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 $A\boldsymbol{x} = \mathbf{b}$.

8. (KS) Przykład: rozkład $PA = LU$ z wyborem elementu głównego dla macierzy

$$A = \begin{bmatrix} 2 & -2 & -2 \\ -2 & 0 & 1 \\ 4 & 1 & -2 \end{bmatrix} =: \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

Szukamy:

$$L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix},$$

oraz macierzy P permutacji σ, którą wypiszemy na końcu. Będziemy jednakże wykonywać potrzebne permutacje. Kolejne elementy macierzy L oraz U będziemy znajdowali z zależności wynikającej wprost z iloczynu macierzy $A = LU$:

$$a_{\sigma(i),j} = \sum_k l_{ik} u_{kj},$$

przechodząc elementy macierzy A w następującej kolejności:

$$\begin{bmatrix} [1] & (2) & (3) \\ (4) & [6] & (7) \\ (5) & (8) & [9] \end{bmatrix}.$$

Taki sposób zapewnia, że wykorzystamy istniejące zera w macierzach L i U powyższy wzór będzie zawierać za każdym razem tylko jeden nieznany element. Przy elementach diagonalnych $[1], [6], [9]$ wykonamy potrzebną permutację.

Zaczynamy od (1). Otrzymujemy

$$a_{\sigma(1),1} = \sum_k l_{1k}u_{k1} = 1u_{11} + 0 + 0 = u_{11}.$$

Zatem mamy zależność na u_{11}. Teraz dokonujemy permutacji. Mamy takie możliwości:

- Nie wykonujemy zamiany: $u_{11} = 2$.
- Zamieniamy $1 \leftrightarrow 2$, wtedy $a_{\sigma(1),1}$ będzie -2, czyli też $u_{11} = -2$.
- Zamieniamy $1 \leftrightarrow 3$, wtedy $a_{\sigma(1),1}$ będzie 4, czyli też $u_{11} = 4$.

Wybieramy taką zamianę, żeby u_{11} miało jak największy moduł, czyli $1 \leftrightarrow 3$. Po zamianie mamy taką sytuację:

$$T_1A = \begin{bmatrix} 4 & 1 & -2 \\ -2 & 0 & 1 \\ 2 & -2 & -2 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix},$$

Liczymy dalej zgodnie z ustaloną wcześniej kolejnością:

$$1 = a_{\sigma_1(1),2} = l_{11}u_{12} + 0 = u_{12} \Rightarrow u_{12} = 1$$

$$-2 = a_{\sigma_1(1),3} = u_{13},$$

$$-2 = a_{\sigma_1(2),1} = a_{21} = l_{21}u_{11} + 0 = 4l_{21} \Rightarrow l_{21} = -0.5,$$

$$2 = a_{\sigma_1(3),1} = l_{31}u_{11} + 0 = 4l_{31} \Rightarrow l_{21} = 0.5,$$

Wykonaliśmy obliczenia dla pierwszego rzędu i kolumny. Mamy

$$T_1A = \begin{bmatrix} 4 & 1 & -2 \\ -2 & 0 & 1 \\ 2 & -2 & -2 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ -0.5 & 1 & 0 \\ 0.5 & l_{32} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & 1 & -2 \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix},$$

Teraz element diagonalny:

$$a_{\sigma(2)2} = l_{21}u_{12} + l_{22}u_{22} + l_{23}u_{32} = 1l_{21} + u_{22}$$

Rozważamy możliwe permutacje:

- Nie wykonujemy zamiany: $u_{22} = a_{22} - l_{21} = 0.5$.
- Zamieniamy $2 \leftrightarrow 3$, wtedy $a_{\sigma(2),2}$ będzie -2, ale nie tylko. Wówczas musimy też zamienić obliczone już wartości w macierzy L, tak jakby permutacja ta była wykonana od początku obliczeń. Zatem byłoby $l_{21} = 0.5$, $l_{31} = -0.5$. Wówczas $u_{22} = -2 - 0.5 = -2.5$
- Wierszy już wykonanych nie zmieniamy, także $2 \leftrightarrow 1$ nie rozważamy.

Wybieramy taką zamianę, żeby u_{22} miało jak największy moduł, czyli $2 \leftrightarrow 3$. Otrzymujemy

$$T_2T_1A = \begin{bmatrix} 4 & 1 & -2 \\ 2 & -2 & -2 \\ -2 & 0 & 1 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.5 & l_{32} & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & 1 & -2 \\ 0 & -2.5 & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix},$$

Liczymy dalej:

$$-2 = a_{\sigma(2)3} = -2 \cdot 0.5 + u_{23} \Rightarrow u_{23} = -1.$$

$$0 = a_{\sigma(3),2} = -0.5 \cdot 1 - 2.5l_{32} \Rightarrow l_{32} = -0.2.$$

Otrzymujemy

$$T_2T_1A = \begin{bmatrix} 4 & 1 & -2 \\ 2 & -2 & -2 \\ -2 & 0 & 1 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.5 & -0.2 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & 1 & -2 \\ 0 & -2.5 & -1 \\ 0 & 0 & u_{33} \end{bmatrix}.$$

Następnie mamy element diagonalny. Nie ma już czego permutować, ponieważ pozostał jeden wiersz. Także liczymy

$$1 = a_{\sigma(3),3} = (-0.5) \cdot (-2) + (-0.2) \cdot (-1) + 1u_{33}, \Rightarrow u_{33} = -0.2.$$

Finalnie mamy

$$PA = T_2T_1A = \begin{bmatrix} 4 & 1 & -2 \\ 2 & -2 & -2 \\ -2 & 0 & 1 \end{bmatrix}, \qquad L = \begin{bmatrix} 1 & 0 & 0 \\ 0.5 & 1 & 0 \\ -0.5 & -0.2 & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} 4 & 1 & -2 \\ 0 & -2.5 & -1 \\ 0 & 0 & -0.2 \end{bmatrix}.$$

Tylko jest to rozkład macierzy A o przestawionych wierszach. Kolejność wierszy z oryginalnej macierzy to 3,1,2, dedukujemy stąd postać P:

$$P = T_2 T_1 = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

Zadania

1. Niech

$$A = \begin{bmatrix} 1 & 2 & 3 \\ -1 & 0 & 3 \\ 1 & 2 & 3 \\ 1 & 4 & 1 \\ 0 & 1 & 1 \end{bmatrix}.$$

Znajdź rozkład ortogonalno-trójkątny tej macierzy a) metodą odbić Householdera, b) metodą ortonormalizacji Grama–Schmidta, c) metodą modyfikowaną ortonormalizacji Grama–Schmidta.

Symbolami $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ oznaczymy kolumny macierzy A.

a) Konstruujemy kolejno 3 odbicia. Pierwsze, względem hiperpłaszczyzny, której wektorem normalnym jest wektor $\boldsymbol{v}_1 = \mathbf{a}_1 + \|\mathbf{a}_1\|_2 \boldsymbol{e}_1$ — znak „+” jest wybrany dlatego, bo pierwsza współrzędna wektora $\mathbf{a}_1$ jest dodatnia. Zatem, $\boldsymbol{v}_1 = [3, -1, 1, 1, 0]^T$. Obliczamy

$$\gamma_1 = \frac{2}{\boldsymbol{v}_1^T \boldsymbol{v}_1} = \frac{2}{12} = \frac{1}{6}.$$

Obrazem wektora $\mathbf{a}_1$ w tym odbiciu jest wektor $-2\boldsymbol{e}_1$ (to wiemy bez stosowania wzoru na odbicie). Ponadto $H_1\mathbf{a}_i = \mathbf{a}_i - \boldsymbol{v}_1\gamma_1\boldsymbol{v}_1^T\mathbf{a}_i$, $\boldsymbol{v}_1^T\mathbf{a}_2 = 12$, $\boldsymbol{v}_1^T\mathbf{a}_3 = 10$, stąd

$$H_1\mathbf{a}_2 = \begin{bmatrix} 2 \\ 0 \\ 2 \\ 4 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ -1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \frac{12}{6} = \begin{bmatrix} -4 \\ 2 \\ 0 \\ 2 \\ 1 \end{bmatrix},$$

$$H_1\mathbf{a}_3 = \begin{bmatrix} 3 \\ 3 \\ 3 \\ 1 \\ 1 \end{bmatrix} - \begin{bmatrix} 3 \\ -1 \\ 1 \\ 1 \\ 0 \end{bmatrix} \frac{10}{6} = \begin{bmatrix} -2 \\ \frac{14}{3} \\ \frac{4}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix}.$$

Po pierwszym odbiciu mamy zatem macierz

$$A^{(1)} = H_1 A = \begin{bmatrix} -2 & -4 & -2 \\ 0 & 2 & \frac{14}{3} \\ 0 & 0 & \frac{4}{3} \\ 0 & 2 & -\frac{2}{3} \\ 0 & 1 & 1 \end{bmatrix}$$

Konstruujemy drugie odbicie: część drugiej kolumny macierzy $A^{(1)}$ otrzymana po odrzuceniu pierwszego współczynnika to wektor $\overline{\mathbf{a}}_2^{(1)} = [2, 0, 2, 1]^T$, jest $\|\overline{\mathbf{a}}_2^{(1)}\|_2 = 3$. Pierwszy współczynnik jest dodatni, zatem weźmiemy wektor normalny hiperpłaszczyzny odbicia $\overline{\mathbf{v}}_2 = [5, 0, 2, 1]^T$ (to jest wektor normalny hiperpłaszczyzny odbicia w $\mathbb{R}^4$, w „całej" przestrzeni $\mathbb{R}^5$ odbijamy za pomocą wektora $\mathbf{v}_2 = [0, 5, 0, 2, 1]^T$). Obrazem wektora $\overline{\mathbf{a}}_2^{(1)}$ jest oczywiście wektor $[-3, 0, 0, 0]^T$; znowu nie musimy go obliczać z ogólnego wzoru, zatem wystarczy obliczyć

$$\gamma_2 = \frac{2}{\overline{\mathbf{v}}_2^T \overline{\mathbf{v}}_2} = \frac{1}{15},$$

potem $\overline{\mathbf{v}}_2^T \overline{\mathbf{a}}_3^{(1)} = 23$ i dalej

$$\overline{H}_2 \overline{\mathbf{a}}_3^{(1)} = \begin{bmatrix} \frac{14}{3} \\ \frac{4}{3} \\ -\frac{2}{3} \\ 1 \end{bmatrix} - \begin{bmatrix} 5 \\ 0 \\ 2 \\ 1 \end{bmatrix} \frac{23}{15} = \begin{bmatrix} -3 \\ \frac{4}{3} \\ -\frac{56}{15} \\ -\frac{8}{15} \end{bmatrix}.$$

Po dwóch odbiciach powstała macierz

$$A^{(2)} = H_2 H_1 A = \begin{bmatrix} -2 & -4 & -2 \\ 0 & -3 & -3 \\ 0 & 0 & \frac{4}{3} \\ 0 & 0 & -\frac{56}{15} \\ 0 & 0 & -\frac{8}{15} \end{bmatrix}.$$

Ostatnie odbicie konstruujemy dla wektora złożonego z ostatnich trzech współczynników trzeciej kolumny macierzy $A^{(2)}$, czyli wektora $\overline{\mathbf{a}}_3^{(3)} = [\frac{4}{3}, -\frac{56}{15}, -\frac{8}{15}]^T$, który ma długość 4. Znów, pierwsza współrzędna jest dodatnia, zatem przyjmiemy $\overline{\mathbf{v}}_3 = [\frac{16}{3}, -\frac{56}{15}, -\frac{8}{15}]^T$. Mamy

$$\gamma_3 = \frac{2}{\overline{\mathbf{v}}_3^T \overline{\mathbf{v}}_3} = \frac{3}{64}.$$

Do dokończenia znajdowania rozkładu liczba γ_3 nie jest potrzebna, bo wiemy, co wyjdzie: macierz

$$R = A^{(3)} = \begin{bmatrix} -2 & -4 & -2 \\ 0 & -3 & -3 \\ 0 & 0 & -4 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

ale γ_3 przyda się później w rozwiązywaniu LZNK.

Dla ciekawości, macierz $Q = H_1H_2H_3$ (odwrotność złożenia wykonanych odbić, czyli złożenie tych odbić w odwrotnej kolejności) jest taka:

$$Q = \begin{bmatrix} -\frac{1}{2} & 0 & -\frac{1}{2} & -\frac{7}{10} & -\frac{1}{10} \\ \frac{1}{2} & -\frac{2}{3} & -\frac{1}{2} & \frac{1}{30} & -\frac{7}{30} \\ -\frac{1}{2} & 0 & -\frac{1}{2} & \frac{7}{10} & \frac{1}{10} \\ -\frac{1}{2} & -\frac{2}{3} & \frac{1}{2} & \frac{1}{30} & -\frac{7}{30} \\ 0 & -\frac{1}{3} & 0 & -\frac{4}{30} & \frac{28}{30} \end{bmatrix}.$$

Ale jak nie mamy wyraźnego powodu, to nie wyznaczamy tej macierzy w postaci jawnej — wystarczy zapamiętać wektory $\boldsymbol{v}_1$, $\overline{\boldsymbol{v}}_2$, $\overline{\boldsymbol{v}}_3$ i liczby γ_1, γ_2, γ_3.

b) Wykonujemy (standardową) ortonormalizację Grama–Schmidta. Długość pierwszej kolumny, $\|\boldsymbol{a}_1\|_2 = 2$ jest współczynnikiem r_{11} macierzy R_1. Pierwszą kolumnę macierzy Q_1 otrzymamy, dzieląc wektor $\boldsymbol{a}_1$ przez jego długość. Stąd mamy

$$\boldsymbol{q}_1 = [\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, 0]^T.$$

Dalej, obliczamy współczynnik $r_{12} = \boldsymbol{q}_1^T\boldsymbol{a}_2 = 4$ i drugą kolumnę macierzy Q_1:

$$\hat{\boldsymbol{q}}_2 = \boldsymbol{a}_2 - \boldsymbol{q}_1 r_{12} = [0, 2, 0, 2, 1]^T, \qquad \boldsymbol{q}_2 = \tfrac{1}{\|\hat{\boldsymbol{q}}_2\|_2}\hat{\boldsymbol{q}}_2 = [0, \tfrac{2}{3}, 0, \tfrac{2}{3}, \tfrac{1}{3}]^T,$$

przy czym po drodze obliczyliśmy też $r_{22} = \|\hat{\boldsymbol{q}}_2\|_2 = 3$.

Wreszcie obliczamy trzecią kolumnę:

$$\hat{\boldsymbol{q}}_3 = \boldsymbol{a}_3 - \boldsymbol{q}_1 r_{13} - \boldsymbol{q}_2 r_{23},$$

gdzie

$$r_{13} = \boldsymbol{q}_1^T\boldsymbol{a}_3 = 2, \quad r_{23} = \boldsymbol{q}_2^T\boldsymbol{a}_3 = 3,$$

jest też $r_{33} = \|\hat{\boldsymbol{q}}_3\|_2 = \sqrt{\hat{\boldsymbol{q}}_3^T\hat{\boldsymbol{q}}_3}$ i w końcu

$$\boldsymbol{q}_3 = \tfrac{1}{r_{33}}\hat{\boldsymbol{q}}_3 = [\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, -\tfrac{1}{2}, 0]^T.$$

Mamy zatem czynniki rozkładu macierzy A:

$$Q_1 = \begin{bmatrix} \frac{1}{2} & 0 & \frac{1}{2} \\ -\frac{1}{2} & \frac{2}{3} & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{2}{3} & -\frac{1}{2} \\ 0 & \frac{1}{3} & 0 \end{bmatrix}, \quad R_1 = \begin{bmatrix} 2 & 4 & 2 \\ 0 & 3 & 3 \\ 0 & 0 & 4 \end{bmatrix}.$$

c) W algorytmie modyfikowanym Grama–Schmidta pierwszą kolumnę macierzy Q_1 obliczamy tak samo, jak w metodzie standardowej, czyli dzielimy ją przez długość, r_{11}, otrzymując $\mathbf{q}_1 = [\frac{1}{2}, -\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 0]^T$. Następnie rzutujemy pozostałe kolumny na hiperpłaszczyznę prostopadłą do $\mathbf{q}_1$, otrzymując

$$
\begin{aligned}
&r_{12} = \mathbf{q}_1^T \mathbf{a}_2 = 4,\\
&\mathbf{a}_2^{(1)} = \mathbf{a}_2 - \mathbf{q}_1 r_{12} = [2,0,2,4,1]^T - 4[\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, 0]^T = [0,2,0,2,1]^T,\\
&r_{13} = \mathbf{q}_1^T \mathbf{a}_3 = 2,\\
&\mathbf{a}_3^{(1)} = \mathbf{a}_3 - \mathbf{q}_1 r_{13} = [3,3,3,1,1]^T - 2[\tfrac{1}{2}, -\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, 0]^T = [2,4,2,0,1]^T,
\end{aligned}
$$

Teraz normalizujemy (dzielimy przez długość, r_{22}) wektor $\mathbf{a}_2^{(1)}$ i otrzymujemy wektor $\mathbf{q}_2$:

$$
\begin{aligned}
&r_{22} = \|\mathbf{a}_2^{(1)}\|_2 = 3,\\
&\mathbf{q}_2 = \tfrac{1}{r_{22}} \mathbf{a}_2^{(1)} = [0, \tfrac{2}{3}, 0, \tfrac{2}{3}, \tfrac{1}{3}]^T.
\end{aligned}
$$

Pozostałe kolumny (tylko jedną, tj. $\mathbf{a}_3^{(1)}$) rzutujemy na hiperpłaszczyznę prostopadłą do $\mathbf{q}_2$:

$$
\begin{aligned}
&r_{23} = \mathbf{q}_2^T \mathbf{a}_2^{(1)} = 3,\\
&\mathbf{a}_2^{(2)} = \mathbf{a}_2^{(1)} - \mathbf{q}_2 r_{23} = [2,4,2,0,1]^T - 3[0, \tfrac{2}{3}, 0, \tfrac{2}{3}, \tfrac{1}{3}]^T = [2,2,2,-2,0]^T.
\end{aligned}
$$

Wektor $\mathbf{q}_3$ otrzymamy, dzieląc $\mathbf{a}_2^{(2)}$ przez $r_{33} = \|\mathbf{a}_2^{(2)}\|_2 = 4$.

Otrzymaliśmy te same czynniki rozkładu, co w wyniku użycia metody standardowej, ale po drodze do wektora $\mathbf{q}_3$ mieliśmy inne wyniki pośrednie. W implementacji korzystającej z arytmetyki zmiennopozycyjnej algorytm modyfikowany jest dokładniejszy (koszt jest dokładnie taki sam), bo wektory, które trzeba rzutować na hiperpłaszczyzny normalne wcześniej znalezionych elementów bazy ortonormalnej są położone bliżej tych hiperpłaszczyzn, przez co następuje mniejsze znoszenie się składników (w tym przykładzie odległość wektora $\mathbf{a}_3^{(1)}$ od $\mathbf{q}_3$ jest mniejsza niż wektora $\mathbf{a}_3$ od $\mathbf{q}_3$). Błędy zaokrągleń sprawiają, że znalezione kolumny macierzy Q_1 nie są wzajemnie prostopadłe (czyli $Q_1^T Q_1$ nie jest dokładnie macierzą jednostkową). Algorytm modyfikowany dostarcza wektory, między którymi kąty są bliższe kąta prostego.

2. W przykładzie pokazanym wyżej można zauważyć, że pierwsze trzy kolumny macierzy $Q = H_1 H_2 H_3$ opisującej złożenie odbić Householdera różnią się tylko znakami od kolumn macierzy Q_1 otrzymanych przy użyciu ortogonalizacji (z kolei macierze R_1 mają wiersze identyczne z dokładnością do znaku). Zatem, proszę

udowodnić, że tak jest zawsze: w dowolnym (uzyskanym dowolną metodą) rozkładzie macierzy A o wymiarach $m \times n$, której kolumny są liniowo niezależne, na czynniki ortogonalny Q i trójkątny R, kierunki (i długości, zawsze 1) pierwszych n kolumn macierzy Q (czyli blok Q_1 macierzy Q) są określone jednoznacznie.

3. Rozwiąż liniowe zadanie najmniejszych kwadratów z macierzą A z zadania 1 i z wektorem prawej strony $\mathbf{b} = [-3, 33, 3, -47, 28]^T$ — a) przez rozwiązanie układu równań normalnych (za pomocą metody Choleskiego), b) metodą odbić Householdera i c) korzystając z macierzy Q_1, R_1 znalezionych przy użyciu ortonormalizacji.

a) Obliczamy

$$A^TA = \begin{bmatrix} 4 & 8 & 4 \\ 8 & 25 & 17 \\ 4 & 17 & 29 \end{bmatrix}, \qquad A^T\mathbf{b} = \begin{bmatrix} -14 \\ -160 \\ -118 \end{bmatrix}.$$

Macierz $M = A^TA$ rozkładamy metodą Choleskiego na czynniki L, L^T. Otrzymujemy

$$L = \begin{bmatrix} 2 & 0 & 0 \\ 4 & 3 & 0 \\ 2 & 3 & 4 \end{bmatrix}.$$

Zauważmy, że dostaliśmy macierz, która jest transpozycją macierzy R_1 otrzymanej w wyniku ortogonalizacji Grama–Schmidta. Rozwiązujemy układy z macierzami trójkątnymi $L\mathbf{y} = A^T\mathbf{b}$ i $L^T\mathbf{x} = \mathbf{y}$. Wychodzi

$$\mathbf{y} = \begin{bmatrix} -40 \\ 0 \\ 40 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} -10 \\ -10 \\ 10 \end{bmatrix}.$$

b) Poddajemy kolejnym odbiciom wektor prawej strony:

$$\mathbf{b}^{(1)} = H_1\mathbf{b} = \mathbf{b} - \boldsymbol{\nu}_1\gamma_1\boldsymbol{\nu}_1^T\mathbf{b} = [40, \tfrac{56}{3}, \tfrac{52}{3}, -\tfrac{98}{3}, 28]^T,$$
$$\mathbf{b}^{(2)} = H_2\mathbf{b}^{(1)} = \mathbf{b}^{(1)} - \boldsymbol{\nu}_2\gamma_2\boldsymbol{\nu}_2^T\mathbf{b}^{(1)} = [40, 0, \tfrac{52}{3}, -\tfrac{602}{15}, \tfrac{364}{15}]^T,$$
$$\mathbf{b}^{(3)} = H_3\mathbf{b}^{(2)} = \mathbf{b}^{(2)} - \boldsymbol{\nu}_3\gamma_3\boldsymbol{\nu}_3^T\mathbf{b}^{(2)} = [40, 0, -40, 0, 3]^T.$$

W ostatnim z tych wektorów wyróżniamy dwa bloki: wektor $\mathbf{y} = [40, 0, -40]^T$,

który przyjmujemy za prawą stronę układu równań

$$\begin{bmatrix} -2 & -4 & -2 \\ 0 & -3 & -3 \\ 0 & 0 & -4 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 40 \\ 0 \\ -40 \end{bmatrix}$$

oraz wektor $[0, 3]^T$, którego długość jest długością residuum, tj. wektora $\mathbf{b} - A\mathbf{x}$ dla rozwiązania LZNK (tego samego, co uzyskanego przez rozwiązanie układu równań normalnych, o ile nie bierzemy pod uwagę błędów zaokrągleń).

c) Mając macierze Q_1 i R_1, obliczamy wektor

$$\mathbf{y} = Q_1^T \mathbf{b} = [-40, 0, 40]^T$$

i rozwiązujemy układ równań $R_1\mathbf{x} = \mathbf{y}$. Rozwiązanie jest takie samo, jak poprzednio (z dokładnością do błędów zaokrągleń, których tu nie było tylko dlatego, że rachunki zostały przeprowadzone na ułamkach reprezentowanych przez całkowite liczniki i mianowniki, a nie w arytmetyce zmiennopozycyjnej).