Zadania

Jeśli w przekształcanych wyrażeniach jest $(1+\varepsilon_1)\cdot\ldots\cdot(1+\varepsilon_n)$, przy czym wszystkie epsilony mają wartości bezwzględne mniejsze niż ν (na przykład 10^{-7} albo 10^{-15}), to w analizie zastępujemy ten iloczyn sumą $(1+\varepsilon_1+\cdots+\varepsilon_n)$. Pominięte składniki są w istocie pomijalne. Na tej samej zasadzie $\sqrt{1+\varepsilon}\approx 1+\varepsilon/2$.

1. Znajdź wskaźnik uwarunkowania zadania obliczania $w=a^2-b^2$.

$$\operatorname{cond}_w a = \left|\frac{\partial w}{\partial a}\cdot\frac{a}{w}\right| = \left|\frac{2a^2}{a^2-b^2}\right| = \left|\frac{2}{1-(b/a)^2}\right|$$

Podobnie można obliczyć $\operatorname{cond}_w b$. Zwróćmy uwagę, że wskaźnik uwarunkowania jest większy lub równy 2, przy czym dla $|b/a|$ bliskiego 1 rośnie nieograniczenie.

2. Zbadaj błędy zaokrągleń wytworzone podczas obliczania wyrażeń $w=a^2-b^2$ i $w=(a+b)(a-b)$.

W pierwszym przypadku będzie obliczone

$$\begin{aligned}
\tilde{w} &= \big(a*a(1+\varepsilon_1)-b*b(1+\varepsilon_2)\big)(1+\varepsilon_3)\\
&= (a^2-b^2)\Big(\frac{a^2(1+\varepsilon_1)(1+\varepsilon_3)-b^2(1+\varepsilon_2)(1+\varepsilon_3)}{a^2-b^2}\Big)\\
&\approx (a^2-b^2)\Big(1+\frac{a^2(\varepsilon_1+\varepsilon_3)-b^2(\varepsilon_2+\varepsilon_3)}{a^2-b^2}\Big)\\
&= (a^2-b^2)\Big(1+\frac{(a/b)^2(\varepsilon_1+\varepsilon_3)-(\varepsilon_2+\varepsilon_3)}{(a/b)^2-1}\Big) = (a^2-b^2)(1+\gamma).
\end{aligned}$$

Dla $|a/b|\approx 1$ błąd względny γ obliczonego wyniku może być bardzo duży.

Z drugiej strony, mamy $\tilde{w}=\tilde{a}^2-\tilde{b}^2$, gdzie $\tilde{a}=a(1+\delta_a)$, $\tilde{b}=b(1+\delta_b$, $|\delta_a|,|\delta_b|\leqslant\frac{3}{2}\nu$ ($\nu=2^{-t}$, gdzie t jest liczbą bitów mantysy). Zatem algorytm jest numerycznie poprawny z małymi stałymi kumulacji — niedokładny wynik jest skutkiem złego uwarunkowania zadania.

W drugim przypadku otrzymamy
$\hat{w}=(a+b)(1+\varepsilon_1)(a-b)(1+\varepsilon_2)(1+\varepsilon_3)\approx(a^2-b^2)(1+\varepsilon_1+\varepsilon_2+\varepsilon_3)$ — wynik jest otrzymany z dużą dokładnością niezależnie od uwarunkowania.

3. Dokonaj analizy błędów w zadaniu obliczania sumy n liczb rzeczywistych metodą „po kolei”.

Obliczymy

$$\tilde{s} = \big(\dots(a_1 + a_2)(1 + \varepsilon_2) + \cdots + a_n\big)(1 + \varepsilon_n) = \tilde{a}_1 + \cdots + \tilde{a}_n,$$

gdzie

$$\tilde{a}_i = a_i(1 + \varepsilon_i) \cdot \ldots \cdot (1 + \varepsilon_n) \approx a_i(1 + \varepsilon_i + \cdots + \varepsilon_n) = a_i(1 + \gamma_i),$$

gdzie $|\gamma_i| \leqslant (n + i - 1)\nu$. Liczba $n + 1 - i$ jest stałą kumulacji, dla składnika a_i, wszystkie te stałe można oszacować z góry przez $n - 1$.

<u>Zadanie domowe:</u> Znajdź stałe kumulacji dla algorytmu sumowania parami, zrealizowanego przez podprogram

```
float Suma ( int n, float a[] )
{   int p;

  if ( n == 1 ) return a[0];
  else {
    p = n/2;
    return Suma ( p, a ) + Suma ( n-p, &a[k] );
  }
} /*Suma*/
```

Dla uproszczenia przyjmij, że $n = 2^k$ dla pewnego $k \in \mathbb{N}$.

4. Schemat Hornera obliczania wartości wielomianu, $w(x) = ax^n + \cdots + a_1x + a_0$, polega na użyciu wzoru

$$w(x) = \big(\dots(a_nx + a_{n-1})x + \ldots + a_1\big)x + a_0.$$

Przy użyciu tego algorytmu otrzymamy

$$\begin{aligned}\tilde{w}(x) = \Big(\big(\dots(a_nx(1 + \varepsilon_n) + a_{n-1})(1 + \delta_{n-1})x(1 + \varepsilon_{n-1}) + \ldots \\ + a_1\big)(1 + \delta_1)x(1 + \varepsilon_1) + a_0)(1 + \delta_0) \\ \tilde{a}_nx^n + \cdots + \tilde{a}_1x + \tilde{a}_0.\end{aligned}$$

Po rozwinięciu mamy

$$\begin{aligned}\tilde{a}_i &= a_i(1+\delta_i)(1+\varepsilon_i)\cdot\ldots\cdot(1+\delta_1)(1+\varepsilon_1)(1+\delta_0)\\ &\approx a_i(1+\delta_i+\varepsilon_i+\cdots+\delta_1+\varepsilon_1+\delta_0) = a_i(1+\gamma_i),\end{aligned}$$

gdzie $|\gamma_i| \leqslant (2i+1)\nu$. A więc obliczamy wartość w punkcie x trochę innego wielomianu.

Zadanie domowe: Znajdź stałe kumulacji dla schematu Hornera obliczania wartości wielomianu danego w bazie Newtona (zobacz s. 7.3 w skrypcie).

5. Wykaż, że algorytm obliczania tzw. niepełnego kwadratu sumy, tj. a^2+ab+b^2 na podstawie wzoru

$$w = \frac{1}{2}\big((a^2+b^2)+(a+b)^2\big)$$

jest numerycznie poprawny ze stałaymi kumulacji danych równymi 0 (a zatem otrzymujemy zaburzony na poziomie reprezentacji wynik dla oryginalnych danych a, b). Uwaga: mnożenie i dzielenie liczb zmiennopozycyjnych, jeśli nie ma nadmiaru ani niedomiaru, nie wprowadza błędów zaokrągleń.

Obliczymy

$$\begin{aligned}w &= (a^2+ab+b^2)\frac{1}{2}\times\\ &\quad\times\Big(\frac{\big((a^2(1+\varepsilon_1)+b^2(1+\varepsilon_2))(1+\varepsilon_3)+(a+b)^2(1+\varepsilon_4)^2(1+\varepsilon_5)\big)(1+\varepsilon_6)}{a^2+ab+b^2}\Big)\\ &= (a^2+ab+b^2)\Big(1+\frac{a^2(1+\beta_1)+ab(1+\beta_2)+b^2(1+\beta_2)}{a^2+ab+b^2}\Big).\end{aligned}$$

Polecam dokończenie tego zadania, tj. oszacowanie β_1, β_2, β_3 (w postaci stała razy ν) i oszacowanie funkcji $a^2/(a^2+ab+b^2)$, $ab/(a^2+ab+b^2)$ i $b^2/(a^2+ab+b^2)$ dla $a, b \in \mathbb{R}$.

Jak należy obliczać niepełny kwadrat różnicy, tj. a^2-ab+b^2?

6*. Udowodnij, że algorytmy obliczania wartości x i y niewiadomych w układzie 2 równań liniowych z dwiema niewiadomymi za pomocą wzorów Cramera (tj. z wyznacznikami) są numerycznie poprawne, tj. istnieją takie dane (współczynniki macierzy i współrzędne wektora prawej strony), zaburzone na poziomie reprezentacji w stosunku do danych oryginalnych, że obliczone $\tilde{x}$ i $\tilde{y}$ są dokładnymi rozwiązaniami zaburzonych układów — ale w obu przypadkach zaburzenia mogą być inne.