c - 1)/((c+1))↑2
end;
xn := r1 := 0;
for i := 1 step 1 until n do
begin
    u := dx [i] := (e× dx [i] -r [i])/q;
    u := x [i] := x [i] + u;
    xn := xn+u×u;
    u := r [i] := r [i] + dr [i];
    r1 := r1 + u × u
end;
if r1 < c3 then
begin
    s := 0;
    go to E3
end redukcja błędu o eps;
if -,H ∧ r1 < t1× xn then
begin
    s := -3;
    H := true
end;
if r1 > 4× c4× r0 ∨ r1 > (a2-a11)/(a2+a11)×r2 ∨ c1 > a1 then
begin
    eps := sqrt (c3/r1);
    a1 := if H then a11 else if k=1 then (a1+c2)/2
        else if c1 > a1 then a1 else c2;
    go to E1
end przejście do metody T na nowym przedziale [a1, a2];
c4 := c4 × c;
if c2 < c1 then c2 := c1;
r2 := r1;
go to E2;
comment wyniki metody mr-T;
E3 : eps := sqrt (r1);
a1 := a11;
if s=0 ∨ s=-3 then macheps := (eps + t × sqrt (xn))/a1
end procedury mr-T;
```

Bibliografia

[1] G. Golub, R. S. Varga, *Chebyshev semi iterative methods, succesive overrelaxation iterative methods*, Part I, II, Numerishe Math. 3 (1961), str. 147-169.

[2] C. Lanczos, *Solution of systems of Linear Equations by Minimized Iteration*, NBS, J. of Res. 49, (1952), str. 33-53.

[3] A. Kiełbasiński, *Algorytm sumowania z poprawkami i niektóre jego zastosowania*, Matematyka Stosowana I (1973), str. 23-41.

[4] M. A. Krasnosielski i inni, *Przybliżone rozwiązywanie równań operatorowych*, Moskwa 1969.

[5] H. Rutishauser, E. Stiefel i inni, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self – Adjoint Boundary Value Problem,* Chapter II (1959).
[6] E. Stiefel, *Kernel Polynomials in Linear Algebra and their Numerical Applications,* NBS. Appl. Math. Series 49, (1958), str. 1–22.
[7] E. L. Wachpress, *Iterative Solution of elliptic systems,* Prentice – Hall (1956).
[8] J. H. Wilkinson, *Błędy zaokrągleń w procesach algebraicznych,* Warszawa 1967.