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

Anna GÓRAJ, M. JANKOWSKI, A. KIEŁBASIŃSKI, H. WOŹNIAKOWSKI (Warszawa)

Oszacowanie błędu rozwiązania układu równań liniowych i zastosowanie poprawianego sumowania w algorytmach algebry liniowej

1. Rozwiązując nieosobliwy układ równań liniowych

$$A\vec{x} = \vec{b}, \qquad \text{dane: } A\ (n \times n),\ \vec{b}\ (n \times 1) \tag{1}$$

numerycznie stabilnym algorytmem, realizowanym w zmiennopozycyjnej arytmetyce (fl), otrzymujemy przybliżone rozwiązanie $\vec{x}$, zamiast prawdziwego rozwiązania $\vec{x}^* = A^{-1}\vec{b}$.

Wiadomo (por. [4], str. 130-155), że błąd tego przybliżenia możemy oszacować:

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

t — oznacza tu ilość cyfr mantysy liczby rzeczywistej w fl,

$$\|\vec{x}\| = \sqrt{\sum_i x_i^2}, \qquad \|A\| = \sqrt{\sum_{i,j} a_{ij}^2}.$$

Najmniejsza znana wartość K charakteryzuje stabilność użytego algorytmu.

Podobne do (2) zależności otrzymujemy dla algorytmów realizujących kumulację iloczynów skalarnych na rejestrze podwójnej precyzji (fl_2, por. [4], str. 37) lub z poprawianym sumowaniem ($\widetilde{\mathrm{fl}}$, por. [2]). Odpowiednie wielkości K są wtedy na ogół mniejsze. Wskaźnik K gra również istotną rolę w oszacowaniu błędu a posteriori:

$$\text{jeśli } \beta \overset{\mathrm{df}}{=} 2^{-t}\ \|A\| \cdot \|B\| \cdot K < 1, \tag{3}$$

$$\text{to } \|\vec{x}^* - \vec{x}\| \leqslant (\beta/(1-\beta)) \cdot \|\vec{x}\|,$$

B oznacza tu przybliżoną macierz A^{-1} (uzyskaną tym samym algorytmem co $\vec{x}$).

Poniższa tablica zawiera wartości wskaźnika K dla kilku częściej używanych metod. Dla ułatwienia porównań podajemy również ilość mnożeń (jako charakterystykę pracochłonności metody). Pomijamy człony z niższymi potęgami n oraz stałe multyplikatywne tam, gdzie nie grają one istotnej roli.

TABLICA 1

	Ilość mnożeń	K (fl)	K ($\widetilde{\mathrm{fl}}$)	K (fl_2)
Eliminacja z wyb. elem. głównego	$n^3/3$	$s_n \cdot n^3$	$s_n \cdot n^2$	$s_n \cdot n^1$
Householder	$2n^3/3$	n^2	n^1	n^1
Givens	$4n^3/3$	$n^{3/2}$	$n^{3/2}$	$n^{3/2}$
Schmidt (z popraw.)	$2n^3$	n^1	$n^{1/2}$	n^0
Banachiewicz ($A = A^T$, $A > 0$)	$n^3/6$	$n^{3/2}$	$n^{1/2}$	$n^{1/4}$

(Oszacowania te pochodzą z [5], a częściowo z [2] i [7]). Wielkość

$$s_n = \max_{i,j,k} |a_{ij}^{(k)}| / \max_{i,j} |a_{ij}^{(1)}|$$

(por. [4], str. 136-139) może być dokładnie znana a posteriori. Oszacowania a priori dla s_n są zbyt pesymistyczne.

Jak wiadomo z doświadczeń (por. np. [4], str. 152–155), wartości K podane w tablicy są na ogół zbyt duże, nierealistyczne. Wynika to między innymi ze statystycznej redukcji błędów, a niekiedy ze szczególnej korelacji błędów.

Wilkinson sugeruje zastąpienie we wzorze (3) wielkości K z tablicy przez jej pierwiastek kwadratowy w celu uzyskania bardziej realistycznej oceny (ale już nie oszacowania z góry) wielkości błędu. Ponieważ wyznaczanie przybliżonej odwrotności B macierzy A podwaja (lub nawet potraja) koszt całego procesu, więc wypada zapytać, czy możliwe jest uzyskanie tańszym kosztem równie dobrych (raczej – równie niedobrych) oszacowań błędu wytworzonego?

2. Pewną odpowiedzią na to pytanie jest próba kontroli „na bieżąco" dokładności algorytmu eliminacji z pełnym wyborem głównego elementu.

Przez „oszacowanie błędu" będziemy rozumieli tutaj dodatnią wielkość, w miarę posiadanej informacji – tego rzędu, co błąd (nie będącą na ogół ograniczeniem górnym błędu). W omawianym systemie kontroli dokładności stosowano następujące reguły wyznaczania takich oszacowań:

(i) jeśli α jest oszacowaniem błędu a, to $|c| \cdot \alpha$ jest oszacowaniem błędu $c \cdot a$;

(ii) jeśli α i β są oszacowaniami błędów a i b, to za oszacowanie błędu $a \pm b$ przyjmujemy

$$\max(\alpha, \beta) \quad \text{lub} \quad \sqrt{\alpha^2 + \beta^2}.$$

Korzystając z tych reguł można koszt kontroli dokładności algorytmu eliminacji sprowadzić do $2n^2$ mnożeń. Eksperymenty wykazały, że otrzymane oszacowania błędu rozwiązania były na ogół 10-100 razy większe od samych błędów.

Obszerniejszą informację o tym systemie kontroli dokładności oraz procedurę algolową „Gauss Control" można znaleźć w [1].

Liczne prace (por. np. [6]) poświęcone są metodom analizy przedziałowej. Koszt takiej kontroli dokładności sięga z reguły n^3 działań, a jakość uzyskiwanych oszacowań (lub błędów) nakazuje daleko idącą rezerwę do takiej realizacji zadania.

3. Jak widać z tablicy 1 zastąpienie arytmetyki fl arytmetyką fl_2 (lub $\widetilde{\mathrm{fl}}$) poprawia w niektórych przypadkach wydatnie charakterystykę stabilności algorytmu. Badania eksperymentalne różnego typu potwierdzają tę teoretyczną ocenę.

Realizacja arytmetyki fl_2 jest jednak w niektórych maszynach bardzo kosztowna (np. w typowym algorytmie algebry liniowej na maszynie GIER koszt wzrasta o 150%). Natomiast realizacja arytmetyki $\widetilde{fl}$ wydaje się stosunkowo prosta i mało kosztowna (na maszynie GIER koszt wzrasta o około 15%).

W przypadku zadań algebry liniowej arytmetyka $\widetilde{fl}$ różni się od fl jedynie bardziej skomplikowanym algorytmem obliczania iloczynu skalarnego dwu wektorów.

Jeśli chcemy obliczyć

$$p = \sum_{i=1}^{n} a_i \cdot b_i$$

(przy a_i, b_i danych numerycznie w fl), to należy posłużyć się następującym algorytmem Møllera (por. [3]), zapisanym w pesudo-algolu:

```
c := s := 0;
for i := 1 step 1 until n do
begin
    v := a_i × b_i;
    u := s + v;
    c := s − u + v + c;
    s := u
end;
    p := s + c;
```

Pewne istotne własności tego algorytmu są opisane w [2], tu ograniczymy się do podania przykładu eksperymentu numerycznego, wykonanego na maszynie GIER.

Przeprowadzono obliczenia wartości własnych licznych macierzy testowych. Na przykład, dla macierzy:

$$A_{80} = \begin{bmatrix} 80 & 79 & \dots & 1 \\ 79 & 79 & \dots & 1 \\ \dots & \dots & \dots & \dots \\ 1 & 1 & \dots & 1 \end{bmatrix}.$$

Transformacją Ortegi-Householdera sprowadzono A_{80} do postaci trójprzekątniowej, a następnie metodą bisekcji uzyskano przybliżenia wartości własnych (por. [8], [9], [11]).

Tablica 2 pokazuje przykładowo błędy otrzymanych kilku przybliżeń wartości własnych macierzy A_{80}.

TABLICA 2

Wartość własna	Błędy: fl	$\widetilde{fl}$	$fl_{4/3}$
$2.50095215_{10}-1$	$1.9_{10}-7$	$-7.8_{10}-8$	$-7.8_{10}-8$
$2.87776346_{10}-1$	$5.1_{10}-6$	$3.4_{10}-8$	$-1.5_{10}-7$
$6.12980701_{10}-1$	$2.4_{10}-6$	$-4.1_{10}-8$	$-4.1_{10}-8$
105.137230	$1.4_{10}-5$	$1.2_{10}-6$	$7.2_{10}-7$

Badania teoretyczne i eksperymentalne wydają się wskazywać na szczególną użyteczność arytmetyki $\widetilde{fl}$ (lub fl_2) w algorytmach wykorzystujących rozkład unitarno-trójkątny (bądź inne rozkłady pokrewne) macierzy. Ta klasa algorytmów z kolei jest obecnie szeroko stosowana do konstrukcji uniwersalnych procedur bibliotecznych.

Ze względu na niesymetryczne zaokrąglenia w arytmetyce GIER - algolu zjawisko kumulacji błędów występuje tu w sposób szczególnie dobitny i korzyść ze stosowania arytmetyki $\widetilde{fl}$ jest tu oczywista. Uważaliśmy więc za wskazane opracować wersje procedur biblioteki programów ZON pracujące w arytmetyce $\widetilde{fl}$.

Prace te, wraz z testowaniem, są w toku (zmiany i ulepszenia dotyczą często nie tylko przejścia od arytmetyki fl do $\widetilde{fl}$). W spisie na końcu tego artykułu podajemy kilka tak dopracowanych procedur.

Uważamy, że przeniesienie tych procedur na inne maszyny niż GIER (np. na ODRĘ 1204) wymaga przeprowadzenia odpowiednich badań.

Bibliografia

[1] A. Kiełbasiński, *Oszacowania błędu w metodzie eliminacji*, Matematyka Stosowana I (1973), str. 9–21.

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

[3] O. Møller, *On a quasi double-precision in floating point addition*, BIT 5 (1965), str. 37-50, 251-255.

[4] J. H. Wilkinson, *Błędy zaokrągleń w procesach algebraicznych*, Warszawa 1967.

[5] J. H. Wilkinson, *The algebraic eigenvalue problem*, Clarendon Press - Oxford 1965.

[6] P. Wisskirchen, *Ein Steuerung Prinzip der Intervallrechnung u.d. Anwendungen auf d. Gaussschen Algorithmus*, Bonn 1969.

[7] H. Woźniakowski, *Analiza algorytmu ortogonalizacji Schmidta dla układu równań liniowych*, Sprawozdania IMM i ZON UW (w przygotowaniu).

Procedury

[8] M. Jankowski, *Procedury Kombajn* 1, 2, s2. *Wyznaczanie wartości i wektorów własnych symetrycznych macierzy, stosując trójdiagonalizację metodą Householdera, metodę bisekcji i Wielandta*, ZON UW 171, 172, 162.

[9] M. Jankowski, *Procedura OL* – przeróbka procedur tred 2 oraz imtql 2 – Num. Math. 11, 12, z włączeniem arytmetyki $\widetilde{fl}$, ZON UW 221.

[10] M. Jankowski, *Procedura OR real* – przeróbka procedur: balance, orthes, hqr 2, ortbak z Num. Math. 12-16 z włączeniem arytmetyki $\widetilde{fl}$, ZON UW – 219.

[11] A. Góraj, *Procedura bisect* – przeróbka procedury bisect z Num. Math. 9, ZON UW 224.