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 $w = a^2 - b^2$.

$$\mathrm{cond}_w\, a = \Big|\frac{\partial w}{\partial a}\cdot\frac{a}{w}\Big| = \Big|\frac{2a^2}{a^2-b^2}\Big| = \Big|\frac{2}{1-(b/a)^2}\Big|$$

Podobnie można obliczyć $\mathrm{cond}_w\, b$. 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

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

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 $n-1$.

Zadanie domowe: 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

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

Przy użyciu tego algorytmu otrzymamy

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

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.

Zadania

1. Wykaż, że jeśli macierz A jest symetryczna i dodatnio określona, to można eliminację Gaussa wykonać bez wyboru elementu głównego, a ponadto w macierzy $A^{(k)}$ otrzymanej w k-tym kroku eliminacji dolny prawy blok o wymiarach $(n-k)\times(n-k)$ jest macierzą symetryczną i dodatnio określoną.

Dowód I: Przyjmujemy założenie indukcyjne, że dolny prawy blok o wymiarach $(n-k+1)\times(n-k+1)$ macierzy $A^{(k-1)}$ jest symetryczny i dodatnio określony (dla $k=1$ jest $A^{(0)}=A$). Zatem $a_{kk}^{(k-1)}>0$, czyli dzielenie przez ten współczynnik jest wykonalne. Dla $i,j>k$ obliczymy

$$\begin{aligned} a_{ij}^{(k)} &= a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)} = a_{ji}^{(k-1)} - \frac{a_{ki}^{(k-1)}}{a_{kk}^{(k-1)}} a_{jk}^{(k-1)} = a_{ji}^{(k-1)} - \frac{a_{jk}^{(k-1)}}{a_{kk}^{(k-1)}} a_{ki}^{(k-1)} \\ &= a_{ji}^{(k)}. \end{aligned}$$

Zatem dolny prawy blok macierzy $A^{(k)}$ jest symetryczny. Jeśli nie jest dodatnio określony, to istnieje wektor $x\in\mathbb{R}^n$, o początkowe i współrzędnych równych 0, taki że $x^T A^{(k)} x \leqslant 0$. Ale k-ty krok eliminacji jest równoważny pomnożeniu macierzy $A^{(k-1)}$ przez macierz L_k^{-1}, która ma jedynki na diagonali, pewne liczby (jakie?) w k-tej kolumnie pod diagonalą i zera wszędzie indziej. Możemy sprawdzić, re wektor $y = L_k^{-T}x$ ma $k-1$ początkowych współrzędnych równych 0 i wtedy

$$0 \geqslant x^T A^{(k)} x = x^T L_k^{-1} A^{(k)} L_k^{-T} x = y^T A^{(k-1)} y,$$

co przeczy założeniu indukcyjnemu. □

Dowód II: Pomnożenie macierzy z lewej strony przez *dowolną* macierz L_k^{-1} i z prawej strony przez L_k^{-T} zachowują symetrię macierzy i określoność: jeśli $S^{(k-1)}$ jest symetryczna i dodatnio określona, to $S^{(k)} = L_k^{-1} S^{(k-1)} L_k^{-T}$ też jest taka. Jeśli zatem $S^{(0)}=A$ jest symetryczna i dodatnio określona, to

$$S^{(k)} = L_k^{-1}\dots L_1^{-1} S L_1^{-T}\dots L_k^{(k)}$$

też jest taka. W szczególności macierz A i wszystkie macierze symetryczne otrzymane w ten sposób mają wszystkie współczynniki na diagonali dodatnie (czyli niezerowe). Możliwe jest więc skonstruowanie dla macierzy $S^{(k-1)}$ macierzy L_k realizującej kolejny krok eliminacji Gaussa. Ale mnożenie przez czynnik L_k^{-1} z lewej strony zachowuje pierwsze k wierszy, mnożenie przez L_k^{-T} z prawej strony pozostawia niezmienione k początkowych kolumn. Stąd wynika, że macierze $S^{(k)}$ i $A^{(k)}$ mają identyczne współczynniki na diagonali i w prawym dolnym bloku $(n-k)\times(n-k)$ i macierz L_k realizująca krok eliminacji Gaussa jest taka sama dla macierzy $A^{(k-1)}$ i $S^{(k-1)}$. □

2. Powołując się na fakt dowiedziony w poprzednim zadaniu wykaż, że macierz symetryczna A jest dodatnio określona wtedy i tylko wtedy, gdy eliminacja Gaussa bez wyboru elementu głównego dla tej macierzy jest wykonalna i otrzymana w jej wyniku macierz trójkątna górna U ma wszystkie współczynniki na diagonali dodatnie.

Jeśli $A = LU$, gdzie macierze L i U są trójkątne (odpowiednio dolna i górna), to macierz $S = L^{-1}AL^{-T}$ jest symetryczna i diagonalna, a jej współczynniki na diagonali są takie jak w macierzy U. Macierz A jest dodatnio określona wtedy i tylko wtedy gdy S jest dodatnio określona wtedy i tylko wtedy gdy współczynniki na diagonali S są dodatnie. □

3. Jak można stwierdzić, że dana macierz symetryczna A jest ujemnie określona?

4. Wykaż, że jeśli macierz kwadratowa A jest diagonalnie dominująca (tj. zachodzą nierówności $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ dla $i = 1, \ldots, n$), to eliminacja Gaussa jest wykonalna bez wyboru elementu głównego.

Wystarczy udowodnić, że wszystkie kolejne macierze $A^{(k)}$ są diagonalne dominujące — stąd wszystkie współczynniki na diagonali są niezerowe. Zatem, z założenia indukcyjnego,

$$|a_{ii}^{(k-1)}| > \sum_{j\neq i} |a_{ij}^{(k-1)}| \quad \text{czyli} \quad \tau_i^{(k-1)} = \sum_{j\neq i} \frac{|a_{ij}^{(k-1)}|}{|a_{ii}^{(k-1)}|} < 1 \quad \text{dla } i \geqslant k.$$

Sumując powyższe nierówności stronami, mamy

$$\begin{aligned}\sum_{j=k+1}^{n} |a_{ij}^{(k)}| &\leqslant \sum_{j=k+1}^{n} |a_{ij}^{(k-1)}| + |a_{ik}^{(k-1)}| \sum_{j=k+1}^{n} \frac{|a_{kj}^{(k-1)}|}{|a_{kk}^{(k-1)}|} = \sum_{j=k+1}^{n} |a_{ij}^{(k-1)}| + |a_{ik}^{(k-1)}|\tau_k^{(k-1)} \\ &\leqslant \sum_{j=k}^{n} |a_{ij}^{(k-1)}|. \qquad (*)\end{aligned}$$

Z powyższej nierówności jest dodatkowy wniosek, że ciąg norm $\|A\|_\infty, \|A^{(1)}\|_\infty, \ldots, \|A^{(n-1)}\|_\infty$ jest nierosnący, a to oznacza małą stałą kumulacji (czyli dobrą dokładność końcowego wyniku — zobacz analizę błędu eliminacji Gaussa w skrypcie). Oznaczmy $p = |a_{ik}^{(k-1)}|/|a_{ii}^{(k-1)}|$ oraz $q = |a_{ki}^{(k-1)}|/|a_{kk}^{(k-1)}|$. Mamy $0 \leqslant p, q < 1$. Ze wzoru

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{kj}^{(k-1)}}{a_{kk}^{(k-1)}} a_{ik}^{(k-1)}$$

wynika nierówność

$$|a_{ii}^{(k)}| \geqslant |a_{ii}^{(k-1)}| - \frac{|a_{ik}^{(k-1)}||a_{ki}^{(k-1)}|}{|a_{kk}^{(k-1)}|} > |a_{ii}^{(k-1)}|(1-pq) > 0.$$

Stąd i na podstawie (*)

$$\begin{aligned}\frac{\sum_{j>k,j\neq i}|a_{ij}^{(k)}|}{|a_{ii}^{(k)}|} &\leqslant \frac{\sum_{j>k,j\neq i}|a_{ij}^{(k-1)}| + |a_{ik}^{(k-1)}|(\tau_k^{(k-1)} - q)}{|a_{ii}^{(k-1)}|(1-pq)} \\ &= \frac{\tau_i^{(k-1)} - p + p(\tau_k^{(k-1)} - q)}{1-pq} \\ &= \tau_i^{(k-1)} - p\frac{q(1-\tau_i^{(k-1)}) + (1-\tau_k^{(k-1)})}{1-pq} \leqslant \tau_i^{(k-1)} < 1.\end{aligned}$$

Zatem, macierz $A^{(k)}$, która ma k początkowych wierszy takich jak $A^{(k-1)}$, też jest diagonalnie dominująca. □

5. Macierz $n \times n$ o następującej budowie:

$$A = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 & a \\ -1 & 1 & 0 & \dots & \dots & 0 & a \\ -1 & -1 & 1 & 0 & \dots & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & a \\ -1 & -1 & \dots & \dots & -1 & 1 & a \\ -1 & -1 & \dots & \dots & -1 & -1 & a \end{bmatrix}$$

zostanie w wyniku eliminacji Gaussa bez wyboru elementu głównego, lub z wyborem częściowym w kolumnie, przekształcona tak, że powstanie macierz

$$U = \begin{bmatrix} 1 & 0 & \dots & \dots & \dots & 0 & a \\ 0 & 1 & 0 & \dots & \dots & 0 & 2a \\ 0 & 0 & 1 & 0 & \dots & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 & 2^{n-2}a \\ 0 & 0 & \dots & \dots & 0 & 1 & 2^{n-1}a \\ 0 & 0 & \dots & \dots & 0 & 0 & 2^n a \end{bmatrix}$$

Jeśli $|a|$ jest duże, to $\|A\|_\infty = |a| + n - 1$, zaś $\|U\|_\infty = 2^n|a|$. To w tym przypadku stała kumulacji algorytmu eliminacji Gaussa jest rzędu 2^n, zatem numeryczna poprawność staje się iluzoryczna już wtedy, gdy n jest rzędu kilkanaście.

Oczywiście, to jest przykład akademicki, w układach pochodzących z praktycznych zadań jest mała szansa, aby spotkać taką macierz.

6. Niech

$$A = \begin{bmatrix} 4 & -2 & 0 & -2 \\ -2 & 2 & 1 & 1 \\ 0 & 1 & 5 & 2 \\ -2 & 1 & 2 & 3 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 8 \\ -3 \\ 13 \\ 2 \end{bmatrix}.$$

Znajdź macierz trójkątną dolną L, taką że $A = LL^T$. Korzystając z macierzy L rozwiąż układ równań $A\boldsymbol{x} = \mathbf{b}$.

Zwróć uwagę, że jeśli współczynniki $a_{i1}, \ldots, a_{ik}$ (dla $k < i$) macierzy $A = LL^T$ są równe 0, to również $l_{i1} = \cdots = l_{ik} = 0$.

7. Niech

$$A = \begin{bmatrix} -1 & 3 & 162 & 21 \\ -1 & -8 & -261 & -188 \\ 1 & 5 & -81 & 77 \\ -1 & -8 & 18 & 244 \end{bmatrix}, \qquad \begin{bmatrix} 185 \\ -458 \\ 2 \\ 253 \end{bmatrix}.$$

Znajdź wektory $\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3 \in \mathbb{R}^4$ wyznaczające odbicia symetryczne (reprezentowane przez macierze $H_i = I - \boldsymbol{v}_i \gamma_i \boldsymbol{v}_i^T$, $\gamma_i = \frac{2}{\boldsymbol{v}_i^T \boldsymbol{v}_i}$), takie że macierz $R = H_3 H_2 H_1 A$ jest trójkątna górna. Korzystając z tgo rozkładu, rozwiąż układ równań liniowych $A\boldsymbol{x} = \mathbf{b}$.