ROCZNIKI POLSKIEGO TOWARZYSTWA MATEMATYCZNEGO
Seria III: MATEMATYKA STOSOWANA VIII (1976)

ANDRZEJ KIEŁBASIŃSKI (Warszawa) i KRYSTYNA ZIĘTAK (Wrocław)

Analiza numeryczna typowych zadań z unitarnym przekształceniem Householdera

(*Praca przyjęta do druku 24.4.1975*)

0. Wstęp. W pracy tej przedstawiamy dowód numerycznej poprawności unitarnego przekształcenia Householdera. Rozważamy różne arytmetyki numeryczne rzeczywiste lub zespolone oraz typowe lub ogólne warunki definiujące transformację. Pewne szczegóły tej analizy wydają się nam nowe, zasadnicza jednak jej struktura pochodzi z pracy J. H. Wilkinsona [6].

Główną zatem przyczyną naszej publikacji jest chęć spopularyzowania tej ważnej transformacji oraz metod analizy Wilkinsonowskiej.

Numeracja wzorów jest oddzielna w każdym z paragrafów. Cytując wzory podane w innym paragrafie podajemy dwie liczby: numer paragrafu i numer wzoru.

1. Zespolone i rzeczywiste przekształcenia Householdera. Niech K^n oznacza zespoloną lub rzeczywistą kartezjańską przestrzeń wektorów, utożsamianych z macierzami jednokolumnowymi, z normą euklidesową:

$$||\vec{x}|| = (\vec{x}^H \vec{x})^{1/2}, \qquad \vec{x} \in K^n.$$

Unitarnym przekształceniem Householdera nazywamy przekształcenie dające się zapisać w postaci

$$P = I - \vec{w} \cdot 2 \cdot \vec{w}^H, \qquad ||\vec{w}|| = 1. \tag{1}$$

Celowo zastosowaliśmy tu zapis postaci $\vec{w} \cdot 2 \cdot \vec{w}^H$, gdyż jest to poprawnie określony iloczyn macierzy wymiarów $n \times 1$, 1×1, $1 \times n$ (przez B^H rozumiemy macierz sprzężoną z macierzą B).

Przekształcenie (1) przyporządkowuje wektorowi $\vec{x} \in K^n$ jego odbicie zwierciadlane $\vec{y}$ względem hiperpłaszczyzny prostopadłej do wektora $\vec{w}$. Istotnie,

$$\vec{y} = P\vec{x} = \vec{x} - \vec{r} \cdot 2,$$

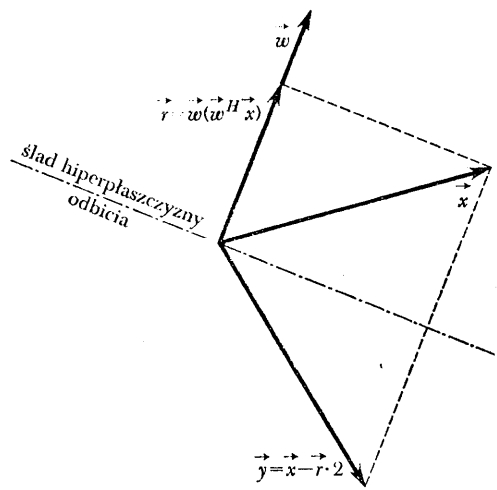
gdzie

$$\vec{r} = \vec{w}(\vec{w}^H \vec{x})$$

jest rzutem prostopadłym wektora $\vec{x}$ na wektor $\vec{w}$.

Unitarne przekształcenia Householdera nazywamy więc niekiedy *przekształceniami odbić zwierciadlanych*. Z tej interpretacji geometrycznej natychmiast wynika unitarność przekształcenia P (tzn. $\forall \vec{x} \in K^n$, $||P\vec{x}|| = ||\vec{x}||$) oraz dalsze, łatwo sprawdzalne własności:

$$PP = I, \quad P = P^{-1} = P^H.$$



Rys. 1 (płaszczyzna rysunku rozpięta na wektorach $\vec{w}$ i $\vec{x}$)

Możemy również sprawdzić, że $\det(P) = -1$. Rozważmy bowiem macierz

$$B = [\vec{w}, \vec{a}_2, \vec{a}_3, \ldots, \vec{a}_n],$$

gdzie $\vec{a}_2, \vec{a}_3, \ldots, \vec{a}_n$ są dowolnymi liniowo niezależnymi wektorami, rozpinającymi hiperpłaszczyznę prostopadłą do wektora $\vec{w}$. Wówczas $\det(B) \neq 0$.

Zbadajmy wyznacznik macierzy

$$PB = [P\vec{w}, P\vec{a}_2, \ldots, P\vec{a}_n].$$

Ponieważ $P\vec{w} = -\vec{w}$, $P\vec{a}_i = \vec{a}_i$, więc

$$\det(PB) = \det[-\vec{w}, \vec{a}_2, \ldots, \vec{a}_n] = -\det(B).$$

Wobec tego, że $\det(PB) = \det(P) \cdot \det(B)$, otrzymujemy $\det(P) = -1$. Przekształcenie Householdera może być oczywiście zapisane również za pomocą dowolnego wektora niezerowego $\vec{u}$: $P = I - \vec{u}\frac{2}{||\vec{u}||^2}\vec{u}^H$, gdy $\vec{w} = \vec{u}/||\vec{u}|| \cdot c$, $|c| = 1$, $c \in K$.

2. Podstawowe zadanie (H) z przekształceniem Householdera. Rozważmy następujące zadanie:

Dla danej pary niezerowych wektorów zespolonych $\vec{a}$ i $\vec{e}$ należących do K^n chcemy wyznaczyć przekształcenie unitarne Householdera, które przeprowadza wektor $\vec{a}$ na wektor $P\vec{a}$ o kierunku wektora $\vec{e}$, tzn.

$$P\vec{a} = \vec{e}k, \quad |k| = ||\vec{a}||/||\vec{e}||. \tag{1}$$

Ponieważ przekształcenie P jest hermitowskie, więc wyrażenie $\vec{a}^H P\vec{a} = \vec{a}^H\vec{e}k$ musi być rzeczywiste. Okazuje się (zob. [6], str. 59), że jest to zarazem warunek dostateczny istnienia przekształcenia spełniającego (1.1) i (1). Zatem należy przyjąć

$$k = \begin{cases} \pm \dfrac{\vec{e}^H\vec{a}}{|\vec{e}^H\vec{a}|}\dfrac{||\vec{a}||}{||\vec{e}||}, & \text{gdy} \quad \vec{e}^H\vec{a} \neq 0, \\ \pm \dfrac{||\vec{a}||}{||\vec{e}||}, & \text{gdy} \quad \vec{e}^H\vec{a} = 0. \end{cases} \tag{2}$$

Z wzoru (1) wynika

$$\vec{w} = (\vec{a}-\vec{e}k)/||\vec{a}-\vec{e}k||.$$

Niech

$$\vec{u} = \vec{a}-\vec{e}k, \tag{3}$$

$$R = ||\vec{u}||^2/2. \tag{4}$$

Wówczas przekształcenie (1.1) możemy zapisać w postaci

$$P = I-\vec{u}\frac{1}{R}\vec{u}^H. \tag{5}$$

Parametr k określiliśmy z dokładnością do znaku. Znak ten wybieramy tak, by wektor $\vec{u}$ miał największą długość. Łatwo sprawdzić, że

$$R = ||\vec{a}||^2 \mp |\vec{e}^H\vec{a}|\,||\vec{a}||/||\vec{e}||. \tag{6}$$

Wobec tego we wzorach (2) wybieramy znak minus. A więc z (2), (3), (6) mamy

$$\vec{u} = \begin{cases} \vec{a}+\vec{e}\dfrac{\vec{e}^H\vec{a}}{|\vec{e}^H\vec{a}|}\dfrac{||\vec{a}||}{||\vec{e}||}, & \text{gdy} \quad \vec{e}^H\vec{a} \neq 0, \\ \vec{a}+\vec{e}\dfrac{||\vec{a}||}{||\vec{e}||}, & \text{gdy} \quad \vec{e}^H\vec{a} = 0, \end{cases} \tag{7}$$

$$R = ||\vec{a}||^2+|\vec{e}^H\vec{a}|\,||\vec{a}||/||\vec{e}||. \tag{8}$$

Taki wybór znaku uzasadnimy później. Na razie ograniczymy się do uwagi, że zapewnia on numeryczną poprawność (zob. [3]) najważniejszych algorytmów korzystających z przekształcenia Householdera.

Z (2), (4), (6) otrzymujemy dla przypadku rzeczywistych wektorów $\vec{a}$, $\vec{e}$ następujące wzory na rzeczywiste przekształcenie Householdera:

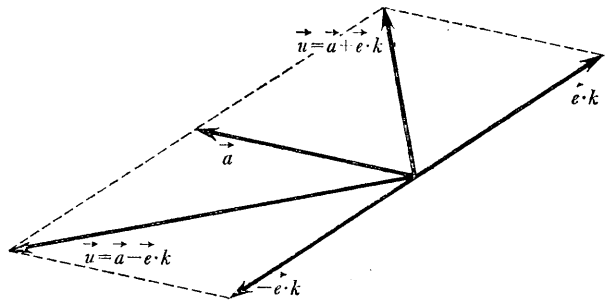
$$\vec{u} = \begin{cases} \vec{a}+\vec{e}\,\mathrm{sign}\,(\vec{e}^T\vec{a})\,||\vec{a}||/||\vec{e}||, & \text{gdy} \quad \vec{e}^T\vec{a} \neq 0, \\ \vec{a}+\vec{e}\,||\vec{a}||/||\vec{e}||, & \text{gdy} \quad \vec{e}^T\vec{a} = 0, \end{cases} \tag{9}$$

$$R = ||\vec{a}||^2+|\vec{e}^T\vec{a}|\,||\vec{a}||/||\vec{e}||. \tag{10}$$

Konstrukcja rzeczywistego przekształcenia Householdera w rozważanym zadaniu ma prostą interpretację geometryczną. Mianowicie, wektor $\vec{u}$ leży na dwusiecznej kąta między wektorami $\vec{a}$ i $-\vec{e}$ lub $\vec{a}$ i $\vec{e}$ (patrz rys. 2).

Dla powiększenia dokładności przekształcenia, wybieramy dwusieczną mniejszego kąta (spośród kątów między wektorami $\vec{a}$ i $\vec{e}$ oraz $\vec{a}$ i $-\vec{e}$). Wówczas kierunek odpowiedniej dwusiecznej jest „lepiej wyznaczony".

Podstawowym zadaniem korzystającym z przekształcenia Householdera jest wyznaczenie wektora $\vec{c} = P\vec{b}$ dla danego wektora $\vec{b}$ i danych wektorów $\vec{a}$, $\vec{e}$, defi-



Rys. 2 (płaszczyzna rysunku rozpięta na wektorach $\vec{a}$ i $\vec{e}$)

niujących przekształcenie P. Zadanie to nazywać będziemy *zadaniem* (H). W dalszej części pracy wykażemy, że wyznaczenie wektora $\vec{c}$ może być wykonane w sposób numerycznie poprawny, gdy przekształcenie P jest określone zależnościami (5), (7), (8). Chcąc wyznaczyć obraz $\vec{c}$ dowolnego wektora zespolonego $\vec{b}$, nie musimy w sposób efektywny konstruować macierzy przekształcenia (5), gdyż

$$\vec{c} = P\vec{b} = \vec{b} - \vec{u}\left(\frac{1}{R}\vec{u}^H\vec{b}\right). \tag{11}$$

Wygodnie jest zatem reprezentować przekształcenie P współrzędnymi wektora $\vec{u}$ i liczbą R. Reprezentacja P zajmuje wtedy tylko około n miejsc (liczb rzeczywistych lub zespolonych) w pamięci maszyny (macierz P zajęłaby około $n^2/2$ miejsc) zaś koszt otrzymania $\vec{c}$ ze wzoru (11) jest rzędu tylko $2n$ mnożeń. (Zamiast n^2 przy posługiwaniu się pełną macierzą P.)

3. Zespolona arytmetyka zmiennopozycyjna cfl, cfl_2, $\widetilde{\mathrm{cfl}}$. Zakładamy, że dysponujemy dwójkową t-cyfrową arytmetyką zmiennopozycyjną fl (zob. np. [5], str. 16, [6], str. 110) określoną tak, że obliczone wyniki operacji arytmetycznych i pierwiastkowania mogą być zapisane w postaci:

$$\mathrm{fl}(x \Diamond y) = (x \Diamond y)(1-\varepsilon), \qquad |\varepsilon| \leqslant 2^{-t},$$

$$\mathrm{fl}(\sqrt{x}) = \sqrt{x}\,(1-\delta), \qquad |\delta| \leqslant p \cdot 2^{-t},$$

gdzie $\Diamond$ oznacza dowolny z operatorów $\pm$, $\times$, $/$, liczba p zaś należy na ogół do przedziału $[1, 2]$. Będziemy stosować relację $\overset{1}{\leqslant}$ (wprowadzoną w [2] lub w [4]), co pozwala na pomijanie w oszacowaniach błędów mniej istotnych składników rzędu 2^{-2t}, 2^{-3t}, itd. w obecności składników rzędu 2^{-t}.

Niech $z_k = x_k + \bar{i}y_k$ (x_k, y_k rzeczywiste, $\bar{i}$ — jednostka urojona). Wówczas dodawanie w arytmetyce cfl jest wykonane zgodnie z regułą

$$\begin{aligned}\mathrm{cfl}(z_1+z_2) &= \mathrm{fl}(x_1+x_2)+\bar{i}\,\mathrm{fl}(y_1+y_2) = \\ &= (x_1+x_2)(1-\varepsilon_1)+\bar{i}(y_1+y_2)(1-\varepsilon_2) = (z_1+z_2)(1-\varepsilon_3),\end{aligned}$$

gdzie $|\varepsilon_j| \leqslant 2^{-t}$ ($\varepsilon_1, \varepsilon_2$ — rzeczywiste, ε_3 na ogół zespolona).

Podobnie wykonywane są w cfl pozostałe podstawowe operacje. Mianowicie, łatwo pokazać, że

$$\begin{aligned}\mathrm{cfl}(z_1z_2) &= (z_1z_2)(1-\varepsilon), && |\varepsilon| \leqslant_1 (1+\sqrt{2})2^{-t},\\ \mathrm{cfl}(z_1/z_2) &= (z_1/z_2)(1-\varphi), && |\varphi| \leqslant_1 (4+\sqrt{2})2^{-t},\\ \mathrm{cfl}(|z_1|) &= |z_1|(1-\nu), && |\nu| \leqslant_1 (p+2{,}25)2^{-t},\\ \mathrm{cfl}(rz_1) &= (rz_1)(1-\delta), && |\delta| \leqslant 2^{-t},\end{aligned} \tag{1}$$

gdzie r jest liczbą rzeczywistą, a $|z_1|$, z_1/z_2 obliczamy według następujących wzorów (por. [7]):

$$|z_1| = \begin{cases} 0, & \text{gdy} \quad x_1 = x_2 = 0,\\ |x_1|\sqrt{1+(y_1/x_1)^2}, & \text{gdy} \quad |x_1| \geqslant |y_1|,\\ |y_1|\sqrt{1+(x_1/y_1)^2}, & \text{gdy} \quad |x_1| < |y_1|, \end{cases}$$

$$z = \frac{z_1}{z_2} = \begin{cases} \dfrac{x_1+y_1h}{x_2+y_2h}+\bar{i}\,\dfrac{y_1-x_1h}{x_2+y_2h}, & h = \dfrac{y_2}{x_2}, \quad \text{gdy} \quad |x_2| \geqslant |y_2|,\\ \dfrac{x_1h+y_1}{x_2h+y_2}+\bar{i}\,\dfrac{y_1h-x_1}{x_2h+y_2}, & h = \dfrac{x_2}{y_2}, \quad \text{gdy} \quad |x_2| < |y_2|. \end{cases}$$

Algorytmy te są „bezpieczne". Pozwalają na uniknięcie niedomiaru lub nadmiaru, o ile dane i wyniki są dobrze reprezentowane w cfl. Dla przykładu podamy analizę numeryczną algorytmu dzielenia. Ze względu na „symetrię" algorytmu wystarczy zbadać jeden przypadek. Np. $|x_2| \geqslant |y_2|$. Niech

$$h = y_2/x_2, \quad m = x_2+y_2h, \quad u = \frac{x_1+y_1h}{m}, \quad v = \frac{y_1-x_1h}{m}.$$

Wówczas

$$\tilde{z} = \mathrm{cfl}\left(\frac{z_1}{z_2}\right) = \frac{x_1+y_1h(1-\delta_1)}{x_2+y_2h(1-\delta_2)}(1-\delta_3)+\bar{i}\,\frac{y_1-x_1h(1-\delta_4)}{x_2+y_2h(1-\delta_2)}(1-\delta_5), \tag{2}$$

gdzie

$$|\delta_1|, |\delta_2|, |\delta_4| \leqslant_1 2\cdot 2^{-t}, \quad |\delta_3|, |\delta_5| \leqslant_1 3\cdot 2^{-t}.$$

Ponieważ $\mathrm{sign}(x_2) = \mathrm{sign}(y_2h)$, a ponadto $|y_2h/m| \leqslant \frac{1}{2}$, więc

$$x_2+y_2h(1-\delta_2) = m\left(1-\frac{y_2h}{m}\delta_2\right) = m(1-\varphi_2), \quad |\varphi_2| \leqslant_1 2^{-t}.$$

Zatem z (2) mamy

$$\tilde{z} = \frac{1}{m}\{[x_1 + y_1 h(1-\delta_1)](1-\varphi_3) + \bar{i}[y_1 - x_1 h(1-\delta_4)](1-\varphi_5)\}, \quad |\varphi_i| \underset{1}{\leqslant} 4\cdot 2^{-t}.$$

(3)

Niech

$$w = \frac{1}{m}[(x_1 + y_1 h)(1-\varphi_3) + i\,(y_1 - x_1 h)(1-\varphi_5)].$$

Wówczas z (3) otrzymujemy

$$\frac{|z-w|^2}{|z|^2} = \frac{(x_1+y_1h)^2\varphi_3^2 + (y_1-x_1h)^2\varphi_5^2}{(x_1+y_1h)^2 + (y_1-x_1h)^2} \underset{1}{\leqslant} 16\cdot 2^{-2t},$$

$$\frac{|\tilde{z}-w|^2}{|z|^2} = \frac{y_1^2h^2\delta_1^2(1-\varphi_3)^2 + x_1^2h^2\delta_4^2(1-\varphi_5)^2}{(x_1+y_1h)^2 + (y_1-x_1h)^2} \underset{1}{\leqslant} \frac{h^2}{1+h^2}\,4\cdot 2^{-2t} \leqslant 2\cdot 2^{-2t}.$$

Stąd, wykorzystując nierówność $|\tilde{z}-z| \leqslant |\tilde{z}-w| + |w-z|$, otrzymujemy

$$\frac{|\tilde{z}-z|}{|z|} \underset{1}{\leqslant} (4+\sqrt{2})2^{-t},$$

co należało pokazać.

Zbadamy teraz błąd wytworzony przy obliczaniu w arytmetyce cfl iloczynu skalarnego dwóch wektorów $\vec{w}$ i $\vec{z}$ o składowych w_j, z_j. Niech

$$s = \sum_{j=1}^{n} z_j \overline{w}_j. \tag{4}$$

Łatwo pokazać, że (por. [5], str. 30 i następne)

$$\tilde{s} = \mathrm{cfl}\Big(\sum_{j=1}^{n} z_j\overline{w}_j\Big) = \Big[\sum_{j=1}^{n} z_j\overline{w}_j(1-\xi_j)\Big](1-\xi), \tag{5}$$

$$|\xi_j| \underset{1}{\leqslant} (n-j+1+\sqrt{2})2^{-t}, \quad |\xi| \leqslant 2^{-t}.$$

A więc obliczony wynik $\tilde{s}$ jest zaokrąglonym iloczynem skalarnym nieco zniekształconych wektorów $\vec{w}$ lub $\vec{z}$, gdyż czynnik $(1-\xi_j)$ możemy uważać za zaburzenie współrzędnej w_j lub z_j. Dla dowolnych wektorów $\vec{w}$ i $\vec{z}$ nie możemy twierdzić, że iloczyn skalarny jest obliczony z małym błędem względnym. Dużą dokładność względną obliczania $\vec{w}^H\vec{z}$ otrzymujemy w szczególnych przypadkach, np. gdy $\vec{w} = \vec{z}$. Mianowicie,

$$\mathrm{cfl}(\vec{z}^H\vec{z}) = \mathrm{fl}\Big(\sum_{j=1}^{n}[(\mathrm{re}\,z_j)^2 + (\mathrm{im}\,z_j)^2]\Big) = (\vec{z}^H\vec{z})(1-\varphi), \tag{6}$$

$$|\varphi| \underset{1}{\leqslant} (n+1)2^{-t}.$$

Zajmiemy się teraz arytmetyką cfl_2 i $\mathrm{cf\tilde{l}}$. Zakładamy, że czytelnik zna zasady arytmetyki fl_2 (zob. [5], str. 37) i $\mathrm{f\tilde{l}}$ (zob. [1]). Niech

$$u = \sum_{j=1}^{n} u_j = \sum_{j=1}^{n} (a_j + ib_j).$$

Wówczas w arytmetyce cfl_2 mamy

$$\tilde{u} = \mathrm{cfl}_2\Big(\sum_{j=1}^{n} (a_j+ib_j)\Big) = \mathrm{fl}_2\Big(\sum_{j=1}^{n} a_j\Big) + i\,\mathrm{fl}_2\Big(\sum_{j=1}^{n} b_j\Big) = \Big[\sum_{j=1}^{n} z_j(1-\psi_j)\Big](1-\psi), \tag{7}$$

$$|\psi_j| \underset{1}{\leqslant} \tfrac{3}{2}(n-j+1)2^{-2t}, \quad |\psi| \underset{1}{\leqslant} 2^{-t}.$$

Dla arytmetyki $\mathrm{cf\tilde{l}}$ otrzymujemy

$$\tilde{u} = \mathrm{cf\tilde{l}}\Big(\sum_{j=1}^{n} (a_j+ib_j)\Big) = \Big[\sum_{j=1}^{n} z_j(1-\psi_j)\Big](1-\psi_0), \quad |\psi_j| \underset{1}{\leqslant} 2^{-t}. \tag{8}$$

Ostatecznie, łatwo pokazać, że dla iloczynu (4) obliczonego w arytmetyce cfl_2 i $\mathrm{cf\tilde{l}}$ otrzymujemy zależności (por. (5), (7), (8)):

$$\tilde{s} = \mathrm{cfl}_2\Big(\sum_{j=1}^{n} z_j\overline{w}_j\Big) = \Big(\sum_{1}^{n} z_j\overline{w}_j(1-\varphi_j)\Big)(1-\varphi), \tag{9}$$

$$|\varphi_j| \underset{1}{\leqslant} \tfrac{3}{2}\big(n-j+1+\sqrt{2}\big)2^{-2t}, \quad |\varphi| \underset{1}{\leqslant} 2^{-t},$$

$$\tilde{s} = \mathrm{cf\tilde{l}}\Big(\sum_{j=1}^{n} z_j\overline{w}_j\Big) = \Big[\sum_{1}^{n} z_j\overline{w}_j(1-\psi_j)\Big](1-\psi), \tag{10}$$

$$|\psi_j| \underset{1}{\leqslant} \big(2+\sqrt{2}\big)2^{-t}, \quad |\psi| \underset{1}{\leqslant} 2^{-t}.$$

Natomiast, jeśli $\vec{z} = \overline{\vec{w}}$, to mamy (por. (6))

$$\mathrm{cfl}_2(\vec{z}^H\vec{z}) = (\vec{z}^H\vec{z})(1-\sigma), \quad |\sigma| \underset{1}{\leqslant} 2^{-t}, \tag{11}$$

$$\mathrm{cf\tilde{l}}(\vec{z}^H\vec{z}) = (\vec{z}^H\vec{z})(1-\psi), \quad |\psi| \underset{1}{\leqslant} 3\cdot 2^{-t}. \tag{12}$$

Wobec tego kwadrat normy wektora $\vec{z}$, a więc również jego norma, jest obliczany w arytmetykach cfl_2, $\mathrm{cf\tilde{l}}$ z dużą dokładnością względną, niezależną od wymiaru n przestrzeni.

Pojęcie arytmetyki fl_2, a więc i cfl_2, nie jest ściśle określone. Najistotniejszą cechą arytmetyk objętych tą nazwą jest opisana powyżej zdolność kumulowania iloczynów skalarnych na rejestrze podwójnej precyzji. W niektórych emc jest to jednak kosztowne, więc stosuje się kumulację jedynie w przypadku sum (iloczynów skalarnych) wielkiej liczby składników. Jeśli dostęp do rejestru podwójnej precyzji nie jest kosztowny, to możemy wykorzystywać go również do kumulacji nawet sum kilku składników, do operacji dzielenia z dzielną daną w podwójnej precyzji oraz do pierwiastkowania argumentu danego w podwójnej precyzji (por. [5], str. 38). Będziemy w takim przypadku mówili o *pełnej arytmetyce* fl_2 (cfl_2).

4. Zespolone zadanie (H) **dla** $\vec{e} = \vec{e}_1$. Niech a_j, b_j, u_j oznaczają składowe wektorów $\vec{a}$, $\vec{b}$, $\vec{u}$. Zakładamy, że składowe $\vec{a}$ i $\vec{b}$ nie są obarczone żadnym błędem i są równe swoim reprezentacjom w arytmetyce cfl. Niech

$$\vec{e} = \vec{e}_1 = (1, 0, 0, \ldots, 0)^T.$$

Jest to przypadek o szczególnie ważnych zastosowaniach. Wzory (2.7), (2.8) można obecnie przekształcić do postaci (zob. [6], str. 308)

$$R = A+S, \qquad u_1 = a_1\left(1+\frac{A}{S}\right), \qquad u_j = a_j \quad (j = 2, 3, \ldots, n),$$

gdzie

$$A = \sum_{j=1}^{n} |a_j|^2, \qquad S = \sqrt{|a_1|^2 A}.$$

Wyznaczając R i u_1 z powyższych wzorów (zamiast z zależności

$$R = \|\vec{a}\|^2 + |a_1|\,\|\vec{a}\|, \qquad u_1 = a_1\left(1+\frac{\|\vec{a}\|}{|a_1|}\right))$$

wykonujemy o jedno pierwiastkowanie mniej.

Postaramy się podać teraz realistyczne oszacowania błędów wytworzonych przy numerycznej realizacji zespolonego zadania (H). Przez $\tilde{F}$ będziemy oznaczać numerycznie obliczoną wartość wyrażenia F (w pseudo-algolu możemy więc zapisać $\tilde{F} := F$), a $\hat{\varphi}2^{-t}$ oznacza oszacowanie błędu φ (tzn. $|\varphi| \underset{1}{\leqslant} \hat{\varphi}2^{-t}$).

Niech α, β, ϱ, μ oznaczają błędy względne, z jakimi są wyznaczone w danej arytmetyce wielkości $\tilde{A}$, $\tilde{S}$, $\tilde{R}$, $\tilde{u}_1$, tzn.

$$(1) \qquad \tilde{A} = A(1-\alpha), \quad \tilde{S} = S(1-\sigma), \quad \tilde{R} = R(1-\varrho), \quad \tilde{u}_1 = u_1(1-\mu),$$
$$\vec{v} = [\tilde{u}_1, u_2, \ldots, u_n].$$

Wektor $\vec{v}$ jest identyczny z numerycznie wyznaczonym wektorem $\tilde{\vec{u}}$. Zbadajmy związki zachodzące pomiędzy błędami α, β, ϱ, μ. Pozwoli nam to lepiej oszacować normę macierzy błędów

$$(2) \qquad B = \tilde{P}-P, \qquad \tilde{P} = I-\vec{v}\frac{1}{R}\vec{v}^H$$

niż gdybyśmy próbowali od razu korzystać ze związków pomiędzy oszacowaniami $\hat{\alpha}$, $\hat{\sigma}$, $\hat{\varrho}$, $\hat{\mu}$. Łatwo sprawdzić, że

$$\sigma = \tfrac{1}{2}\alpha+\varepsilon_1+\Delta_1, \qquad |\varepsilon_1| \leqslant (1{,}5+p)2^{-t},$$

$$(3) \qquad \mu = \frac{\frac{1}{2}\alpha-\varepsilon_1+\varepsilon_2}{1+h}+\varepsilon_3+\Delta_2, \qquad |\varepsilon_2| \leqslant 2^{-t}, \qquad |\varepsilon_3| \underset{1}{\leqslant} 2\cdot 2^{-t},$$

$$\varrho = \frac{\alpha+(\frac{1}{2}\alpha+\varepsilon_1)h}{1+h}+\varepsilon_4+\Delta_3, \qquad |\varepsilon_4| \leqslant 2^{-t},$$

gdzie

$$h = \frac{|a_1|}{||\vec{a}||},$$

a błędy Δ_1, Δ_2, Δ_3 mają oszacowania na poziomie 2^{-2t}. Błędy α, ε_1, ε_2, ε_4, σ, ϱ są w istocie wyrażeniami rzeczywistymi, chociaż rozważamy tutaj zespolony przypadek zadania (H).

Zbadamy teraz, jak zaburzenia wektora $\vec{u}$ i liczby R wpływają na przekształcenie P. Z (1) i (2) łatwo otrzymamy następujące wyrażenie dla B:

$$B = \frac{1}{R}[-\vec{u}\vec{u}^H\varrho_1 + \vec{u}\vec{e}_1^T\bar{u}_1\bar{\mu} + \vec{e}_1\vec{u}^H u_1\mu - \vec{e}_1\vec{e}_1^T u_1\bar{u}_1\mu\bar{\mu}], \tag{4}$$

gdzie

$$\varrho_1 = \varrho + \varrho^2 + \varrho^3 + \ldots = \varrho + \Delta_4, \quad |\Delta_4| \leqslant 2^{-2t},$$

$\bar{u}_1$, $\bar{\mu}$ są to liczby zespolone sprzężone z liczbami u_1, μ. Macierz B jest więc hermitowska i ma następującą budowę:

(5)

$$B = \frac{1}{R}\begin{pmatrix} (\mu+\bar{\mu}-\mu\bar{\mu}-\varrho_1)u_1\bar{u}_1 & (\mu-\varrho_1)u_1\bar{u}_2 & (\mu-\varrho_1)u_1\bar{u}_3 & \ldots & (\mu-\varrho_1)u_1\bar{u}_n \\ (\bar{\mu}-\varrho_1)u_2\bar{u}_1 & -\varrho_1 u_2\bar{u}_2 & -\varrho_1 u_2\bar{u}_3 & \ldots & -\varrho_1 u_2\bar{u}_n \\ (\bar{\mu}-\varrho_1)u_3\bar{u}_1 & -\varrho_1 u_3\bar{u}_2 & -\varrho_1 u_3\bar{u}_3 & \ldots & -\varrho_1 u_3\bar{u}_n \\ \ldots\ldots & \ldots\ldots & \ldots\ldots & \ldots & \ldots\ldots \\ (\bar{\mu}-\varrho_1)u_n\bar{u}_1 & -\varrho_1 u_n\bar{u}_2 & -\varrho_1 u_n\bar{u}_3 & \ldots & -\varrho_1 u_n\bar{u}_n \end{pmatrix}.$$

Skorzystamy teraz z tej szczególnej postaci B. Zauważmy, że pomijając elementy pierwszego wiersza, wszystkie kolumny tej macierzy są równoległe do wektora $(0, u_2, u_3, \ldots, u_n)^T$. Jeśli więc pomnożymy lewostronnie macierz B przez macierz odbicia zwierciadlanego G, przeprowadzającego $(n-1)$-wymiarowy wektor $[u_2, u_3, \ldots, u_n]^T$ na wektor o kierunku wektora $\vec{e}_2$, to otrzymamy macierz GB, mającą niezmieniony pierwszy wiersz, zaś wiersz drugi postaci

$$\frac{1}{R}(U\bar{u}_1(\mu-\varrho_1) \quad -U\bar{u}_2\varrho_1 \quad \ldots \quad -U\bar{u}_n\varrho_1),$$

gdzie

$$U = \left(\sum_{j=2}^{n} |u_j|^2\right)^{1/2}.$$

Następne wiersze zawierają wyłącznie elementy zerowe. Zatem wszystkie wiersze macierzy GB, z pominięciem elementów pierwszej kolumny, są bądź równoległe do wektora $(0, \bar{u}_2, \bar{u}_3, \ldots, \bar{u}_n)^T$, bądź składają się z samych zer. Zatem macierz GBG^H ma wszystkie elementy równe zeru z wyjątkiem czterech elementów w lewym górnym rogu. Ten niezerowy blok ma postać

$$C = \frac{1}{R}\begin{pmatrix} (\mu+\bar{\mu}-\mu\bar{\mu}-\varrho_1)u_1\bar{u}_1 & (\mu-\varrho_1)u_1 U \\ (\bar{\mu}-\varrho_1)\bar{u}_1 U & -\varrho_1 U^2 \end{pmatrix}. \tag{6}$$

Oczywiście norma spektralna macierzy B jest równa normie spektralnej macierzy C. Aby wyznaczyć tę normę, musimy znaleźć największy moduł wartości własnej macierzy C. Wielomian charakterystyczny ma postać

$$\lambda^2+\lambda[2\varrho_1-(1+h)(\mu+\bar{\mu}-\mu\bar{\mu})]-(1-h^2)(1-\varrho_1)\mu\bar{\mu}=0. \tag{7}$$

Widzimy więc (zob. (3)), że maksymalny moduł wartości własnej jest równy

$$|\lambda|_{\max}=\tfrac{1}{2}\{|2\varrho_1-(1+h)(\mu+\bar{\mu}-\mu\bar{\mu})|+ \\ +\sqrt{[2\varrho_1-(1+h)(\mu+\bar{\mu}-\mu\bar{\mu})]^2+4(1-h^2)(1-\varrho_1)\mu\bar{\mu}}\}.$$

Skorzystamy teraz ze związków (3)

$$|\lambda|=\left|\frac{\frac{1}{2}\alpha+\varepsilon_1 h}{1+h}+\varepsilon_1-\varepsilon_2+\varepsilon_4-(1+h)\frac{\varepsilon_3+\bar{\varepsilon}_3+\Delta_5}{2}\right|+ \\ +\sqrt{\left(\frac{\frac{1}{2}\alpha+\varepsilon_1+h}{1+h}+\varepsilon_1-\varepsilon_2+\varepsilon_4-(1+h)\frac{\varepsilon_3+\bar{\varepsilon}_3+\Delta_5}{2}\right)^2+ \\ +(1-h)^2\left|\frac{\frac{1}{2}\alpha+\varepsilon_2-\varepsilon_1}{1+h}+\varepsilon_3+\Delta_6\right|^2}, \tag{8}$$

gdzie Δ_i są rzędu wielkości 2^{-2t}.

Oznaczamy oszacowanie prawej strony powyższej równości przez $\hat{\gamma}2^{-t}$. Zatem dla $\hat{\alpha}$ (mnożnik z oszacowania błędu α) i $\hat{\gamma}$ otrzymujemy dla $p=1$ (zob. (3.6), (3.11), (3.12), (3)), wartości przedstawione w poniższej tabelce:

	cfl	cfl̃	cfl_2
$\hat{\alpha}$	$n+1$	3	1
$\hat{\gamma}$	$\frac{1}{2}n+9+\sqrt{(\frac{1}{2}n+9)^2+(\frac{1}{2}n+6)^2} \leqslant$ $\leqslant(1+\sqrt{2})(\frac{1}{2}n+9)$	24	23,5

Z powyższych rozważań mamy oszacowanie postaci

$$\|P-\tilde{P}\|_2=\|B\|_2\leqslant 2^{-t}\hat{\gamma}. \tag{6}$$

Niech $\vec{g}$ oznacza numerycznie wyznaczony obraz wektora $\vec{b}$. Porównajmy go z wektorem $\vec{c}=P\vec{b}$:

$$\|\vec{g}-\vec{c}\|\leqslant\|\vec{g}-\tilde{P}\vec{b}\|+\|\tilde{P}\vec{b}-P\vec{b}\|\leqslant\|\vec{g}-\tilde{P}\vec{b}\|+2^{-t}\hat{\gamma}\|\vec{b}\|. \tag{10}$$

Oszacujemy teraz pierwszy składnik. Wektor $\vec{g}$ wyznaczamy w dwóch krokach (zob. (2.11))

$$\tilde{w}:=\frac{\vec{v}^H\vec{b}}{\tilde{R}},\qquad \vec{g}:=\vec{b}-\vec{v}\tilde{w}.$$

Wobec tego z (3.1), (3.5), (3.9), (3.10) otrzymujemy

$$\tilde{w}=\frac{\vec{v}^H(I-D_1)\vec{b}}{\tilde{R}},\qquad \vec{g}=(I-D_2)(\vec{b}-(I-D_3)\vec{v}\tilde{w}), \tag{11}$$

gdzie

$$D_1 = \operatorname{diag}(\delta_j), \quad D_2 = \operatorname{diag}(\beta_j), \quad D_3 = \operatorname{diag}(\alpha_j),$$

a błędy δ_j, β_j, α_j, mają oszacowania z mnożnikami $\hat{\delta}_j$, $\hat{\beta}_j$, $\hat{\alpha}_j$ odpowiednio równymi:

	cfl	cfĩ	cfl_2
$\hat{\delta}_j$	$n-j+3+\sqrt{2}$	$4+\sqrt{2}$	2

$\hat{\alpha}_j = 1+\sqrt{2}$, $\hat{\beta}_j = 1$ dla każdej z rozważanych arytmetyk. Zatem stąd oraz z (2), (3) mamy następujące oszacowanie:

$$(12) \qquad \|\vec{g}-\tilde{P}\vec{b}\| \underset{1}{\leqslant} \|D_2\|\,\|\vec{b}\| + \frac{\|\vec{v}\vec{v}^H\|}{\tilde{R}}(\|D_1\|+\|D_2\|+\|D_3\|)\|\vec{b}\| \underset{1}{\leqslant}$$

$$\underset{1}{\leqslant} (2\hat{\delta}_1+3\hat{\beta}_1+2\hat{\alpha}_1)\|\vec{b}\|2^{-t} \equiv \hat{\delta}\|\vec{b}\|2^{-t}.$$

Ostatecznie więc z (10), (12) wynika, że

$$\|\vec{g}-\vec{c}\| \underset{1}{\leqslant} (\hat{\gamma}+\hat{\delta})\|\vec{b}\|2^{-t} \equiv K_w\|\vec{c}\|2^{-t},$$

przy czym (przyjęliśmy $p = 1$)

	cfl	cfĩ	cfl_2
$\hat{\delta}$	$2n+11+3\sqrt{2}$	$13+4\sqrt{2}$	$9+2\sqrt{2}$
K_w	$3{,}2n+36$	42,6	35

Widzimy więc, że sądząc z oszacowań błędu, arytmetyka cfĩ jest prawie równie dobra, jak arytmetyka cfl_2.

W obu arytmetykach wektor $\vec{c}$ jest wyznaczany z dużą dokładnością względną, niezależną od wymiaru n.

5. Rzeczywiste zadanie H dla $\vec{e} = \vec{e}_1$. Jeśli $\vec{e} = \vec{e}_1$, to wzory (2.9), (2.10) przyjmują postać

$$R = A+|a_1|N, \quad u_1 = \begin{cases} a_1+N, & \text{gdy} \quad a_1 \geqslant 0, \\ a_1-N, & \text{gdy} \quad a_1 < 0, \end{cases} \quad u_j = a_j \ (j = 2, 3, \ldots, n),$$

gdzie

$$A = \|\vec{a}\|^2, \quad N = \|\vec{a}\|.$$

J. H. Wilkinson przedstawia w [6], str. 148 i następne, analizę błędu wytworzonego przy numerycznej realizacji w pełnej arytmetyce fl_2 rzeczywistego zadania H dla $\vec{e} = \vec{e}_1$. W niniejszej pracy podamy oszacowania potencjalnie nieco lepsze od oszacowania otrzymanego przez Wilkinsona (zob. również [1]). Nadal stosujemy oznaczenia wprowadzone w § 4. Niech

$$\tilde{A} = A(1-\alpha), \quad \tilde{N} = N(1-\nu), \quad \tilde{u}_1 = u_1(1-\mu), \quad \tilde{R} = R(1-\varrho).$$

Wprowadzimy obecnie zależności pomiędzy błędami α, ν, ϱ, μ. Łatwo sprawdzić, że

$$1-\nu = \sqrt{1-\alpha}\,(1-\varepsilon_1), \qquad |\varepsilon_1| \leqslant p2^{-t},$$

$$1-\mu = \left(1-\frac{\nu N}{a_1+\operatorname{sign}(a_1)N}\right)(1-\varepsilon_2), \qquad |\varepsilon_2| \leqslant 2^{-t},$$

$$1-\varrho = \left(1-\frac{\alpha A+(\nu+\varepsilon_3-\varepsilon_3\nu)|a_1|N}{A+|a_1|N}\right)(1-\varepsilon_4), \qquad |\varepsilon_3|, |\varepsilon_4| \leqslant 2^{-t}.$$

Stąd wynika, że

$$\begin{aligned} \nu &= \tfrac{1}{2}\alpha+\varepsilon_1+\Delta_1, \\ \mu &= \frac{\frac{1}{2}\alpha+\varepsilon_1}{1+h}+\varepsilon_2+\Delta_2, \\ \varrho &= \frac{\alpha+(\frac{1}{2}\alpha+\varepsilon_1+\varepsilon_3)h}{1+h}+\varepsilon_4+\Delta_3, \end{aligned} \tag{1}$$

gdzie

$$h = \frac{|a_1|}{||\vec{a}||},$$

a błędy Δ_j mają oszacowania na poziomie 2^{-2t}. Chcemy podkreślić, że rozważamy tu obliczenie R w „niepełnej" arytmetyce fl_2, w następujący sposób

$$\tilde{R} = \mathrm{fl}\left[\mathrm{fl}\left(\mathrm{fl}_2(||\vec{a}||^2)\right)+\sqrt{\mathrm{fl}\left(\mathrm{fl}_2(||\vec{a}||^2)\right)}\,|a_1|\right]. \tag{2}$$

Pokażemy teraz, jak zniekształcenie wektora $\vec{u}$ i parametru $\tilde{R}$ wpływa na przekształcenie P. Niech tak jak w § 4

$$||B|| = ||\tilde{P}-P|| \leqslant 2^{-t}\hat{\gamma}.$$

Podobnie jak w poprzednim punkcie

$$B = \frac{1}{R}(-\vec{u}\vec{u}^T\varrho_1+\vec{e}_1\vec{u}^Tu_1\mu+\vec{u}\vec{e}_1^Tu_1\mu-\vec{e}_1\vec{e}_1^Tu_1^2\mu^2),$$

gdzie

$$\varrho_1 = \varrho+\varrho^2+\varrho^3+\ \ldots\ = \varrho+\Delta_4, \qquad |\Delta_4| \underset{1}{\leqslant} 2^{-2t}.$$

Błąd α dla arytmetyki fl_2 i $\tilde{\mathrm{fl}}$ ma takie samo oszacowanie jak w § 4 dla arytmetyki cfl_2 i $\tilde{\mathrm{cfl}}$, a dla arytmetyki fl możemy pokazać, że $|\alpha| \underset{1}{\leqslant} n2^{-t}$. Odpowiednia tabelka dla $\hat{\gamma}$ wygląda więc następująco:

	fl	$\tilde{\mathrm{fl}}$	fl_2
$\hat{\gamma}$	$\frac{1}{2}n+4+\sqrt{(\frac{1}{2}n+4)^2+(\frac{1}{2}n+2)^2} \leqslant (1+\sqrt{2})(\frac{1}{2}n+4)$	12	9,7

Uwzględniając analogiczne zmiany w oszacowaniach elementów macierzy D_i otrzymujemy dla błędu otrzymanego obrazu $\vec{g}$ wektora $\vec{b}$ oszacowanie:

$$||\vec{g}-\vec{c}|| \underset{1}{\leqslant} (\hat{\gamma}+\hat{\delta})||\vec{b}||\,2^{-t} \equiv K_w||\vec{c}||\,2^{-t},$$

przy czym

	fl	$\tilde{\text{fl}}$	fl_2
$\hat{\delta}$	$2n+7$	13	9
K_w	$3{,}2n+17$	25	18,7

Otrzymane wartości wskaźnika kumulacji algorytmu, K_w, są dla arytmetyk $\tilde{\text{fl}}$ i fl_2 większe od wartości podanych w [1] (porównaj również [6], str. 154). Jak już wspominaliśmy, jest to związane m.in. ze sposobem obliczania wartości parametru R (zob. (2)). Ponadto, ze względu na ograniczenie zasięgu stosowania podwójnej precyzji w artmetyce fl_2 jedynie do kumulowania sum wielu składników, otrzymujemy większe oszacowania błędu wytworzonego przy obliczaniu $\tilde{P}\vec{b}$. Mianowicie, pokazaliśmy, że mnożnik $\hat{\delta} = 9$, gdy tymczasem Wilkinson, zakładając pełne wykorzystanie podwójnej precyzji w arytmetyce fl_2, otrzymuje $\hat{\delta} = 3{,}35$. Z tych samych względów istnieje możliwość obniżania oszacowań otrzymanych dla zespolonego zadania H, gdybyśmy rozszerzyli zakres stosowania podwójnej precyzji w arytmetyce cfl_2.

Tak więc pokazaliśmy, że zespolone i rzeczywiste zadanie (H) dla $\vec{e} = \vec{e}_1$ może być realizowane w sposób numerycznie poprawny w każdej z rozważanych arytmetyk.

6. Numeryczna realizacja ogólnego zadania (H). Niech P_a oznacza przekształcenie Householdera przeprowadzające wektor $\vec{a}$ na wektor o kierunku wektora $\vec{e}$. Wówczas P_a jest reprezentowane przez wektor $\vec{u}(a)$ i parametr R_a (zob. (2.5)–(2.7)).

Wprowadzimy pomocnicze oznaczenia: $\vec{v}$ — numerycznie wyznaczony wektor $\vec{u}(a)$, $\vec{f}$ — zaburzony wektor $\vec{a}$, taki, by $(\vec{e}^H\vec{a})_{\text{obl}} = \vec{e}^H\vec{f}$, $\tilde{P}_a$ — numerycznie wyznaczane przekształcenie Householdera reprezentowane przez $\vec{v}$ i $\tilde{R}_a$, $\vec{g}$ — numerycznie wyznaczony wektor $\tilde{P}_a\vec{b}$.

W przeprowadzonej przez nas analizie numerycznej ogólnego zadania (szczegóły pomijamy) otrzymaliśmy następujące oszacowania

$$\|\vec{f}-\vec{a}\| \underset{1}{\leqslant} K_d\|\vec{a}\|2^{-t}, \qquad \|\vec{v}-\vec{u}(f)\| \underset{1}{\leqslant} \hat{\mu}\|\vec{u}(f)\|2^{-t},$$

$$\|\tilde{P}_a-P_f\| \underset{1}{\leqslant} \hat{\gamma}2^{-t}, \qquad \|\vec{g}-P_f\vec{b}\| \underset{1}{\leqslant} K_w\|\vec{b}\|2^{-t} \underset{1}{=} K_w\|\vec{c}\|2^{-t},$$

gdzie $\vec{c} = P_a\vec{b}$, P_f oznacza przekształcenie Householdera skonstruowane dla wektora $\vec{f}$ i $\vec{e}$, a K_d, K_w, $\hat{\gamma}$, $\hat{\mu}$ są odpowiednio równe:

	fl	$\tilde{\text{fl}}$	fl_2	cfl	$\text{cf}\tilde{\text{l}}$	cfl_2
$\hat{\mu}$	$7n+2{,}8$	4,9	3,5	$2{,}1n+11{,}6$	15,8	9,6
$\hat{\gamma}$	$6{,}8n+13$	33,4	22,8	$14{,}5n+56$	91,6	50,2
K_d	0	0	0	$n+\sqrt{2}$	$2+\sqrt{2}$	$\langle 2^{-t}\rangle$
K_w	$8{,}8n+20$	47	32	$16{,}5n+71$	110	62

Widzimy więc, że w przypadku zespolonym otrzymany obraz $\vec{g}$ wektora $\vec{b}$ dla transformacji P_a może być interpretowany jako trochę zaburzony obraz wektora $\vec{b}$ dla transformacji $P_{\vec{f}}$, przy czym wektor $\vec{f}-\vec{a}$ jest małym zaburzeniem wektora $\vec{a}$. W tym sensie realizację ogólnego zadania (H) możemy uważać za numerycznie poprawną. W przypadku rzeczywistym $\vec{f}=\vec{a}$, w przypadku zespolonym zaś dla arytmetyki cfl_2 zaburzenia $\vec{f}-\vec{a}$ są na poziomie $2^{-2t}\|\vec{a}\|$. W trakcie analizy numerycznej zespolonego zadania (H) okazało się, że o dokładności obliczonego wektora $P_a\vec{b}$ decyduje przede wszystkim dokładność, z jaką oblicza się iloczyn skalarny $\vec{e}^H\vec{a}$. O wrażliwości wektora $\vec{u}(\vec{a})$ na zaburzenia wektora $\vec{a}$ decyduje bowiem wrażliwość ilorazu $\vec{e}^H\vec{a}/|\vec{e}^H\vec{a}|$ na małe zmiany składowych wektora $\vec{a}$ (a niekiedy może być ona duża).

Gdyby z jakichś powodów potrzebna była (również dla przypadku zespolonego) wysoka dokładność względna wektora $\vec{g}$, należałoby wtedy stosować jedynie arytmetykę cfl_2.

Bibliografia

[1] A. Kiełbasiński, *Algorytm sumowania z poprawkami i niektóre jego zastosowania*, Matematyka Stosowana 1 (1973), str. 23–41.

[2] — *Analiza numeryczna algorytmu ortogonalizacji Grama–Schmidta*, ibid. 2 (1974), str. 15–35.

[3] — *Podstawowe pojęcia analizy błędów w metodach numerycznych algebry liniowej*, ibid. 4 (1975), str. 5–27.

[4] J. Majchrowska i A. Smoktunowicz, *Analiza numeryczna algorytmu Ortegi–Householdera w dziedzinie zespolonej*, ten tom, str. 55–66.

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

[6] — *Алгебраическая проблема собственных значений*, Москва 1970.

[7] J. H. Wilkinson and C. Reinsch, *Handbook for automatic computation*, Springer-Verlag, 1971, str. 359–371 (lub Numer. Math. 12 (1968), str. 349–368).