ZADANIA Z MATEMATYKI OBLICZENIOWEJ

(1) Zaproponuj algorytm obliczania wartości wyrażenia

$$f(x) = (1+x)^3 - 1 \quad \text{dla} \quad x > 0$$

z błędem względnym na poziomie dokładności arytmetyki.

(2) Niech $f(x) = \sqrt{2x+30} - \sqrt{2x+10}$ dla $x \geqslant -5$. Rozpatrzmy dwa algorytmy, A1 i A2, obliczania $f(x)$.
A1: zgodnie z powyższym wzorem,
A2: korzystając z równoważnego wzoru

$$f(x) = \frac{20}{\sqrt{2x+30} + \sqrt{2x+10}}.$$

Który z tych algorytmów należy zastosować do obliczenia $f(x)$ w arytmetyce fl_ν dla x rzędu 10^{16}? Podaj krótkie uzasadnienie.

(3) Jak w arytmetyce fl_ν policzyć wartość funkcji $1 - \cos(4x)$ dla x na poziomie dokładności arytmetyki, aby otrzymać mały błąd względny? (Zakładamy, że funkcje trygonometryczne potrafimy liczyć z błędem względnym na poziomie błędu reprezentacji.)

• Wobec tożsamości

$$1 - \cos(4x) = (\cos^2(2x) + \sin^2(2x)) - (\cos^2(2x) - \sin^2(2x)) = 2\sin^2(2x),$$

podane wyrażenie możemy policzyć z błędem na poziomie reprezentacji stosując wzór po prawej stronie tego równania.

(4) Czy występowanie zjawiska redukcji cyfr znaczących oznacza, że zadanie odejmowania liczb 'podobnych' nie jest dobrze uwarunkowane?

• Tak, bo skoro odejmowanie jest numerycznie poprawne (ze stałymi kumulacji 2) to dobre uwarunkowanie oznaczałoby też wynik w fl_ν z błędem względnym na poziomie ν. Mielibyśmy sprzeczność.
□

(5) Niech $a_i \geqslant 0$ dla $1 \leqslant i \leqslant n$. Czy z punktu widzenia błędów w fl_ν lepiej jest policzyć sumę tych liczb w kolejności od najmniejszej do największej czy odwrotnie?

• Mamy

$$\left| fl_\nu\Big(\sum_{i=1}^n a_i\Big) - \sum_{i=1}^n a_i \right| \leqslant \sum_{i=1}^n |\varepsilon_i| a_i \lessapprox \nu\Big(n a_1 + \sum_{i=2}^n (n-i+2) a_i\Big),$$

przy czym to oszacowanie górne jest ostre. Prawa strona jest oczywiście najmniejsza gdy kolejne a_i dodajemy w kolejności od najmniejszej do największej.
□

(6) Niech A będzie nieosobliwą macierza kwadratową. Pokaż, że jeśli $\|E\| \leqslant K_1\nu\|A\|$ i $\|\vec{e}\| \leqslant K_2\nu\|\vec{x}\|$ to

$$\|(A+E)(\vec{x}+\vec{e})\| \lessapprox (K_1+K_2)(\|A\|\|A^{-1}\|)\,\nu\,\|A\vec{x}\|.$$

(Zakładamy, że norma macierzowa jest zgodna z normą wektorową, $\|A\vec{x}\| \leqslant \|A\|\|\vec{x}\|$.)

• Mamy

$$\begin{aligned}\|(A+E)(\vec{x}+\vec{e})\| &= \|A\vec{e}+E\vec{x}+E\vec{e}\| \leqslant \|A\|\|\vec{e}\| + \|E\|\|\vec{x}\| + \|E\|\|e\| \\ &\leqslant (K_1+K_2+K_1K_2\nu)\,\nu\,\|A\|\|\vec{x}\| \lessapprox (K_1+K_2)(\|A\|\|A^{-1}\|)\,\nu\,\|A\vec{x}\|,\end{aligned}$$

gdzie użyliśmy nierówności $\|\vec{x}\| = \|A^{-1}A\vec{x}\| \leqslant \|A^{-1}\|\|A\vec{x}\|$.
□

(7) Wykaż, że naturalny algorytm obliczania cosinusa kąta pomiędzy wektorami $\vec{a}, \vec{b} \in \mathbb{R}^n$,

$$\cos(\vec{a},\vec{b}) \;=\; \frac{\sum_{j=1}^n a_j b_j}{\sqrt{\Big(\sum_{j=1}^n a_j^2\Big)\Big(\sum_{j=1}^n b_j^2\Big)}},$$

jest numerycznie poprawny. Oszacuj błąd względny wyniku w fl_ν.

• Dla iloczynu skalarnego mamy

$$fl_\nu\Big(\sum_{j=1}^n a_j b_j\Big) = \sum_{j=1}^n a_j b_j(1+\varepsilon_j), \qquad |\varepsilon_j| \lessapprox (n+2)\nu.$$

W mianowniku dodajemy liczby nieujemne, więc wynik obliczony jest nieco zaburzonym dokładnym wynikiem,

$$fl_\nu\Bigg(\sqrt{\Big(\sum_{j=1}^n a_j^2\Big)\Big(\sum_{j=1}^n b_j^2\Big)}\Bigg) \;=\; \sqrt{\Big(\sum_{j=1}^n a_j^2\Big)\Big(\sum_{j=1}^n b_j^2\Big)}(1+\alpha), \qquad |\alpha| \lessapprox (n+3)\nu.$$

Przyjmując $\tilde{b}_j = b_j(1+\varepsilon_j)$ za dane zaburzone dostajemy

$$fl_\nu\Big(\cos(\vec{a},\vec{b})\Big) \;=\; \frac{\sum_{j=1}^n a_j\tilde{b}_j}{\sqrt{\Big(\sum_{j=1}^n a_j^2\Big)\Big(\sum_{j=1}^n \tilde{b}_j^2\Big)}}\,(1+e), \qquad |e| \lessapprox 2(n+3)\nu,$$

gdzie do e 'wciągnęliśmy' błąd z dzielenia, α i sztucznie dorzucone błędy przy b_j w mianowniku. Wynik w fl_ν jest więc nieco zaburzonym dokładnym wynikiem dla dokładnych a_j i nieco zaburzonych b_j.

Aby oszacować błąd względny, przenosimy błąd $(1+\alpha)$ do licznika i dostajemy

$$\Big|fl_\nu\Big(\cos(\vec{a},\vec{b})\Big) - \cos(\vec{a},\vec{b})\Big| \;=\; \frac{\Big|\sum_{j=1}^n a_j b_j \beta_j\Big|}{\|\vec{a}\|_2\|\vec{b}\|_2} \;\lessapprox\; (2n+5)\Bigg(\frac{\sum_{j=1}^n |a_j b_j|}{\Big|\sum_{j=1}^n a_j b_j\Big|}\Bigg)\,\nu\,\Big|\cos(\vec{a},\vec{b})\Big|.$$

□

(8) Pokaż, że naturalny algorytm obliczania $\|A\vec{x}\|_2$ dla danej macierzy $A \in \mathbb{R}^{n\times n}$ i wektora $\vec{x} \in \mathbb{R}^n$ jest numerycznie poprawny. Dokładniej,

$$fl_\nu(\|A\vec{x}\|_2) \;=\; \|(A+E)\vec{x}\|_2,$$

gdzie $\|E\|_2 \lessapprox k(n)\nu\|A\|_2$ i $k(n) = 3\sqrt{n}(\frac{n}{2}+1)$. Ponadto, jeśli A jest nieosobliwa to

$$|fl_\nu(\|A\vec{x}\|_2) - \|A\vec{x}\|_2| \;\lessapprox\; k(n)\,\nu\,\Big(\|A\|_2\|A^{-1}\|_2\Big)\;\|A\vec{x}\|_2.$$

• Mamy

$$fl_\nu(\|A\vec{x}\|_2) = \sqrt{\sum_{i=1}^n \Big|\sum_{j=1}^n a_{i,j}x_j(1+\varepsilon_{i,j})\Big|^2(1+\eta_i)}\;(1+\delta) = (*),$$

gdzie $|\varepsilon_{i,j}| \lessapprox (n+2)\nu$, $|\eta_i| \lessapprox n\nu$, $|\delta| \leqslant \nu$, a stąd

$$(*) = \sqrt{\sum_{i=1}^n \Big|\sum_{j=1}^n a_{i,j}(1+\omega_{i,j})x_j\Big|^2},$$

gdzie $(1+\omega_{i,j}) = (1+\varepsilon_{i,j})(1+\eta_i)^{1/2}(1+\delta)$, $|\omega_{i,j}| \lessapprox 3(\frac{n}{2}+1)\nu$. Stąd traktując $\tilde{a}_{i,j} = a_{i,j}(1+\omega_{i,j})$ jako dane zaburzone dostajemy

$$\|E\|_2 = \Big(\sum_{i,j=1}^n |a_{i,j}|^2|\omega_{i,j}|^2\Big)^{1/2} \lessapprox 3(\tfrac{n}{2}+1)\|A\|_E \leqslant 3\sqrt{n}(\tfrac{n}{2}+1)\|A\|_2.$$

Aby pokazać drugą część zadania, wystarczy wykorzystać nierówność trójkąta;

$$\Big|\,\|(A+E)\vec{x}\|_2 - \|A\vec{x}\|_2\Big| \leqslant \|E\vec{x}\|_2 \leqslant \|E\|_2\|A^{-1}A\vec{x}\|_2 \leqslant k(n)\nu\|A\|_2\|A^{-1}\|_2\|A\vec{x}\|_2.$$

□

(9) Wykaż, że algorytmy obliczania wartości każdej z niewiadomych x i y w układzie dwóch równań liniowych z dwiema niewiadomymi za pomocą wzorów Cramera (tj. z wyznacznikami) są numerycznie poprawne.

(10) Rozpatrzmy dwa zadania, $S : F \to G$ i $T : G \to H$. Niech Φ i Ψ będą algorytmami dokładnymi rozwiązującymi, odpowiednio, zadania S i T. Oba algorytmy są numerycznie poprawne w zbiorach danych $F_0 \subset F$ i $G_0 \subset G$, przy czym $S(F_0) \subset G_0$, tzn. istnieją $\nu_0 > 0$, $K_1, K_2, K_3, K_3 > 0$, takie że dla $\nu \leqslant \nu_0$ i dowolnych $f \in F_0$ i $g \in G_0$ można wskazać $\tilde{f} \in F$ i $\tilde{g} \in G$ spełniające:

$$\|\tilde{f}-f\| \leqslant K_1\nu\|f\|, \qquad \|fl_\nu(\Phi(f)) - S(\tilde{f})\| \leqslant K_2\nu\|S(\tilde{f})\|,$$
$$\|\tilde{g}-g\| \leqslant K_3\nu\|g\|, \qquad \|fl_\nu(\Psi(g)) - T(\tilde{g})\| \leqslant K_4\nu\|T(\tilde{g})\|.$$

W danej przestrzeni unormowanej X, niech $\mathcal{B}(x,r) = \{x_1 \in X : \|x_1 - x\| \leqslant r\|x\|\}$. Załóżmy, że istnieją $\delta_0 > 0$ i $M > 0$ takie, że

(a) $\forall\delta \leqslant \delta_0 \quad \forall f \in F \qquad \mathcal{B}\Big(S(f),\delta)\Big) \subseteq S\Big(\mathcal{B}(f,\delta M)\Big)$, albo

(b) $\forall\delta \leqslant \delta_0 \quad \forall g \in G \qquad T\Big(\mathcal{B}(g,\delta)\Big) \subseteq \mathcal{B}\Big(T(g),\delta M)\Big)$.

Wykaż, że wtedy złożenie $\Psi \circ \Phi$ jest algorytmem dokładnym dla zadania złożenia $T \circ S : F \to H$, który jest też numerycznie poprawny w F_0. Dokładniej dla $f \in F_0$ istnieje $f^* \in F$ takie że: w przypadku (a)

$$\|f^* - f\| \leqslant K\nu\|f\|, \qquad \|fl_\nu\Big((\Psi\circ\Phi)(f)\Big) - (T\circ S)(f^*)\| \leqslant K_4\nu\|(T\circ S)(f^*)\|,$$

gdzie $K \lessapprox (K_2 + K_3)M + K_1$, a w przypadku (b)

$$\|f^* - f\| \leqslant K_1\nu\|f\|, \qquad \|fl_\nu\Big((\Psi \circ \Phi)(f)\Big) - (T \circ S)(f^*)\| \leqslant K\nu\|(T \circ S)(f^*)\|,$$

gdzie $K \lessapprox (MK_5 + K4)$.

• Ustalmy $f \in F_0$ i przyjmijmy $g = fl_\nu(\Phi(f))$. Wtedy

$$fl_\nu\Big((\Psi \circ \Phi)(f)\Big) = fl_\nu\Big(\Psi\Big(fl_\nu(\Phi(f))\Big)\Big) = fl_\nu(\Psi(g)),$$

co implikuje

$$\begin{aligned}\|\tilde{g} - S(\tilde{f})\| &\leqslant \|\tilde{g} - g\| + \|g - S(\tilde{f})\| \leqslant K_3\nu\|g\| + K_2\nu\|S(\tilde{f})\| \\ &\leqslant K_3\nu\Big(\|g - S(\tilde{f})\| + \|S(\tilde{f})\|\Big) + K_2\nu\|S(\tilde{f})\| \\ &\leqslant (K_2 + K_3 + K_2K_3\nu)\nu\|S(\tilde{f})\|.\end{aligned}$$

Przyjmijmy $K_5 = K_2 + K_3 + K_2K_3\nu \approx K_2 + K_3$. W przypadku (a) istnieje $f^* \in F$ takie, że $\|f^* - \tilde{f}\| \leqslant MK_5\nu\|\tilde{f}\|$ oraz $S(f^*) = \tilde{g}$. Stąd

$$\|f^* - f\| \leqslant \|f^* - \tilde{f}\| + \|\tilde{f} - f\| \leqslant MK_5\nu\|\tilde{f}\| + K_1\nu\|f\| \leqslant (MK_5 + K_1 + K_1K_5\nu)\nu\|f\|,$$

$$\|fl_\nu(\Psi \circ \Phi)(f) - (T \circ S)(f^*)\| = \|fl_\nu(\Psi(g)) - T(\tilde{g})\| \leqslant K_4\nu\|T(\tilde{g})\| = K_4\nu\|(T \circ S)(f^*)\|.$$

W przypadku (b) zachodzi $\|T(\tilde{g}) - T(S(\tilde{f}))\| \leqslant MK_5\nu\|T(S(\tilde{f}))\|$. Stąd, przyjmując $f^* = \tilde{f}$ mamy $\|f^* - f\| \leqslant K_1\nu\|f\|$ i

$$\begin{aligned}&\|fl_\nu(\Psi \circ \Phi)(f) - (T \circ S)(f^*)\| = \|fl_\nu(\Psi(g)) - T(S(\tilde{f}))\| \\ &\quad \leqslant \|fl_\nu(\Psi(g)) - T(\tilde{g})\| + \|T(\tilde{g}) - T(S(\tilde{f}))\| \leqslant K_4\nu\|T(\tilde{g})\| + MK_5\nu\|T(S(\tilde{f}))\| \\ &\quad \leqslant (MK_5 + K_4 + MK_4K_5\nu)\nu\|(T \circ S)(f^*)\|.\end{aligned}$$

□

(11) Rozpatrzmy arytmetykę stałoprzecinkową fx_ν, gdzie różnica pomiędzy liczbą rzeczywistą x a jej reprezentacją $fx_\nu(x)$ wynosi $|x - fx_\nu(x)| \leqslant \nu$. Powiemy, że algorytm Φ realizujący przekształcenie $S : F \to G$ jest numerycznie poprawny w zbiorze danych $F_0 \subseteq F$ gdy istnieją K_1, K_2 o następującej własności: dla dowolnych danych $f \in F_0$ i dostatecznie silnej arytmetyki ($\nu \leqslant \nu_0$) istnieją dane $\tilde{f} \in F$ takie, że

$$\|\tilde{f} - f\| \leqslant K_1\nu \quad \text{oraz} \quad \|fx_\nu(\Phi(f)) - S(\tilde{f})\| \leqslant K_2\nu.$$

($fx_\nu(\Phi(f))$ jest tu wynikiem zwracanym przez Φ w arytmetyce fx_ν dla danych f.)

Niech teraz Φ_1, Φ_2 będą algorytmami numerycznie poprawnymi realizującymi odpowiednio przekształcenia $S_1 : X \to Y$ oraz $S_2 : Y \to Z$. Wykaż, że jeśli S_2 spełnia warunek Lipschitza w Y to złożenie $\Phi_2 \circ \Phi_1$ realizujące przekształcenie $S = S_2 \circ S_1 : X \to Z$ jest też algorytmem numerycznie poprawnym.

• Z numerycznej poprawności Φ_1 mamy istnienie $\tilde{f} \in X$ takiego, że

$$\|\tilde{f} - f\| \leqslant K_1\nu \quad \text{i} \quad \|fx_\nu(\Phi_1(f)) - S_1(\tilde{f})\| \leqslant K_2\nu.$$

Oznaczmy $g = S_1(\tilde{f})$. Z numerycznej poprawności Φ_2 mamy z kolei istnienie $\tilde{g} \in Y$ takiego, że

$$\|\tilde{g} - fx_\nu(\Phi_1(f))\| \leqslant K_3\nu \quad \text{i} \quad \|fx_\nu((\Phi_2 \circ \Phi_1)(f)) - S_2(\tilde{g}))\| \leqslant K_4\nu.$$

Zauważmy, że

$$\|\tilde{g} - g\| \leqslant \|\tilde{g} - fx_\nu(\Phi_1(f))\| + \|fx_\nu(\Phi_1(f)) - g\| \leqslant (K_3 + K_2)\nu.$$

Stąd, oznaczając przez L stałą Lipschitza dla S_2 dostajemy

$$\begin{aligned}
&\|fx_\nu((\Phi_2 \circ \Phi_1)(f)) - (S_2 \circ S_1)(\tilde{f})\| \\
&\quad\leqslant \|fx_\nu((\Phi_2 \circ \Phi_1)(f)) - S_2(\tilde{g})\| + \|S_2(\tilde{g}) - S_2(g)\| \\
&\quad\leqslant K_4\nu + L\|\tilde{g} - g\| \leqslant K_4\nu + L(K_3 + K_2)\nu = K\nu,
\end{aligned}$$

gdzie $K = K_4 + L(K_3 + K_2)$, czyli numeryczną poprawność złożenia $\Phi_2 \circ \Phi_1$.
□

(12) Wykaż, że dla normy pierwszej macierzy $A = (a_{i,j}) \in \mathbb{R}^{m\times n}$ mamy

$$\|A\|_1 \;=\; \max_{1\leqslant j\leqslant n} \sum_{i=1}^{m} |a_{i,j}|,$$

a dla normy nieskończoność

$$\|A\|_\infty \;=\; \max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{i,j}|.$$

• Dla normy pierwszej mamy

$$\begin{aligned}
\|A\vec{x}\|_1 \;&=\; \sum_{i=1}^{m} \Big| \sum_{j=1}^{n} a_{i,j}x_j \Big| \leqslant \sum_{j=1}^{n}\sum_{i=1}^{m} |a_{i,j}||x_j| \\
&=\; \Big(\max_{1\leqslant j\leqslant n} \sum_{i=1}^{m} |a_{i,j}| \Big) \Big(\sum_{j=1}^{n} |x_j| \Big) = \Big(\max_{1\leqslant j\leqslant n} \sum_{i=1}^{m} |a_{i,j}| \Big) \|\vec{x}\|_1.
\end{aligned}$$

Z drugiej strony dla wektora $\vec{x}^* = \vec{e}_{j^*}$, gdzie j^* jest indeksem kolumny dla której max jest osiągnięte, mamy

$$\|A\vec{x}^*\|_1 = \Big(\sum_{i=1}^{m} |a_{i,j^*}| \Big) \|\vec{x}^*\|_1.$$

Z kolei dla normy nieskończoność,

$$\begin{aligned}
\|A\vec{x}\|_\infty \;&=\; \max_{1\leqslant i\leqslant m} \Big| \sum_{j=1}^{n} a_{i,j}x_j \Big| \leqslant \max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{i,j}||x_j| \\
&\leqslant\; \Big(\max_{1\leqslant j\leqslant n} |x_j| \Big) \Big(\max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{i,j}| \Big) = \Big(\max_{1\leqslant i\leqslant m} \sum_{j=1}^{n} |a_{i,j}| \Big) \|\vec{x}\|_\infty
\end{aligned}$$

i z drugiej strony mamy wszędzie powyżej równości dla $\vec{x}^*_j = |a_{i^*,j}|/a_{i^*,j}$, $1 \leqslant j \leqslant n$, (albo $\vec{x}^*_j = 0$ jeśli $a_{i^*,j} = 0$), mamy

$$\|A\vec{x}^*\|_\infty = \Big(\sum_{j=1}^{n} |a_{i^*,j}| \Big) \|\vec{x}^*\|_\infty.$$

□

(13) Wykaż, że normę drugą macierzy można wyrazić jako

$$\|A\|_2 = \max_{\lambda \in \mathrm{Sp}(A^TA)} \sqrt{\lambda},$$

gdzie $\mathrm{Sp}(B)$ jest zbiorem wszystkich wartości własnych macierzy B.

• Z algebry liniowej wiadomo, że jeśli macierz kwadratowa $B \in \mathbb{R}^{n\times n}$ spełnia $B = B^T \geqslant 0$ to istnieje w $\mathbb{R}^n$ baza ortonormalna wektorów własnych macierzy B, a odpowiednie wartości własne są rzeczywiste i nieujemne. Oczywiście, $(A^TA) = (A^TA)^T \geqslant 0$. Niech $\{\vec{x}_i, \lambda_i\}$, $1 \leqslant i \leqslant n$, będą parami własnymi macierzy A^TA, przy czym $\langle \vec{x}_i, \vec{x}_j\rangle_2 = \delta_{i,j}$ i $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_n \geqslant 0$. Zapisując dowolny wektor w tej bazie, $\vec{x} = \sum_{i=1}^n a_i\vec{x}_i$ mamy

$$\begin{aligned}\|A\vec{x}\|_2^2 &= \langle A\vec{x}, A\vec{x}\rangle_2 = \langle A^TA\vec{x}, \vec{x}\rangle_2 = \Big\langle \sum_{i=1}^n a_i\lambda_i\vec{x}_i, \sum_{j=1}^n a_j\vec{x}_j \Big\rangle_2 \\ &= \sum_{i,j=1}^n \lambda_i a_i a_j \langle \vec{x}_i, \vec{x}_j\rangle_2 = \sum_{i=1}^n \lambda_i |a_i|^2 \leqslant \lambda_1 \sum_{i=1}^n |a_i|^2 = \lambda_1 \|\vec{x}\|_2^2.\end{aligned}$$

Stąd $\|A\|_2 \leqslant \sqrt{\lambda}$. Z drugiej strony, mamy

$$\|A\vec{x}_1\|_2 = \sqrt{\langle A\vec{x}_1, A\vec{x}_1\rangle} = \sqrt{\langle A^TA\vec{x}_1, \vec{x}_1\rangle_2} = \sqrt{\lambda}_1\sqrt{\langle \vec{x}_1, \vec{x}_1\rangle_2} = \sqrt{\lambda}_1\|\vec{x}_1\|_2.$$

□

(14) Wykaż, że dla macierzy $A \in \mathbb{R}^{m\times n}$ mamy

$$\|A\|_2 = \sup_{\|\vec{z}\|_2=1} \sup_{\|\vec{y}\|_2=1} \left|\vec{y}^TA\vec{z}\right|.$$

Wywnioskuj stąd, że $\|A\|_2 = \|A^T\|_2$.

• Dla $\|\vec{y}\|_2 = \|\vec{z}\|_2 = 1$ mamy $|\vec{y}^TA\vec{z}| = |\langle A\vec{z}, \vec{y}\rangle_2| \leqslant \|A\vec{z}\|_2$, co dowodzi nierówności '$\geqslant$'. Dla dowodu odwrotnej nierówności zauważmy, że jeśli dla $\vec{z}$ takiego, że $\|\vec{z}\|_2 = 1$ i $A\vec{z} \neq \vec{0}$ weźmiemy $\vec{y} = A\vec{z}/\|A\vec{z}\|_2$ to

$$\vec{y}^TA\vec{z} = \frac{(A\vec{z})^T(A\vec{z})}{\|A\vec{z}\|_2} = \|A\vec{z}\|_2.$$

Aby pokazać, że $\|A^T\|_2 = \|A\|_2$, wystarczy zauważyć, że

$$\vec{y}^TA\vec{z} = \left(\vec{y}^TA\vec{z}\right)^T = \vec{z}^TA^T\vec{y}.$$

□

(15) Pokaż, że dla $A \in \mathbb{R}^{m\times n}$ macierze A^TA i AA^T mają takie same niezerowe wartości własne, a podprzestrzenie własne im odpowiadające mają ten sam wymiar. Wywnioskuj stąd, że $\|A^T\|_2 = \|A\|_2$.

• Jeśli $\lambda > 0$ jest wartością własną macierzy A^TA, a $\vec{x} \neq \vec{0}$ odpowiadającym jej wektorem własnym, czyli $A^TA\vec{x} = \lambda\vec{x}$, to $A\vec{x} \neq \vec{0}$ oraz $(AA^T)(A\vec{x}) = \lambda(A\vec{x})$, czyli λ jest też wartością własną macierzy AA^T, a $A\vec{x}$ jest jej odpowiadającym wektorem własnym. Podobnie, jeśli λ i $\vec{x}$ jest parą własną macierzy AA^T, przy czym $\lambda > 0$, to λ i $A^T\vec{x}$ jest parą własną macierzy A^TA.

Niech V_1 i V_2 będą podprzestrzeniami własnymi macierzy A^TA i AA^T odpowiadającymi wartości własnej $\lambda \neq 0$. Z poprzednich rozważań mamy, że $A(V_1) \subseteq V_2$ i $A^T(V_2) \subseteq (V_1)$. Macierz A jest różnowartościowa na V_1. Jeśli bowiem $\vec{x}_1, \vec{x}_2 \in V_1$ i $A\vec{x}_1 = A\vec{x}_2$ to $A(\vec{x}_1 - \vec{x}_2) = \vec{0}$, co implikuje $\vec{x}_1 = \vec{x}_2$, bo $(A^TA)(\vec{x}_1 - \vec{x}_2) = \lambda(\vec{x}_1 - \vec{x}_2)$. Stąd $\dim V_1 = \dim A(V_1) \leqslant \dim V_2$. Podobnie $\dim V_2 \leqslant \dim V_1$, czyli $\dim V_1 = \dim V_2$.

Równość wymiarów podprzestrzeni własnych odpowiadających niezerowym wartościom własnym λ wynika z faktu, że zarówno A jak i A^T są różnowartościowe na

Druga część zadania wynika teraz z (13).

□

(16) Wykaż, że dla dowolnej macierzy $A \in \mathbb{R}^{m\times n}$ mamy

$$\|A\|_2 \leqslant \| \, |A| \, \|_2 \leqslant \|A\|_E \leqslant \sqrt{\min(m,n)}\, \|A\|_2,$$

gdzie $|A| = (|a_{i,j}|)_{i,j}$. Stąd wywnioskuj, że

$$\|A - rd(A)\|_2 \leqslant \|A - rd(A)\|_E \leqslant \nu\, \|A\|_E \leqslant \sqrt{\min(m,n)}\, \nu\, \|A\|_2.$$

• Pierwsza nierówność wynika bezpośrednio z definicji, bowiem

$$\begin{aligned} \|A\|_2^2 &= \sup_{\|\vec{x}\|_2=1} \|A\vec{x}\|_2^2 = \sup_{\|\vec{x}\|_2=1} \sum_{i=1}^{m} \Big| \sum_{j=1}^{n} a_{i,j}x_j \Big|^2 \leqslant \sup_{\|\vec{x}\|_2=1} \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} |a_{i,j}||x_j| \Big)^2 \\ &= \sup_{\|\vec{x}\|_2=1} \sum_{i=1}^{m} \Big(\sum_{j=1}^{n} |a_{i,j}|x_j \Big)^2 = \| \, |A| \, \|_2^2. \end{aligned}$$

Niech $\vec{a}_i^T$ będą wektorami-wierszami macierzy A, tak że $A = (\vec{a}_1, \vec{a}_2, \ldots, \vec{a}_m)^T$. Wykorzystując nierówność Schwarza dla ciagów mamy

$$\|A\vec{x}\|_2^2 = \sum_{i=1}^{m} \Big| \langle \vec{a}_i, \vec{x} \rangle_2 \Big|^2 \leqslant \sum_{i=1}^{m} \|\vec{a}_i\|_2^2 \|\vec{x}\|_2^2 = \|A\|_E^2 \|\vec{x}\|_2^2.$$

Ponieważ A jest dowolne to powyższa nierówność zachodzi też dla $A = |A|$.

Aby pokazać ostatnią nierówność, niech $\vec{b}_j$ będą wektorami-kolumnami macierzy A, tak że $A = (\vec{b}_1, \vec{b}_2, \ldots, \vec{b}_n)$. Wtedy

$$\|A\|_E^2 = \sum_{j=1}^{n} \|\vec{b}_j\|_2^2 = \sum_{j=1}^{n} \|A\vec{e}_j\|_2^2 \leqslant \sum_{j=1}^{n} \|A\|_2^2 = n\|A\|_2^2,$$

czyli $\|A\|_E \leqslant \sqrt{n}\|A\|_2$. Biorąc macierz transponowaną i wykorzystując wynik z (15) mamy z kolei

$$\|A\|_E = \|A^T\|_E \leqslant \sqrt{m}\|A^T\|_2 = \sqrt{m}\|A\|_2.$$

□

(17) Pokaż, że dla dowolnej macierzy $A \in \mathbb{R}^{m\times n}$ mamy

$$\|A\|_2^2 \leqslant \|A\|_1 \, \|A\|_\infty.$$

• Jeśli λ jest wartością własną macierzy B, a $\vec{x} \neq \vec{0}$ jej odpowiadającym wektorem własnym to $\|B\vec{x}\| = |\lambda|\|\vec{x}\|$. To oznacza, że dla norm macierzowych indukowanych normami wektorowymi zachodzi $\|B\| \geqslant |\lambda|$.

Wobec (13) mamy teraz

$$\|A\|_2^2 = \max_{\lambda \in \mathrm{Sp}(A^TA)} \lambda \leqslant \|A^TA\|_\infty \leqslant \|A^T\|_\infty \|A\|_\infty = \|A\|_1 \|A\|_\infty.$$

□

(18) Macierz A dana jest wpostaci blokowej

$$A = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,n} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,n} \\ \vdots & \vdots & & \vdots \\ A_{m,1} & A_{m,2} & \cdots & A_{m,n} \end{pmatrix}.$$

Wykaż, że dla dowolnych i, j mamy $\|A_{i,j}\|_p \leqslant \|A\|_p$ dla $1 \leqslant p \leqslant +\infty$.

• Wykorzystamy wprost definicję normy macierzy. Ustalmy i, j. Niech $\vec{x} = (\vec{0}, \vec{z}, \vec{0})^T$ będzie wektorem o wymiarze równym liczbie kolumn macierzy A, który ma elementy niezerowe jedynie w miejscach odpowiadających numerom kolumn macierzy $A_{i,j}$. Wtedy $\vec{z}$ ma wymiar równy liczbie kolumn macierzy $A_{i,j}$ oraz

$$A\vec{x} = (A_{1,j}\vec{z}, \ldots, A_{m,j}\vec{z})^T.$$

Stąd

$$\|A\|_p = \sup_{\vec{x}\neq 0} \frac{\|A\vec{x}\|_p}{\|\vec{x}\,\|_p} \geqslant \sup_{\vec{x}=(\vec{0},\vec{z},\vec{0})^T \neq \vec{0}} \frac{\left(\sum_{k=1}^m \|A_{k,j}\vec{z}\,\|_p^p\right)^{1/p}}{\|\vec{z}\,\|_p} \geqslant \sup_{\vec{z}\neq\vec{0}} \frac{\|A_{i,j}\vec{z}\,\|_p}{\|\vec{z}\,\|_p} = \|A_{i,j}\|_p.$$

□

(19) Pokaż, że dla danej macierzy A i permutacji P istnieje co najwyżej jeden rozkład $PA = LU$ na macierz trójkątną dolną L z jedynkami na przekątnej i na macierz trójkątną górną U.

• Jeśli $L_1U_1 = L_2U_2$ to $U_1U_2^{-1} = L_2^{-1}L_1$. Teraz wystarczy zauważyć, że macierz odwrotna do górnej (dolnej) trójkątnej jest też górną (dolną) trójkątną, a jeśli dodatkowo na głównej przekątnej miała jedynki, to odwrotna też zachowuje tę własność. Podobnie, iloczyn dwóch macierzy górnych (dolnych) trójkątnych jest macierzą górną (dolną) trójkatną, a jeśli dodatkowo miały one jedynki na głównej przekątnej to iloczyn tę własność zachowuje. Stąd macierz $U_1U_2^{-1}$ jest trójkątna górna, a $L_1^{-1}L_2$ jest trójkatna dolna z jedynkami na głównej przekątnej. Z ich równości wynika, że obie muszą być macierzą identycznościową i w konsekwencji $U_1 = U_2$ i $L_1 = L_2$.
□

(20) Pokaż wykonalność algorytmu przeganiania dla macierzy trójdiagonalnych

$$A = \begin{pmatrix} a_1 & c_1 & & & \\ b_2 & a_2 & c_2 & & \\ & b_3 & a_3 & c_3 & \\ & & \ddots & \ddots & \ddots \\ & & & b_n & a_n \end{pmatrix}$$

z dominującą przekątną, tzn. gdy $|a_i| > |b_i| + |c_i|$ dla $1 \leqslant i \leqslant n$ ($b_0 = 0 = c_n$). Ponadto, w każdym kroku eliminacyjnym, element maksymalna zwiększa się co najwyżej dwukrotnie.

• W pierwszym kroku eliminacyjnym mamy $l_{2,1} = b_2/a_1$, przy czym $a_1 \neq 0$ bo $|a_1| > |c_1|$. Przy eliminacji elementu $(2,1)$ macierzy A zmienia się jedynie element $(2,2)$ i przyjmuje wartość $a'_2 = a_2 - c_1 l_{2,1} = a_2 - c_1 \frac{b_2}{a_1}$. Wykorzystując dominację przekątnej mamy

$$|a'_2| \geqslant |a_2| - |c_1|\frac{|b_2|}{|a_1|} > \Big(|b_2| + |c_2|\Big) - |c_1|\frac{|b_2|}{|a_1|} = |b_2|\Big(1 - \frac{|c_1|}{|a_1|}\Big) + |c_2| \geqslant |c_2|,$$

a to oznacza, że macierz powstała po pierwszym kroku eliminacji ma też dominującą przekątną. Ponadto,

$$|a'_2| \leqslant |a_2| + |c_1|\frac{|b_2|}{|a_1|} \leqslant |a_2| + |a_1|\frac{|b_2|}{|a_1|} = |a_2| + |b_2|,$$

co z kolei oznacza, że maksymalny element zwiększa się co najwyżej dwukrotnie.

Nietrudno zauważyć, że to samo rozumowanie stosuje się indukcyjnie w kolejnych krokach eliminacyjnych. Algorytm jest więc wykonalny bez przestawień wierszy.
□

(21) Pokaż, że jeśli eliminację Gaussa z wyborem elementu głównego w kolumnie zastosujemy do nieosobliwej macierzy trójdiagonalnej A jak w zadaniu (20) to wzrost elementu maksymalnego macierzy nie będzie zależał od n. Dokładniej,

$$\max_{i,j,k} |a^{(k)}_{i,j}| \;\leqslant\; 2 \max_{i,j} |a_{i,j}|.$$

• Oznaczmy przez M wartość maksymalnego elementu macierzy wyjściowej $A = A^{(0)}$. Użyjemy indukcji po k aby pokazać, że część macierzy $A^{(k)}$, która nie uległa jeszcze eliminacji jest postaci

$$B = \begin{pmatrix} b_{1,1} & b_{1,2} & & & \\ b_{2,1} & b_{2,2} & b_{2,3} & & \\ & b_{3,2} & b_{3,3} & b_{3,4} & \\ & & \ddots & \ddots & \ddots \end{pmatrix} \in \mathbb{R}^{(n-k)\times(n-k)},$$

gdzie $|b_{1,1}| \leqslant 2M$, a moduły pozostałych niezerowych elementów są nie większe od M. Oczywiście, macierz $A^{(0)}$ ma powyższą własność. Załóżmy, że własność zachodzi dla $k \geqslant 0$. W kroku $k+1$ mamy dwa przypadki. Jeśli elementem głównym jest $b_{1,1}$ to po eliminacji zmienione elementy wynoszą $b'_{2,1} = 0$, $|b'_{2,2}| \leqslant |b_{2,2}| + |b_{1,2}| \leqslant 2M$, oraz $|b'_{2,3}| = |b_{2,3}|$. Jeśli zaś elementem głównym jest $b_{2,1}$ to zamieniamy wiersze pierwszy z drugim i mamy to samo, czyli $b'_{2,1} = 0$, $|b'_{2,2}| \leqslant |b_{2,1}| + |b_{2,2}| \leqslant 2M$, oraz $|b'_{2,3}| \leqslant |b_{2,3}|$. Ostatecznie, macierz $U = (u_{i,j}) = A^{(n-1)}$ jest trójdiagonalna, dokładniej $u_{i,j} = 0$ dla $i \leqslant j \leqslant i+2$, gdzie na diagonali mamy $|u_{i,i}| \leqslant 2M$, a pozostałe elementy $|u_{i,j}| \leqslant M$.
□

(22) Pokaż, że dla nieosobliwej macierzy Hessenberga ($a_{i,j} = 0$ dla $i \geqslant j+2$) eliminacja Gaussa z wyborem elementu głównego w kolumnie daje

$$\max_{i,j,k} |a^{(k)}_{i,j}| \;\leqslant\; (k+1) \max_{i,j} |a_{i,j}|.$$

• Podobnie jak w rozwiązaniu zadania (21), niech M będzie maksymalną wartością w macierzy wyjściowej. Użyjemy indukcji po k aby pokazać, że część macierzy $A^{(k)}$, która nie uległa jeszcze eliminacji jest postaci

$$B = \begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} & b_{1,4} & \cdots \\ b_{2,1} & b_{2,2} & b_{2,3} & b_{2,4} & \cdots \\ & b_{3,2} & b_{3,3} & b_{3,4} & \cdots \\ & & b_{4,3} & b_{4,4} & \cdots \\ & & & \ddots & \ddots \end{pmatrix} \in \mathbb{R}^{(n-k)\times(n-k)},$$

gdzie $|b_{1,j}| \leqslant (k-1)M$ dla $1 \leqslant j \leqslant n-k$ i $|b_{i,j}| \leqslant M$ dla $i \leqslant j \leqslant n-k$, $2 \leqslant i \leqslant n-k$. Oczywiście, ta własność zachodzi dla $k = 0$. Załóżmy, że własność zachodzi dla $k \geqslant 0$. Jeśli elementem głównym jest $b_{1,1}$ to nie zamieniamy wierszy i ńowe" elementy wynoszą $b'_{2,1} = 0$ oraz dla $i \leqslant j \leqslant n-k$ mamy $|b'_{2,j}| \leqslant |b_{2,j}| + |b_{1,j}| \leqslant M + (k-1)M = kM$. Podobnie, jeśli zamieniamy wiersze pierwszy z drugim to znowu $b'_{2,1} = 0$ oraz dla $i \leqslant j \leqslant n-k$ mamy $|b'_{2,j}| \leqslant |b_{1,j}| + |b_{2,j}| \leqslant (k-1)M + M = KM$.
□

(23) Pokaż, że jeśli macierz pełna A ma dominującą przekątną, tzn.

$$2\,|a_{i,i}| \;>\; \sum_{j=1}^{n} |a_{i,j}|, \qquad 1 \leqslant i \leqslant n,$$

to jest to macierz nieosobliwa. Ponadto, eliminacja Gaussa jest dla takich macierzy wykonalna bez przestawień wierszy.

• Pokażemy najpierw nieosobliwość macierzy A. Załóżmy, że $A\vec{x} = \vec{0}$ dla pewnego $\vec{x} \neq \vec{0}$. Możemy założyć, że $\|\vec{x}\|_\infty = 1$. Niech $|x_s| = \|\vec{x}\|_\infty$. Wtedy s-ty element wektora $A\vec{x}$ wynosi

$$0 = \Big|\sum_{j=1}^{n} a_{s,j}x_j\Big| \geqslant |a_{s,s}x_s| - \sum_{j\neq s} |a_{s,j}x_j| \geqslant |a_{s,s}| - \sum_{j\neq s} |a_{s,j}|,$$

co przeczy dominacji przekątnej.

Teraz wystarczy pokazać, że macierze powstające w kolejnych krokach eleminacji Gaussa są macierzami o dominującej przekątnej, bo wtedy elementy diagonalne są niezerowe. W tym celu, wystarczy rozpatrzyć pierwszy krok i zobaczyć co się dzieje gdy zerujemy element $(2,1)$. Elementy w drugim wierszu wynoszą $a'_{2,1} = 0$ oraz

$$a'_{2,j} = a_{2,j} - a_{1,j}\frac{a_{2,1}}{a_{1,1}} \qquad \text{dla} \qquad 2 \leqslant j \leqslant n.$$

Stąd

$$\begin{aligned} \sum_{j=3}^{n} |a'_{2,j}| \;&\leqslant\; \sum_{j=3}^{n} |a_{2,j}| + \frac{|a_{2,1}|}{|a_{1,1}|}\sum_{j=3}^{n} |a_{1,j}| \;<\; (|a_{2,2}| - |a_{2,1}|) + \frac{|a_{2,1}|}{|a_{1,1}|}(|a_{1,1}| - |a_{1,2}|) \\ &=\; |a_{2,2}| - |a_{1,2}|\frac{|a_{2,1}|}{|a_{1,1}|} \;\leqslant\; \left|a_{2,2} - a_{1,2}\frac{a_{2,1}}{a_{1,1}}\right| \;=\; |a'_{2,2}|, \end{aligned}$$

co należało pokazać.
□

(24) Wykaż numeryczną poprawność rozkładu $A = LU$ za pomocą eliminacji Gaussa bez przestawień wierszy dla macierzy A z dominującą przekątną (wierszowo).

• Wystarczy zauważyć, że w każdym kroku eliminacyjnym element maksymalny macierzy zwiększa się co najwyżej dwukrotnie. Rzeczywiście, korzystając z zadania (23) mamy, że

$$|a'_{2,j}| \leqslant |a_{2,j}| + |a_{1,j}|\frac{|a_{2,1}|}{|a_{1,1}|} \leqslant |a_{2,j}| + |a_{2,1}|, \qquad 2 \leqslant j \leqslant n,$$

bo $|a_{1,j}| < |a_{1,1}|$.
□

(25) Jeśli

$$(A + E)\vec{z} = \vec{b},$$

gdzie $\|E\|_p \leqslant K\nu\|A\|_p$, to oczywiście dla residuum $\vec{r} = \vec{b} - A\vec{z}$ mamy

$$\|\vec{r}\|_p \leqslant K\nu\|A\|_p\|\vec{z}\|_p.$$

Pokaż, że dla $p = 1, 2, \infty$ zachodzi też twierdzenie odwrotne, tzn. jeśli spełniony jest warunek $\|\vec{r}\|_p \leqslant K\nu\|A\|_p\|\vec{z}\|_p$ to istnieje macierz pozornych zaburzeń E taka, że $\|E\|_p \leqslant K\nu\|A\|_p$ oraz spełniona jest równość $(A + E)\vec{z} = \vec{b}$.

• Dla $p = 1$ bierzemy

$$E = \frac{\vec{r}}{\|\vec{z}\|_1}(\mathrm{sgn}z_1, \ldots, \mathrm{sgn}z_n).$$

Wtedy $\|E\|_1 \leqslant \|\vec{r}\|_1/\|\vec{z}\|_1 \leqslant K\nu\|A\|_1$ i wobec tego, że $(\mathrm{sgn}z_1, \ldots, \mathrm{sgn}z_n)\vec{z} = \|\vec{z}\|_1$, mamy też $(A + E)\vec{z} = A\vec{z} + \vec{r} = \vec{b}$.

Dla $p = 2$ bierzemy

$$E = \frac{\vec{r}\,\vec{z}^T}{\|\vec{z}\|_2^2}.$$

Wtedy znowu $\|E\|_2 \leqslant \|\vec{r}\|_2/\|\vec{z}\|_2 \leqslant K\nu\|A\|_2$ i $(A + E)\vec{z} = A\vec{z} + \vec{r}(\vec{z}^T\vec{z})/\|\vec{z}\|_2^2 = A\vec{z} + \vec{r} = \vec{b}$.

Z kolei dla $p = +\infty$ bierzemy

$$E = \frac{\vec{r}\,(\mathrm{sgn}z_k)\vec{e}_k^T}{\|\vec{z}\|_\infty},$$

gdzie k jest tym indeksem, dla którego $|z_k| = \|\vec{z}\|_\infty$. Wtedy w oczywisty sposób $\|E\|_\infty \leqslant \|\vec{r}\|_\infty/\|\vec{z}\|_\infty \leqslant K\nu\|A\|_\infty$. Mamy też $(\mathrm{sgn}z_k)\vec{e}_k^T\vec{z} = |z_k| = \|\vec{z}\|_\infty$ i dlatego $(A + E)\vec{z} = A\vec{z} + \vec{r} = \vec{b}$.
□

(26) Dana jest macierz

$$A = \begin{pmatrix} 1 & 2 & 0 \\ 2 & 1 & 2 \\ 0 & 2 & 1 \end{pmatrix}.$$

Wyznacz czynniki P, L, U rozkładu $PA = LU$. Za pomocą wyznaczonego rozkładu rozwiąż układ równań $A\vec{x} = \vec{b}$ dla $\vec{b} = [6, 4, -4]^T$.

(27) Stosując (dosłownie) algorytm eliminacji Gaussa z wyborem elementu głównego w kolumnie dokonaj rozkładu macierzy

$$A = \begin{pmatrix} -1 & 1 & 0 & -3 \\ 1 & 0 & 3 & 1 \\ 0 & 1 & -1 & -1 \\ 3 & 0 & 1 & 2 \end{pmatrix}$$

na iloczyn $PA = L\,R$, gdzie P jest macierzą permutacji, L macierzą trójkątną dolną z jedynkami na głównej przekątnej, a R macierzą trójkątną górną.

• Współczynniki $l_{i,j}$ będziemy 'odkładać' w macierzy L, która na początku jest macierzą zerową, a macierz P w wektorze permutacji p o początkowej zawartości $(1,2,3,4)^T$. Najpierw przestawiamy wiersze 1 z 4 w wektorze p i macierzy A (bo element główny jest w $A(4,1)$) i dostajemy konfigurację

$$p = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 3 & 0 & 1 & 2 \\ 1 & 0 & 3 & 1 \\ 0 & 1 & -1 & -1 \\ -1 & 1 & 0 & -3 \end{pmatrix}, \quad L = \begin{pmatrix} & & & \\ & & & \\ & & & \\ & & & \end{pmatrix}.$$

(Puste miejsca oznaczają, że tam są zera.) Wpisujemy odpowiednie współczynniki do pierwszej kolumny macierzy L i eliminujemy wyrazy w pierwszej kolumnie macierzy A. Dostajemy

$$p = \begin{pmatrix} 4 \\ 2 \\ 3 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 3 & 0 & 1 & 2 \\ & 0 & 8/3 & 1/3 \\ & 1 & -1 & -1 \\ & 1 & 1/3 & -7/3 \end{pmatrix}, \quad L = \begin{pmatrix} & & & \\ 1/3 & & & \\ 0 & & & \\ -1/3 & & & \end{pmatrix}.$$

Teraz element główny jest w $A(3,2)$ więc musimy przestawić wiersze 2 z 3 w wektorze p, macierzy A i macierzy L,

$$p = \begin{pmatrix} 4 \\ 3 \\ 2 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 3 & 0 & 1 & 2 \\ & 1 & -1 & -1 \\ & 0 & 8/3 & 1/3 \\ & 1 & 1/3 & -7/3 \end{pmatrix}, \quad L = \begin{pmatrix} & & & \\ 0 & & & \\ 1/3 & & & \\ -1/3 & & & \end{pmatrix}.$$

Po następnym wpisaniu współczynników do L i eliminacji w drugiej kolumnie macierzy A dostajemy konfigurację

$$p = \begin{pmatrix} 4 \\ 3 \\ 2 \\ 1 \end{pmatrix}, \quad A = \begin{pmatrix} 3 & 0 & 1 & 2 \\ & 1 & -1 & -1 \\ & & 8/3 & 1/3 \\ & & 4/3 & -4/3 \end{pmatrix}, \quad L = \begin{pmatrix} & & & \\ 0 & & & \\ 1/3 & 0 & & \\ -1/3 & 1 & & \end{pmatrix}.$$

Element główny jest w $A(3,3)$, więc w ostatnim kroku nie będzie przestawień. Wpisujemy ostatni współczynnik do L, wpisujemy jedynki na jej przekątnej, eliminujemy element $A(4,3)$ i ostatecznie mamy rozkład $PA = LU$ postaci

$$\begin{pmatrix} & & & 1 \\ & & 1 & \\ & 1 & & \\ 1 & & & \end{pmatrix} \begin{pmatrix} -1 & 1 & 0 & -3 \\ 1 & 0 & 3 & 1 \\ 0 & 1 & -1 & -1 \\ 3 & 0 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 1 & & & \\ 0 & 1 & & \\ 1/3 & 0 & 1 & \\ -1/3 & 1 & 1/2 & 1 \end{pmatrix} \begin{pmatrix} 3 & 0 & 1 & 2 \\ & 1 & -1 & -1 \\ & & 8/3 & 1/3 \\ & & & -3/2 \end{pmatrix}$$

□

(28) Dana jest macierz

$$A = \begin{pmatrix} 1 & 1 & 0 & -1 \\ 0 & 2 & -1 & 0 \\ 1 & 1 & 3 & 1 \\ -1 & -3 & -2 & 3 \end{pmatrix} \in \mathbb{R}^{4\times 4}.$$

Stosując eliminację Gaussa, znajdź czynniki rozkładu $PA = LU$ tej macierzy, gdzie $L \in \mathbb{R}^{4\times 4}$ jest macierzą trójkątną dolną z jedynkami na głównej przekątnej i pozostałymi elementami o module nie większym od jedności, $U \in \mathbb{R}^{4,4}$ jest trójkątna górna, a $P \in \mathbb{R}^{4\times 4}$ jest macierzą permutacji.

(29) Niech

$$A = \begin{pmatrix} 3 & 2 & 1 \\ 30 & 22 & 11 \\ 0 & 20 & 110 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

Znajdź metodą eliminacji Gaussa z wyborem elementu głównego w kolumnie macierze L, U i P takie, że $PA = LU$, gdzie $L \in \mathbb{R}^{3\times 3}$ jest macierzą trójjkątną dolną z jedynkami na głównej przekątnej i wszystkimi elementami co do modułu nie większymi od jedności, $R \in \mathbb{R}^{3\times 3}$ jest trójkątna górna, a $P \in \mathbb{R}^{3\times 3}$ jest macierzą permutacji.

(30) Pokaż, że macierz

$$A = \begin{pmatrix} 4 & 2 & -2 \\ 2 & 2 & 1 \\ -2 & 1 & 21 \end{pmatrix} \in \mathbb{R}^{3\times 3}.$$

ma rozkład Choleskiego $A = LDL^T$. Wyznacz ten rozkład i za jego pomocą rozwiąż układ równań $A\vec{x} = \vec{b}$, gdzie $\vec{b} = [0, 2, 4]^T$.

(31) Niech

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

Znajdź macierz trójkątną dolną L z jedynkami na głównej przekątnej i diagonalną D, takie że $A = LDL^T$. Korzystając z tego rozkładu rozwiąż układ równań $A\vec{x} = \vec{b}$.

(32) Wykonaj rozkład Choleskiego $A = LDL^T$ macierzy

$$A = \begin{bmatrix} 4 & 8 & 12 & 16 \\ 8 & 19 & 30 & 41 \\ 12 & 30 & 50 & 70 \\ 16 & 41 & 70 & 100 \end{bmatrix}.$$

(33) Macierz

$$A = \begin{pmatrix} 2 & 6 & -4 \\ 6 & 17 & -17 \\ -4 & -17 & -20 \end{pmatrix}$$

jest nieosobliwa i symetryczna, ale *nie jest* dodatnio określona. Znajdź, jeśli istnieją, macierz trójkątną dolną L z jedynkami na głównej przekątnej oraz macierz diagonalną D takie, że $A = LDL^T$. Czy taki rozkład istnieje dla dowolnej nieosobliwej i symetrycznej macierzy A?

• Przeprowadzamy eliminację Gaussa bez wyboru elementu główngo wykonując kolejno odpowiednie operacje $A^{(1)} = L_1 A L_1^T$ i $A^{(2)} = L_2 A^{(1)} L_2^T$. Formalnie, wystarczy wyliczać wyrazy na i pod diagonalą kolejnych macierzy $A^{(i)}$, bo pozostałe wyrazy dostajemy z symetrii. Otrzymujemy $A = LDL^T$, gdzie

$$L = \begin{pmatrix} 1 & & \\ 3 & 1 & \\ -2 & 5 & 1 \end{pmatrix}, \qquad D = \begin{pmatrix} 2 & & \\ & -1 & \\ & & -3 \end{pmatrix}.$$

Przy okazji stwierdziliśmy, że macierz A nie jest dodatnio określona, bo nie wszystkie elementy diagonalne macierzy D są dodatnie. Taki rozkład nie zawsze istnieje dla macierzy symetrycznych i nie określonych dodatnio, bo w procesie eliminacyjnym może pojawić się zero na głównej diagonali, np. dla macierzy

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

□

(34) Niech macierz $A = A^T > 0$. Wykaż jednoznaczność rozkładu $A = LDL^T$, gdzie L jest macierzą trójkątną dolną z jedynkami na głównej przekątnej, a D macierzą diagonalną z dodatnimi elementami na głównej przekątnej.

• Oznaczając $U = DL^T$ mamy na podstwie tezy zadania (19), że rozkład $A = LU$ jest jednoznaczny, co oznacza, że macierz diagonalna $D = U(L^T)^{-1}$ też jest wyznaczona jednoznacznie. □

(35) Zaproponuj algorytm rozwiązywania układu równań liniowych $A\vec{x} = \vec{b}$ z macierzą nieosobliwą $A = (a_{i,j}) \in \mathbb{R}^{n\times n}$ taką, że $a_{i,j} = 0$ dla $|i - j| \geqslant 2$, $1 \leqslant i \leqslant n - 1$, $1 \leqslant j \leqslant n$. (Zauważ, że ostatni wiersz jest w ogólności pełny.) Algorytm ma działać w czasie liniowym w n i być numerycznie poprawny.

• Macierz A^T ma elementy niezerowe na trzech centralnych diagonalach oraz w ostatniej kolumnie. Nietrudno zauważyć, ze stosując eliminację Gaussa z wyborem elementu głównego w kolumnie można tę macierz kosztem liniowym w n rozłożyć na odpowiedni iloczyn $PA^T = LU$, gdzie macierz L oprócz jedynek na głównej diagonali ma co najwyżej jeden element niezerowy w każdej kolumnie pod tą diagonalą, a macierz U ma elementy niezerowe $u_{i,j}$ dla $i \leqslant j \leqslant i + 2$ i dla $j = n$. Równoważnie mamy $AP^T = U^T L^T$. Układ $A\vec{x} = \vec{b}$ jest więc równoważny układowi

$$(AP^T)P\vec{x} = \vec{b} \quad \text{albo} \quad U^T L^T P\vec{x} = \vec{b}$$

i można go rozwiązać rozwiązując kolejno kosztem liniowym układy:

$$U^T\vec{y} = \vec{b}, \qquad L^T\vec{z} = \vec{y}, \qquad \vec{x} = P^T\vec{z}.$$

Algorytm jest numerycznie poprawny, bo numerycznie poprawna jest eliminacja Gaussa z wyborem elementu głównego.

(36) Dla danych punktów

$$(x_1, y_1) = (-1, 1), \quad (x_2, y_2) = (-1, 0), \quad (x_3, y_3) = (-2, 2), \quad (x_4, y_4) = (0, 0)$$

poszukujemy wielomianu postaci $w(x) = a_0 + a_1 x$ minimalizującego

$$\sum_{i=1}^{4} (w(x_i) - y_i)^2.$$

Sformułuj odpowiednie zadanie wygładzania liniowego (liniowe zadanie najmniejszych kwadratów) i rozwiąż je dowolnym sposobem.

(37) Dla danych

$$(x_1, y_1) = (-1, 0), \quad (x_2, y_2) = (0, -21), \quad (x_3, y_3) = (1, 26), \quad (x_4, y_4) = (2, -8)$$

wyznacz współczynniki $a, b, c \in \mathbb{R}$ takie, że funkcja $f(t) = a + bt + ct(t-1)(16t-56)$ minimalizuje sumę $\sum_{k=1}^{4} |f(x_k) - y_k|^2$. W tym celu sformułuj odpowiednie zadanie wygładzania liniowego (liniowe zadanie najmniejszych kwadratów) i rozwiąż je metodą równań normalnych.

• Sprawdzamy, że $f(-1) = a - b$, $f(0) = a$, $f(1) = a + b$, $f(2) = a + 2b - 48c$. Zadanie ma więc postać $A\vec{x} \cong \vec{b}$,

$$\begin{pmatrix} 1 & -1 & 0 \\ 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & -48 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} \cong \begin{pmatrix} 0 \\ -21 \\ 26 \\ -8 \end{pmatrix},$$

albo przechodząc do równań normalnych $A^T A\vec{x} = A^T\vec{b}$,

$$\begin{pmatrix} 4 & 2 & -48 \\ 2 & 6 & -96 \\ -48 & -96 & 2304 \end{pmatrix} \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} -3 \\ 10 \\ 384 \end{pmatrix}.$$

Rozwiązując ostatni układ dostajemy rozwiązanie

$$a = \frac{5}{4}; \qquad b = 13; \qquad c = \frac{107}{144}.$$

(38) Niech

$$A = \begin{pmatrix} 1 & -1 \\ 1 & -1 \\ 1 & -2 \\ 1 & 0 \end{pmatrix} \in \mathbb{R}^{4\times 2}.$$

Stosując odbicia Householdera $H_i = I - \vec{u}_i\vec{u}_i^T/\gamma_i$ sprowadź macierz A do postaci trójkątnej górnej $R = H_2H_1A$. Wskaż współczynniki macierzy R oraz odpowiednie wektory $\vec{u}_i$ i liczby γ_i, $i = 1, 2$.

(39) Stosując odbicia Householdera $H_i = I - \vec{u}_i\vec{u}_i^T/\gamma_i$ sprowadź macierz

$$A = \begin{pmatrix} 0 & -2 \\ 0 & 0 \\ -5 & 1 \\ 0 & 2 \end{pmatrix}$$

do postaci trójkątnej górnej $R = H_2H_1A$. Wskaż współczynniki macierzy R oraz odpowiednie wektory $\vec{u}_i$ i liczby γ_i, $i = 1, 2$. Następnie, korzystając z rozkładu, znajdź

$$\min_{\vec{x}\in\mathbb{R}^2} \|\vec{b} - A\vec{x}\|_2$$

dla $\vec{b} = (-4, 1, -3, 4)^T$. Jaki wektor realizuje minimum?

• Oznaczmy $A = (\vec{a}_1, \vec{a}_2)$. Mamy $\|\vec{a}_1\|_2 = 5$ i zgodnie ze wzorami $H_1 = I - \vec{u}_1\vec{u}_1^T/\gamma_1$, gdzie $u_{1,1} = a_1 + \|\vec{a}\|_2$, $u_{i,1} = a_{i,1}$, $2 \leqslant i \leqslant 4$, $\gamma_1 = \|\vec{a}_1\|_2^2 + |a_{1,1}|\,\|\vec{a}\|_2$, czyli

$$\vec{u}_1 = (5, 0, -5, 0)^T, \quad \gamma_1 = 25.$$

Dalej $H_1\vec{a}_1 = (-\|\vec{a}_1\|_2, 0, 0, 0)^T = (-5, 0, 0, 0)^T$,

$$H_1\vec{a}_2 = \vec{a}_2 - \vec{u}_1\frac{\vec{u}_1^T\vec{a}_2}{\gamma_1} = \begin{pmatrix} -2 \\ 0 \\ 1 \\ 2 \end{pmatrix} - \begin{pmatrix} 5 \\ 0 \\ -5 \\ 0 \end{pmatrix}\left(\frac{-3}{5}\right) = \begin{pmatrix} 1 \\ 0 \\ -2 \\ 2 \end{pmatrix},$$

czyli

$$H_1A = \begin{pmatrix} -5 & 1 \\ 0 & 0 \\ 0 & -2 \\ 0 & 2 \end{pmatrix}.$$

Teraz przeprowadzamy wektor $(0, -2, 2)^T$ na kierunek $\vec{e}_1 \in \mathbb{R}^3$. Mamy $u_{2,2} = 2\sqrt{2}$, $u_{3,2} = -2$, $u_{4,2} = 2$, $\gamma_2 = 8$, Uzupełniając $u_{2,1} = 0$ dostajemy $H_2(H_1\vec{a}_2) = (1, -2\sqrt{2}, 0, 0)^T$ i ostatecznie

$$R = H_2H_1A = \begin{pmatrix} -5 & 1 \\ 0 & -2\sqrt{2} \\ 0 & 0 \\ 0 & 0 \end{pmatrix}.$$

Aby rozwiązać zadanie wygładzania liniowego, najpierw liczymy $\vec{c} = H_2H_1\vec{b}$;

$$H_1\vec{b} = \begin{pmatrix} -4 \\ 1 \\ -3 \\ 4 \end{pmatrix} - \begin{pmatrix} 5 \\ 0 \\ -5 \\ 0 \end{pmatrix}\left(\frac{-1}{5}\right) = \begin{pmatrix} -3 \\ 1 \\ -4 \\ 4 \end{pmatrix},$$

$$\vec{c} = \begin{pmatrix} -3 \\ 1 \\ -4 \\ 4 \end{pmatrix} - \begin{pmatrix} 0 \\ 2\sqrt{2} \\ -2 \\ 2 \end{pmatrix}\left(2 + \frac{\sqrt{2}}{4}\right) = \begin{pmatrix} -3 \\ -4\sqrt{2} \\ \sqrt{2}/2 \\ -\sqrt{2}/2 \end{pmatrix},$$

a następnie rozwiązujemy układ równań

$$\begin{pmatrix} -5 & 1 \\ 0 & -2\sqrt{2} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} -3 \\ -4\sqrt{2} \end{pmatrix}.$$

Stąd dostajemy $x_1 = 1$, $x_2 = 2$, oraz residuum $\|\vec{c} - R\vec{x}\|_2 = \sqrt{c_3^2 + c_4^2} = 1$.
□

(40) Niech

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

Znajdź wektory $\vec{v}_1, \vec{v}_2, \vec{v}_3 \in \mathrm{R}^4$ wyznaczające odbicia symetryczne, reprezentowane przez macierze ortogonalne $H_i = I - \vec{v}_i\vec{v}_i^T/\gamma_i$, $\gamma_i = \|\vec{v}_i\|_2/2$, takie że macierz $R = H_3H_2H_1A$ jest trójkątna górna. Korzystając z tgo rozkładu, rozwiąż układ równań liniowych $A\vec{x} = \vec{b}$.

(41) Niech $\vec{u}$ będzie niezerowym wektorem w $\mathbb{R}^m$ oraz $\gamma = \|\vec{u}\|_2^2/2$. Uzasadnij, że algorytm obliczania

$$H\vec{x} \;=\; \vec{x} \;-\; s\,\vec{u}, \qquad s \;=\; (\vec{u}^T\vec{x})/\gamma,$$

według powyższego wzoru, jest numerycznie poprawny ze względu na dany wektor $\vec{x} \in \mathbb{R}^m$. Dokładniej, obliczony wynik jest nieco zaburzonym dokładnym wynikiem.

• Mamy

$$fl_\nu(s) = \frac{1}{\gamma}\sum_{j=1}^{n} u_j x_j (1+\varepsilon_j), \qquad |\varepsilon_j| \lessapprox (n+2)\nu,$$

$$\begin{aligned} fl_\nu((H\vec{x})_i) &= x_i(1+\alpha_i) - fl_\nu(s)u_i(1+\beta_i), && |\alpha_i|, |\beta_i| \lessapprox 2\nu, \\ &= x_i(1+\alpha_i) - \frac{1}{\gamma}u_i\sum_{j=1}^{n} u_j x_j (1+\delta_{i,j}), && |\delta_{i,j}| \lessapprox (n+4)\nu. \end{aligned}$$

Stąd

$$\begin{aligned} \|fl_\nu(H\vec{x}) - H\vec{x}\|_2^2 &= \sum_{i=1}^{n}\Big|\alpha_i x_i - \frac{1}{\gamma}u_i\sum_{j=1}^{n}\delta_{i,j}u_j x_j\Big|^2 \\ &\leqslant 2\sum_{i=1}^{n}\left(\alpha_i^2 x_i^2 + \frac{1}{\gamma^2}u_i^2\Big(\sum_{j=1}^{n}|\delta_{i,j}u_j x_j|\Big)^2\right) \\ &\lessapprox 8\nu^2\|\vec{x}\|_2^2 + \frac{2}{\gamma^2}\|\vec{u}\|_2^2(n+4)^2\nu^2\Big(\|\vec{u}\|_2^2\|\vec{x}\|_2^2\Big) \\ &\leqslant 2(n+4)^2\nu^2\|\vec{x}\|_2^2. \end{aligned}$$

Teraz wystarczy wykorzystać fakt, że $\|H\vec{x}\|_2 = \|\vec{x}\|_2$ aby ostatecznie otrzymać

$$\|fl_\nu(H\vec{x}) - H\vec{x}\|_2 \lessapprox (n+4)\sqrt{2}\,\nu\,\|H\vec{x}\|_2.$$

□

(42) Niech $A \in \mathbb{R}^{m\times n}$, gdzie $m \geqslant n = \mathrm{rank}(A)$. Niech $\vec{x}^*$ i $\vec{y}^*$ będą rozwiązaniami zadań wygładzania, odpowiednio, $A\vec{x} \cong \vec{b}$ i $A\vec{y} \cong \vec{c}$, gdzie $\vec{c}$ jest zaburzonym vektorem $\vec{b}$ spełniającym

$$\|\vec{c} - \vec{b}\,\|_2 \leqslant \eta \|\vec{b}\,\|_2.$$

Wykaż, że jeśli $\vec{b} \in \mathrm{Im}(A)$ to

$$\|\vec{y}^* - \vec{x}^*\|_2 \leqslant \eta\, \kappa(A) \|\vec{x}^*\|_2, \qquad \text{gdzie} \quad \kappa(A) = \|A\|_2 \|A^+\|_2,$$

tzn. za uwarunkowanie zadania ze względu na zaburzenia wektora $\vec{b}$ można przyjąć $\kappa(A)$. Czy uwarunkowanie się zmieni gdy zrezygnujemy z założenia, że $\vec{b} \in \mathrm{Im}(A)$? Uzasadnij odpowiedź.

• Mamy $\vec{x}^* = A^+\vec{b}$ i $\vec{y}^* = A^+\vec{c}$. Ponieważ $\vec{b} \in \mathrm{Im}(A)$ to również $A\vec{x}^* = \vec{b}$. Stąd

$$\begin{aligned} \|\vec{y}^* - \vec{x}^*\|_2 &= \|A^+(\vec{c} - \vec{b}\,)\|_2 \leqslant \eta\, \|\vec{b}\,\|_2 \|A^+\|_2 \\ &= \eta\, \|A\vec{x}^*\|_2 \|A^+\|_2 \leqslant \eta\, \|A\|_2 \|A^+\|_2 \|\vec{x}^*\|_2, \end{aligned}$$

co dowodzi zasadniczą część zadania.

Jeśli $\vec{b} \notin \mathrm{Im}(A)$ to uwarunkowanie rośnie wraz ze wzrostem wektora residualnego $\vec{r} = \vec{b} - A\vec{x}^*$. Mamy bowiem $\|\vec{b}\,\|_2^2 = \|A\vec{x}^*\|_2^2 + \|\vec{r}\,\|_2^2$, a stąd (dla $\vec{x}^* \neq \vec{0}$)

$$\|\vec{y}^* - \vec{x}^*\|_2 \leqslant \eta\, \|\vec{b}\,\|_2 \|A^+\|_2 = \eta\, \|A^+\|_2 \|A\|_2 \left(1 + \left(\frac{\|\vec{r}\,\|_2}{\|A\|_2 \|\vec{x}^*\|_2}\right)^2\right)^{1/2} \|\vec{x}^*\|_2.$$

Najlepiej to widać w przypadku gdy $\vec{b}$ jest postopadły do $\mathrm{Im}(A)$, a $\vec{c}$ nie. Wtedy $\vec{x}^* = \vec{0}$, ale $\vec{y}^*$ już nie - uwarunkowanie dla takich danych jest formalnie nieskończone.
□

(43) Rozpatrzmy dwa rozkłady ortogonalno-trójkątne macierzy $A \in \mathrm{R}^{m\times n}$, $\mathrm{rank}(A) = n$. Pierwszy to $A = QR$, gdzie $Q \in \mathrm{R}^{m\times m}$ i $R \in \mathrm{R}^{m\times n}$ (np. pochodzący z algorytmu Householdera), a drugi to $A = Q'R'$, gdzie $Q' \in \mathrm{R}^{m\times n}$ i $R' \in \mathrm{R}^{n\times n}$ (np. pochodzący z algorytmu ortogonalizacji Grama-Schmidta. Pokaż, że n pierwszych kolumn macierzy Q różnią się od n pierwszych kolumn macierzy Q' co najwyżej znakami; tzn. istnieje macierz diagonalna $D \in \mathbb{R}^{m\times n}$ taka, że $d_{i,i} = \pm 1$ dla $1 \leqslant i \leqslant n$ oraz $Q' = D\,Q$.

• Niech $A = (\vec{a}_1, \ldots, \vec{a}_n)$, $Q = (\vec{q}_1, \ldots, \vec{q}_n, \ldots, \vec{q}_m)$, $Q_1 = (\vec{q}\,'_1, \ldots, \vec{q}\,'_n)$. Postępujemy indukcyjnie ze względu na n. Dla $n = 1$ mamy $\vec{q}_1 r_{1,1} = \vec{q}\,'_1 r'_{1,1}$, a ponieważ $\|\vec{q}_1\|_2 = 1 = \|\vec{q}\,'_1\|_2$ to $|r'_{1,1}| = |r_{1,1}|$ i $\vec{q}\,'_1 = \pm\vec{q}_1$. Dla $n \geqslant 2$ mamy

$$\sum_{i=1}^{n-1} \vec{q}_i r_{i,n} + \vec{q}_n r_{n,n} \;=\; \vec{a}_n \;=\; \sum_{i=1}^{n-1} \vec{q}\,'_i r'_{i,n} + \vec{q}\,'_n r'_{n,n}.$$

To jest rozkład wektora $\vec{a}_n$ na składowe w $V_{n-1} = \mathrm{span}(\vec{a}_1, \ldots, \vec{a}_{n-1})$ i w jednoznacznie wyznaczonej jednowymiarowej przestrzeni W_n do niej prostopadłej, takiej, że $V_{n-1} \oplus W_n = \mathrm{span}(\vec{a}_1, \ldots, \vec{a}_n)$. Z jednoznaczności takiego rozkładu dostajemy, że $\vec{q}\,'_n r'_{n,n} = \vec{q}_n r_{n,n}$, a stąd $\vec{q}\,'_n = \pm\vec{q}_n$.
□

(44) Przedyskutuj jednoznaczność rozkładu SVD (*Singular Value Decomposition*) danej macierzy $A \in \mathrm{R}^{m\times n}$,

$$A \,=\, U\,\Sigma\,V^T,$$

gdzie $U \in \mathrm{R}^{m\times m}$ i $V \in \mathrm{R}^{n\times n}$ są macierzami ortogonalnymi, a $\Sigma \in \mathrm{R}^{m\times n}$ macierzą diagonalną, $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2 \ldots, \sigma_k, 0, \ldots, 0)$, gdzie $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_k > 0$ i $k \leqslant \min(m, n)$.

• Jeśli $A = U\Sigma V^T$ to

$$(A^TA)V = (U\Sigma V^T)^T(U\Sigma V^T)V = V\Sigma^TU^TU\Sigma V^TV = V(\Sigma^T\Sigma),$$

a to oznacza, że wyrazy diagonalne macierzy $\Sigma^T\Sigma = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0) \in \mathbb{R}^{n\times n}$ są wartościami własnymi macierzy A^TA (uwzględniając ich krotności), a kolumny macierzy ortogonalnej V są odpowiadającymi im wektorami własnymi. Podobnie, wyrazy diagonalne macierzy $\Sigma\Sigma^T = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_k^2, 0, \ldots, 0) \in \mathbb{R}^{m\times m}$ są wartościami własnymi macierzy AA^T, a kolumny U są odpowiadającymi im wektorami własnymi. (Macierze A^TA i AA^T mają te same niezerowe wartości własne.) Ponieważ wartości własne są wyznaczone jednoznacznie, macierz Σ jest też wyznaczona jednoznacznie. Natomiast macierze U i V można dobrać na tyle sposobów, na ile sposobów można dobrać odpowiednie ortonormalne bazy wektorów własnych.

W szczególności, gdy macierz A jest kwadratowa, nieosobliwa i wszystkie wartości własne macierzy A^TA są jednokrotne to macierz V jest wyznaczona jednoznacznie z dokładnością do znaków jej kolumn i $U = AV\Sigma^{-1}$. Z dugiej strony, gdy A jest macierzą zerową to U i V są dowolnymi macierzami ortogonalnymi.
□

(45) Zastosujmy ortogonalizację Grama-Schmidta do dwóch wektorów liniowo niezależnych $\vec{a}$ i $\vec{b}$ o normach $\|\vec{a}\|_2 = 1 = \|\vec{b}\|_2$. Załóżmy dla uproszczenia, że w obliczeniach *jedynie* iloczyn skalarny tych wektorów $s = \langle\vec{a}, \vec{b}\rangle_2$ liczy się z błędem, $fl_\nu(s) \,=\, s(1+\varepsilon)$, gdzie $|\varepsilon|$ jest dodatnie i na poziomie ν. Pokaż, że wtedy dla otrzymanej w wyniku ortogonalizacji unormowanej pary wektorów $\vec{a}$ i $\vec{c}$ mamy

$$\langle\vec{a}, \vec{c}\rangle_2 \,=\, \frac{-\varepsilon s}{\sqrt{1 - s^2(1-\varepsilon^2)}}.$$

Wywnioskuj stąd, że gdy $\vec{a}$ i $\vec{b}$ są "prawie liniowo zależne", to $\vec{a}$ i $\vec{c}$ są dalekie od ortogonalnych.

• Wobec powyższych założeń, obliczony wektor $\vec{c}$ wynosi

$$\vec{c} = \frac{\vec{b} - s(1+\varepsilon)\vec{a}}{\|\vec{b} - s(1+\varepsilon)\vec{a}\|_2}.$$

Mamy

$$\begin{aligned}\|\vec{b} - s(1+\varepsilon)\vec{a}\|_2^2 \;&=\; \|\vec{b}\|_2^2 + s^2(1+\varepsilon)^2\|\vec{a}\|_2^2 - 2\langle\vec{b}, \vec{a}\rangle_2 s(1+\varepsilon)\\ &=\; 1 + s^2(1+\varepsilon)^2 - 2s^2(1+\varepsilon) = 1 - s^2(1-\varepsilon^2),\end{aligned}$$

a stąd

$$\langle\vec{a}, \vec{c}\rangle_2 \,=\, \frac{\left\langle\vec{a}, \vec{b} - s(1+\varepsilon)\vec{a}\right\rangle_2}{\sqrt{1 - s^2(1-\varepsilon^2)}} = \frac{\langle\vec{a}, \vec{b}\rangle_2 - s(1+\varepsilon)\|\vec{a}\|_2^2}{\sqrt{1 - s^2(1-\varepsilon^2)}} = \frac{-\varepsilon s}{\sqrt{1 - s^2(1-\varepsilon^2)}}.$$

Jeśli $\vec{a}$ i $\vec{b}$ są "prawie liniowo zależne" to znaczy, że $|s|$ jest bliskie 1. Ale wobec tego, że $\lim_{|s|\to 1^-} |\langle \vec{a}, \vec{c}\rangle_2| = 1$, wektory $\vec{a}$ i $\vec{c}$ są też "prawie liniowo zależne."
□

(46) Pokaż numeryczną poprawność algorytmu Hornera obliczania wartości wielomianu w punkcie ξ ze względu na dane współczynniki a_j tego wielomianu.

• W algorytmie Hornera mamy $v_n = a_n$, $v_{n-1} = v_n\xi + a_{n-1} = a_n\xi + a_{n-1}$ i ogólniej

$$v_k = a_n\xi^{n-k} + a_{n-1}\xi^{n-k-1} + \cdots + a_{k+1}\xi + a_k \quad \text{dla} \quad k = n-1, n-2, \ldots, 1, 0.$$

Pokażemy indukcyjnie, że obliczone v_k jest dokladnym v_k dla nieco zaburzonych współczynników $a_n, \ldots, a_{n-k}$. Rzeczywiście, $fl_\nu(v_n) = a_n(1+\varepsilon_{n,n})$, gdzie błąd wynika jedynie z reprezentacji, $|\varepsilon_{n,n}| \leqslant \nu$. Wykorzystując założenie indukcyjne, dla $k \leqslant n-1$ mamy

$$\begin{aligned} fl_\nu(v_k) &= fl_\nu(v_{k+1})\xi(1+\alpha_k) + a_k(1+\beta_k) \\ &= \Big(\sum_{i=k+1}^{n} a_i(1+\varepsilon_{k+1,i})\xi^{i-k-1}\Big)\xi(1+\alpha_k) + a_k(1+\beta_k) \\ &= \sum_{i=k}^{n} a_i(1+\varepsilon_{k,i})\xi^{i-k}, \end{aligned}$$

gdzie $|\alpha_k|, |\beta_k| \lessapprox 2\nu$, $(1+\varepsilon_{k,i}) = (1+\beta_k)$ oraz $(1+\varepsilon_{k,i}) = (1+\varepsilon_{k+1,i})(1+\alpha_k)$ dla $k+1 \leqslant i \leqslant n$. Biorąc $k = 0$ dostajemy, że obliczona wartość wielomianu o współczynnikach a_i w punkcie ξ jest dokładną wartością w ξ wielomianu o współczynnikach $\tilde{a}_i = a_i(1+\varepsilon_{0,i})$, gdzie $|\varepsilon_{0,i}| \lessapprox 2(i+1)\nu$.
□

(47) Pokaż, że algorytm Hornera obliczania wartości $w(\xi)$ wielomianu danego w postaci potęgowej jest jednocześnie algorytmem dzielenia tego wielomianu przez jednomian $(x-\xi)$. Dokładniej, jeśli $w(x) = \sum_{j=0}^{n} a_j x^j$ to

$$w(x) \;=\; \Big(\sum_{j=1}^{n} v_j x^{j-1}\Big)\,(x-\xi) \;+\; v_0,$$

gdzie v_j są zdefiniowane tak jak w algorytmie Hornera. Wykorzystaj ten fakt do jednoczesnego obliczania nie tylko wartości $w(\xi)$, ale też pochodnej $w'(\xi)$.

• Niech $w(x) = a_0 + a_1x + \cdots + a_nx^n$. Porównując współczynniki przy x^k dla $k = n, n-1, \ldots, 1, 0$ po obu stronach równania

$$\sum_{i=0}^{n} a_i x^i = \Big(\sum_{i=1}^{n} c_i x^{i-1}\Big)(x-\xi) + c_0$$

dostajemy wzór rekurencyjny $a_n = c_n$ i $a_i = c_i - \xi c_{i+1}$ dla $i = n-1, \ldots, 0$. A ponieważ w algorytmie Hornera mamy $v_n = a_n$ i $v_i = \xi v_{i+1} + a_i$ to $v_i = c_i$.

Jeśli teraz $w(x) = v(x)(x-\xi)+v_0$ to $w'(x) = v(x)+(x-\xi)v'(x)$, czyli $w'(\xi) = v(\xi)$. Pochodną można więc policzyć jako wartość ilorazu $v(x) = \sum_{i=1}^{n} v_j x^{j-1}$ dla $x = \xi$. Realizuje to następujący algorytm:

```
w0 := w1 := a_n;
for j := n − 1 downto 1 do
begin
    w0 := w0 ∗ ξ + a_j;
    w1 := w1 ∗ ξ + w0
end;
w0 := w0 + a_0.
```

Po jego wykonaniu mamy $w0 = w(\xi)$ i $w1 = w'(\xi)$.
□

(48) Zaproponuj algorytm o koszcie proporcjonalnym do n^2 obliczający współczynniki wielomianu $p \in \Pi_n$ w bazie potęgowej dysponując współczynnikami tego wielomianu w bazie Newtona.

• Niech b_i będą odpowiednimi współczynnikami w rozwinięciu wielomianu p w bazie Newtona. Dla $i = n, n-1, \ldots, 0$, niech $a_j^{(i)}$, $i \leqslant j \leqslant n$, będą współczynnikami w rozwinięciu

$$b_i + b_{i+1}(x - x_i) + \ldots + b_n(x - x_i)\cdots(x - x_{n-1}) = a_i^{(i)} + a_{i+1}^{(i)}x + \ldots + a_n^{(i)}x^{n-i}.$$

Dla $i = n$ mamy $a_n^{(n)} = b_n$. Dla $i \leqslant n-1$ rozpisujemy lewą stronę powyższego równania i dostajemy

$$\begin{aligned} & b_i + (x - x_i)\Big(b_{i+1} + b_{i+2}(x - x_{i+1}) + \ldots + b_n(x - x_{i+1})\cdots(x - x_{n-1})\Big) \\ = \;& b_i + (x - x_i)\Big(a_{i+1}^{(i+1)} + a_{i+2}^{(i+1)}x + \ldots + a_n^{(i+1)}x^{n-i-1}\Big) \\ = \;& \Big(b_i - x_i a_{i+1}^{(i+1)}\Big) + \Big(a_{i+1}^{(i+1)} - x_i a_{i+2}^{(i+1)}\Big)x + \ldots + \Big(a_{n-1}^{(i+1)} - x_i a_n^{(i+1)}\Big)x^{n-i-1} + a_n^{(i+1)}x^{n-i}, \end{aligned}$$

a stąd zależność rekurencyjną dla $i = n-1, n-2, \ldots, 1, 0$

$$\begin{cases} a_n^{(i)} = a_n^{(i+1)}, \\ a_k^{(i)} = a_k^{(i+1)} - x_i a_{k+1}^{(i+1)}, & k = n-1, n-2, \ldots, i+1, \\ a_i^{(i)} = b_i - x_i a_{i+1}^{(i+1)}. \end{cases}$$

Uwzględniając fakt, że poszukiwanymi współczynnikami są $a_j = a_j^{(0)}$, algorytm może być zrealizowany następująco:

```
for i := 0 to n do a[i] := b[i];
for i := n − 1 downto 0 do
for k := i to n − 1 do
    a[k] := a[k] − x[i] ∗ a[k + 1].
```

□

(49) Zaproponuj algorytm działający w czasie proporcjonalnym do n^2 rozwiązujący układ równań z macierzą Vandermonde'a $V = \{x_i^{j-1}\}_{i,j=1}^n$, gdzie punkty $x_i \in \mathbb{R}$ są parami różne.

• Rozwiązaniem układu $V\vec{x} = \vec{f}$ z macierzą Vandermonde'a $V = \{x_i^{j-1}\}_{i,j=1}^n$ są współczynniki w bazie potęgowej wielomianu interpolującego dane $\vec{f} = (f_1, \ldots, f_n)^T$ w punktach $x_1, \ldots, x_n$. Możemy więc najpierw znaleźć współczynniki tego wielomianu w bazie Newtona kosztem rzędu n^2, np. przy pomocy algorytmu różnic dzielonych, a

następnie skorzystać z zadania (48) aby również kosztem n^2 przejść z bazy Newtona do bazy potęgowej.
□

(50) Pokaż, że dla parami różnych węzłów $t_0, \ldots, t_n$, gdzie $n \geqslant 1$, mamy

$$f[t_0, t_1, \ldots, t_n] = \sum_{i=0}^{n} \frac{f(t_i)}{(t_i - t_0)\cdots(t_i - t_{i-1})(t_i - t_{i+1})\cdots(t_i - t_n)}.$$

• Zastosujemy indukcję ze względu na n. Dla $n = 1$ mamy

$$f[t_0, t_1] = \frac{f(t_1) - f(t_0)}{t_1 - t_0} = \frac{f(t_0)}{t_0 - t_1} + \frac{f(t_1)}{t_1 - t_0}.$$

Niech $n \geqslant 2$. Z założenia indukcyjnego mamy

$$f[t_1, \ldots, t_n] = \frac{f(t_n)}{\prod_{j=0}^{n-1}(t_n - t_j)} + \sum_{i=1}^{n-1} \frac{f(t_i)}{\prod_{i\neq j=1}^{n}(t_i - t_j)},$$

$$f[t_0, \ldots, t_{n-1}] = \frac{f(t_0)}{\prod_{j=1}^{n-1}(t_0 - t_j)} + \sum_{i=1}^{n-1} \frac{f(t_i)}{\prod_{i\neq j=0}^{n-1}(t_i - t_j)}.$$

Wstawiając powyższe wyrażenia do prawej strony równania

$$f[t_0, \ldots, t_n] = \frac{f[t_1, \ldots, t_n] - f[t_0, \ldots, t_{n-1}]}{t_n - t_0}$$

dostajemy, że:
- współczynnik przy $f(t_n)$ wynosi

$$\frac{1}{(t_n - t_0)\prod_{j=0}^{n-1}(t_n - t_j)} = \frac{1}{\prod_{j=0}^{n-1}(t_n - t_j)},$$

- współczynnik przy $f(t_0)$ wynosi

$$\frac{-1}{(t_n - t_0)\prod_{j=1}^{n-1}(t_0 - t_j)} = \frac{1}{\prod_{j=1}^{n}(t_0 - t_j)},$$

- współczynnik przy $f(t_i)$ dla $1 \leqslant i \leqslant n-1$ wynosi

$$\begin{aligned}
&\frac{1}{t_n - t_0}\left(\frac{1}{\prod_{i\neq j=1}^{n}(t_i - t_j)} - \frac{1}{\prod_{i\neq j=0}^{n-1}(t_i - t_j)}\right) \\
&= \frac{1}{t_n - t_0}\left(\frac{1}{t_i - t_n} - \frac{1}{t_i - t_0}\right)\left(\frac{1}{\prod_{i\neq j=1}^{n-1}(t_i - t_j)}\right) \\
&= \frac{1}{(t_i - t_0)(t_n - t_i)}\left(\frac{1}{\prod_{i\neq j=1}^{n-1}(t_i - t_j)}\right) = \frac{1}{\prod_{i\neq j=0}^{n}(t_i - t_j)},
\end{aligned}$$

co dowodzi tezy zadania.
□

(51) Niech $t_0 < t_1 < \ldots < t_n$. Pokaż, że realizując w fl_ν algorytm różnic dzielonych dla danych $f(t_j)$, zamiast dokładnej wartości $f[t_0, t_1, \ldots, t_n]$ otrzymujemy $fl_\nu\Big(f[t_0, t_1, \ldots, t_n]\Big)$, która jest dokładną różnicą dzieloną dla danych $f(t_j)(1+\varepsilon_j)$, gdzie $|\varepsilon_j| \lessapprox (3n+1)\nu$ dla $0 \leqslant j \leqslant n$.

• Użyjemy indukcji ze względu na n. Dla $n = 0$ teza jest oczywista. Niech $n \geqslant 1$. Mamy

$$fl_\nu\Big(f[t_0, \ldots, t_n]\Big) = \frac{fl_\nu\Big(f[t_1, \ldots, t_n]\Big) - fl_\nu\Big(f[t_0, \ldots, t_{n-1}]\Big)}{t_n - t_0}(1+\gamma_n),$$

gdzie $|\gamma_n| \lessapprox 3\nu$. Z założenia indukcyjnego i z tezy zadania (50) dostajemy, że

$$fl_\nu\Big(f[t_1, \ldots, t_n]\Big) = \frac{f(t_n)(1+\alpha_n)}{\prod_{j=0}^{n-1}(t_n - t_j)} + \sum_{i=1}^{n-1} \frac{f(t_i)(1+\alpha_{i,1})}{\prod_{i\neq j=1}^{n}(t_i - t_j)},$$

$$fl_\nu\Big(f[t_0, \ldots, t_{n-1}]\Big) = \frac{f(t_0)(1+\alpha_0)}{\prod_{j=1}^{n-1}(t_0 - t_j)} + \sum_{i=1}^{n-1} \frac{f(t_i)(1+\alpha_{i,2})}{\prod_{i\neq j=0}^{n-1}(t_i - t_j)},$$

gdzie $|\alpha_n|, |\alpha_0|, |\alpha_{i,1}|, |\alpha_{i,2}| \lessapprox (3(n-1)+1)\nu = (3n-2)\nu$. Wstawiając powyższe wyrażenia do prawej strony równania rekurencyjnego dostajemy, że:
- współczynnik przy $f(t_n)$ wynosi

$$\frac{(1+\alpha_n)(1+\gamma_n)}{(t_n - t_0)\prod_{j=0}^{n-1}(t_n - t_j)} = \frac{(1+\alpha_n)(1+\gamma_n)}{\prod_{j=0}^{n-1}(t_n - t_j)},$$

- współczynnik przy $f(t_0)$ wynosi

$$\frac{-(1+\alpha_0)(1+\gamma_n)}{(t_n - t_0)\prod_{j=1}^{n-1}(t_0 - t_j)} = \frac{(1+\alpha_0)(1+\gamma_n)}{\prod_{j=1}^{n}(t_0 - t_j)},$$

- współczynnik przy $f(t_i)$ dla $1 \leqslant i \leqslant n-1$ wynosi

$$\begin{aligned} &\frac{1+\gamma_n}{t_n - t_0}\left(\frac{1+\alpha_{i,1}}{\prod_{i\neq j=1}^{n}(t_i - t_j)} - \frac{1+\alpha_{i,2}}{\prod_{i\neq j=0}^{n-1}(t_i - t_j)}\right) \\ &= \frac{1+\gamma_n}{t_n - t_0}\left(\frac{1+\alpha_{i,1}}{t_i - t_n} - \frac{1+\alpha_{i,2}}{t_i - t_0}\right)\left(\frac{1}{\prod_{i\neq j=1}^{n-1}(t_i - t_j)}\right). \end{aligned}$$

Jeśli węzły są uporządkowane to środkowy czynnik powstaje przez odjęcie liczb o różnych znakach, a stąd

$$\frac{1+\alpha_{i,1}}{t_i - t_n} - \frac{1+\alpha_{i,2}}{t_i - t_0} = \left(\frac{1}{t_i - t_n} - \frac{1}{t_i - t_0}\right)(1+\alpha_i), \qquad |\alpha_i| \lessapprox (3n-2)\nu.$$

Współczynnik ten wynosi więc

$$\frac{(1+\gamma_n)(1+\alpha_i)}{(t_i - t_0)(t_n - t_i)}\left(\frac{1}{\prod_{i\neq j=1}^{n-1}(t_i - t_j)}\right) = \frac{(1+\gamma_n)(1+\alpha_i)}{\prod_{i\neq j=0}^{n}(t_i - t_j)},$$

Ostatecznie, przyjmując $(1+\varepsilon_i) = (1+\gamma_n)(1+\alpha_i)$ mamy

$$fl_\nu\Big(f[t_0, \ldots, t_n]\Big) = \sum_{i=0}^{n} \frac{f(t_i)(1+\varepsilon_i)}{\prod_{i\neq j=0}^{n}(t_i - t_j)}, \qquad \text{gdzie} \quad |\varepsilon| \lessapprox (3n+1)\nu,$$

co wobec tezy zadania (50) kończy dowód.
□

(52) Uzasadnij, że iloraz różnicowy jest funkcją symetryczną, tzn. dla dowolnej permutacji $t_{i_0}, \ldots, t_{i_s}$ węzłów $t_0, \ldots, t_s$ mamy

$$f[t_{i_0}, \ldots, t_{i_s}] \;=\; f[t_0, \ldots, t_s].$$

• Teza wynika prosto z faktu, że $f[t_0, \ldots, t_s]$ jest współczynnikiem przy najwyższej potędze wielomianu interpolującego f w punktach $t_0, \ldots, t_s$ (uwzględniając krotności) oraz, że ten wielomian nie zależy od kolejności węzłów. (W przypadku węzłów jednokrotnych teza wynika też z tezy zadania (50))
□

(53) Niech w będzie wielomianem stopnia dokładnie n. Pokaż, że dla ustalonych $t_1, t_2, \ldots, t_s$ funkcja

$$w_{t_1,\ldots,t_s}(t) \;=\; w[t_1, t_2, \ldots, t_s, t], \quad t \in \mathbb{R},$$

jest wielomianem stopnia dokładnie $n - s$ dla $s \leqslant n$, oraz jest zerem dla $s \geqslant n + 1$.

• Załóżmy najpierw, że $s \leqslant n$. Niech $w_t \in \Pi_s$ będzie wielomianem interpolującym w w punktach $t_1, t_2, \ldots, t_s$ oraz t (uwzględniając krotności). Wtedy

$$\begin{aligned} w(t) \;=\; w_t(t) \;&=\; \sum_{k=1}^{s} w[t_1, t_2, \ldots, t_k](t - t_1)(t - t_2)\cdots(t - t_{k-1}) \\ &\quad + w[t_1, \ldots, t_s, t](t - t_1)(t - t_2)\cdots(t - t_s), \end{aligned}$$

a stąd

$$\begin{aligned} & w[t_1, \ldots, t_s, t](t - t_1)(t - t_2)\cdots(t - t_s) \\ =\; & w(t) - \sum_{k=1}^{s} w[t_1, t_2, \ldots, t_k](t - t_1)(t - t_2)\cdots(t - t_{k-1}). \end{aligned}$$

Wielomian po prawej stronie zeruje się w punktach $t_1, \ldots, t_s$, jest więc podzielny przez $(t - t_1)\cdots(t - t_s)$. Iloraz wynosi $w[t_1, t_2, \ldots, t_s, t]$ i jest wielomianem stopnia dokładnie $n - s$. To w szczególności oznacza, że $w[t, t_1, \ldots, t_s]$ jest stałą dla $s = n$ i zerem dla $s \geqslant n + 1$.
□

(54) Pokaż, że jeśli funkcja $f \in C^n([a, b])$ to dla każdego t_n

$$\lim_{t \to t_n} f[t_0, t_1, \ldots, t_{n-1}, t] \;=\; f[t_0, t_1, \ldots, t_{n-1}, t_n].$$

• Zastosujemy indukcję ze względu na n. Dla $n = 0$ mamy $\lim_{t\to t_0} f[t] = f(t_0)$. Niech $n \geqslant 1$. Jeśli $t_0 = \cdots = t_n$ to

$$\lim_{t \to t_0} f[t_0, \ldots, t_{n-1}, t] = \lim_{t \to t_0} \frac{f^{(n)}(\xi(t))}{n!} = \frac{f^{(n)}(x_0)}{n!} = f[t_0, \ldots, t_{n-1}, t_n],$$

gdzie środkowa równość wynika z faktu, że $\xi(t)$ należy do przedziału o końcach t_0 i t, oraz z ciągłości $f^{(n)}$. Załóżmy więc, bez zmniejszenia ogólności, że $t_0 \neq t_n$. Wtedy

$$f[t_0, t_1, \ldots, t_{n-1}, t] = \frac{f[t_1, t_2, \ldots, t_{n-1}, t] - f[t_0, t_1, \ldots, t_{n-1}]}{t - t_0}.$$

Biorąc granicę prawej strony tej równości przy $t \to t_n$ i korzystając z założenia indukcyjnego dostajemy, że ta granica wynosi

$$\frac{f[t_1, t_2, \ldots, t_{n-1}, t_n] - f[t_0, t_1, \ldots, t_{n-1}]}{t_n - t_0} = f[t_0, t_1, \ldots, t_{n-1}, t_n].$$

□

(55) Pokaż, że jeśli funkcja $f \in C^n([a,b])$ to

$$\frac{\partial}{\partial x_n} f[x_0, x_1, \ldots, x_{n-1}, x_n] \;=\; Pf[x_0, x_1, \ldots, x_{n-1}, x_n, x_n].$$

• Korzystając z tezy zadania (54) mamy

$$\begin{aligned} &\lim_{x \to x_n} \frac{f[x_0, x_1, \ldots, x_{n-1}, x] - f[x_0, x_1, \ldots, x_{n-1}, x_n]}{x - x_n} \\ =\; &\lim_{x \to x_n} f[x_0, x_1, \ldots, x_{n-1}, x_n, x] \;=\; f[x_0, x_1, \ldots, x_{n-1}, x_n, x_n]. \end{aligned}$$

□

(56) Wykaż, że dla funkcji $f(x) = 1/x$ mamy $f[x_0, x_1, \ldots, x_n] = (-1)^n \prod_{i=0}^n x_i^{-1}$.

• Przeprowadzimy dowód indukcyjny po n. Dla $n = 0$ oczywiście $f[x_0] = x_0^{-1}$. Dla $n \geqslant 1$ skorzystamy ze wzoru rekurencyjnego i założenia indukcyjnego. Mamy

$$\begin{aligned} f[x_0, \ldots, x_n] &= \frac{f[x_1, \ldots, x_n] - f[x_0, \ldots, x_{n-1}]}{x_n - x_0} \\ &= \frac{(-1)^{n-1} \prod_{i=1}^n x_i^{-1} - (-1)^{n-1} \prod_{i=0}^{n-1} x_i^{-1}}{x_n - x_0} \\ &= (-1)^{n-1} \Big(\prod_{i=1}^{n-1} x_i^{-1} \Big) \Big(\frac{x_n^{-1} - x_0^{-1}}{x_n - x_0} \Big) \\ &= (-1)^n \prod_{i=0}^{n} x_i^{-1}. \end{aligned}$$

□

(57) Wykaż, że dla funkcji $f(x) = x^n$ i dowolnych punktów x_j, $0 \leqslant j \leqslant k$, różnica dzielona

$$f[x_0, x_1, \ldots, x_k] = \begin{cases} 0 & \text{jeśli} \quad k \geqslant n+1, \\ 1 & \text{jeśli} \quad k = n, \\ x_0 + x_1 + \cdots + x_k & \text{jeśli} \quad k = n-1. \end{cases}$$

• Dla $k \geqslant n+1$ teza wynika prosto z faktu, że $f[x_0, \ldots, x_k] = f^{(k)}(\xi)/k!$ dla pewnego ξ. W pozostałych przypadkach rozwijamy f w bazie Newtona,

$$x^n = \sum_{i=0}^{n} f[x_0, \ldots, x_i](x - x_0) \cdots (x - x_{i-1}).$$

Porównując współczynniki przy x^n po obu stronach tego równania dostajemy

$$f[x_0, \ldots, x_{n-1}, x_n] = 1,$$

a porównując współczynniki przy x^{n-1} dostajemy

$$0 = f[x_0, \ldots, x_{n-1}] - f[x_0, \ldots, x_{n-1}, x_n](x_0 + x_1 + \ldots + x_n),$$

czyli tezę zadania.
□

(58) Wyznacz wielomian w jak najniższego stopnia taki, że

$$w(x_i) = y_i, \quad w'(x_i) = 0, \qquad 0 \leqslant i \leqslant n,$$

gdzie punkty x_i są parami różne. Odpowiedź podaj 'w języku' odpowiednich wielomianów Lagrange'a.

• To jest zadanie interpolacji Hermite'a, gdzie każdy węzeł interpolacyjny jest podwójny. Zadanie ma więc jednoznaczne rozwiązanie $w \in \Pi_{2n+1}$. Wielomian w interpoujący ogólniejsze dane $w(x_i) = a_i$ i $w'(x_i) = b_i$ najprościej zapisać wykorzystując, podobnie jak dla węzłów jednokrotnych, bazę kanoniczną, tzn.

$$w(x) = \sum_{i=0}^{n} a_i p_i(x) + b_i q_i(x),$$

gdzie $p_i, q_i \in \Pi_{2n+1}$,

$$\begin{aligned} p_i(x_j) &= \delta_{i,j}, \quad & p_i'(x_j) &= 0, \\ q_i(x_j) &= 0, & q_i'(x_j) &= \delta_{i,j}, \end{aligned}$$

$0 \leqslant i, j \leqslant n$. Teraz trzeba wyrazić p_i i q_i przez wielomiany Lagrange'a l_i dla węzłów x_i, $0 \leqslant i \leqslant n$. W tym celu zauważmy, że $l_i^2 \in \Pi_{2n}$ i ma podwójne zera w punktach x_j dla $j \neq i$. Stąd p_i musi być postaci $p_i(x) = (\alpha_i x + \beta_i) l_i^2(x)$. Wykorzystując warunki interpolacyjne dla p_i w x_i dostajemy

$$p(x_i) = \alpha_i x_i + \beta_i = 1, \qquad p'(x_i) = \alpha_i + 2(\alpha_i x_i + \beta_i) l_i'(x_i) = \alpha_i + 2 l_i'(x_i) = 0,$$

a stąd $\alpha_i = -2l_i'(x_i)$, $\beta_i = 1 + 2x_i l_i'(x_i)$. Podobnie wyliczamy $q_i(x)$ i otrzymujemy

$$p_i(x) = \Big(1 - 2(x - x_i) l_i'(x_i)\Big) l_i^2(x), \qquad q_i(x) = (x - x_i) l_i^2(x).$$

Ponieważ w treści zadania mamy zerowe warunki interpolacyjne na pochodne, rozwiązaniem jest wielomian $w(x) = \sum_{i=0}^{n} y_i p_i(x)$.
□

(59) Stosując algorytm różnic dzielonych znajdź w bazie potęgowej wielomian $w \in \Pi_5$ interpolujący funkcję $f(x) = x^6$ w trzech dwukrotnych węzłach równych $0, 1$ i 2.

• Najpierw znajdziemy współczynniki wielomianu w w baie Newtona

$$1, \quad x, \quad x^2, \quad x^2(x-1), \quad x^2(x-1)^2, \quad x^2(x-1)^2(x-2).$$

Stosujemy algorytm różnic dzielonych.

$$\begin{array}{llllll} f[0] = 0 & & & & & \\ f[0] = 0 & f[0,0] = 0 & & & & \\ f[1] = 1 & f[0,1] = 1 & f[0,0,1] = 1 & & & \\ f[1] = 1 & f[1,1] = 6 & f[0,1,1] = 5 & f[0,0,1,1] = 4 & & \\ f[2] = 64 & f[1,2] = 63 & f[1,1,2] = 57 & f[0,1,1,2] = 26 & f[0,0,1,1,2] = 11 & \\ f[2] = 64 & f[2,2] = 192 & f[1,2,2] = 129 & f[1,1,2,2] = 72 & f[0,1,1,2,2] = 23 & f[0,0,1,1,2,2] = 6 \end{array}$$

Stąd

$$\begin{aligned} w(x) &= f[0] + f[0,0]x + f[0,0,1]x^2 + f[0,0,1,1]x^2(x-1) \\ &\quad + f[0,0,1,1,2]x^2(x-1)^2 + f[0,0,1,1,2,2]x^2(x-1)^2(x-2) \\ &= x^2 + 4x^2(x-1) + 11x^2(x-1)^2 + 6x^2(x-1)^2(x-2) \\ &= 6x^5 - 13x^4 + 12x^3 - 4x^2. \end{aligned}$$

Zauważ, że to zadanie można również rozwiązać przez zastosowanie wyniku z zadania (58).
□

(60) Funkcję $f(x) = x^4$ interpolujemy wielomianem Hermite'a w dwóch podwójnych węzłach: $x = 0$ i $x = 1$.
(a) Wyznacz wielomian interpolacyjny w odpowiedniej bazie Newtona.
(b) Uzasadnij, że dla każdego $x \in [0,1]$ błąd interpolacji można oszacować przez $\frac{1}{16}$.

• Baza Newtona to: 1, x, x^2, $(x-1)x^2$. Stosujemy algorytm różnic dzielonych.

$$\begin{array}{llll} f[0] = 0 & & & \\ f[0] = 0 & f[0,0] = 0 & & \\ f[1] = 1 & f[0,1] = 1 & f[0,0,1] = 1 & \\ f[1] = 1 & f[1,1] = 4 & f[0,1,1] = 3 & f[0,0,1,1] = 2 \end{array}$$

Stąd szukany wielomian interpolacyjny w bazie Newtona wynosi

$$w(x) = x^2 + 2(x-1)x^2,$$

albo w wygodniejszej formie $w(x) = 2x^3 - x^2 = x^2(2x-1)$. Ze wzoru na błąd interpolacji mamy

$$f(x) - w(x) = x^2(x-1)^2 f[0,0,1,1,x] = x^2(x-1)^2 \frac{f^{(4)}(\xi)}{4!} = x^2(x-1)^2.$$

Moduł funkcji po prawej stronie tego równania przyjmuje maksimum w $x = \frac{1}{2}$ i to maksimum wynosi $\frac{1}{16}$.
□

(61) Funkcję $f \in C^{n+1}([a,b])$ interpolujemy wielomianem $w_f \in \Pi_n$ w węzłach

$$a = x_0 < x_1 < \cdots < x_n = b.$$

Wykaż, że jeśli $f \notin \Pi_n$ i pochodna $f^{(n+1)}$ nie zmienia znaku w $[a,b]$ to jedynymi rozwiązaniami równania $f(x) = w_f(x)$ w przedziale $[a,b]$ są węzły x_i.

Czy założenie, że $x_0 = a$ i $x_n = b$ ma znaczenie?

(62) Funkcję $f : [0,1] \to \mathbb{R}$ interpolujemy kawałkami wielomianem $\overline{w}_f$ stopnia r na kolejnych podprzedziałach $\left[\frac{i-1}{k}, \frac{i}{k}\right]$, $1 \leqslant i \leqslant k$. Wykaż, że jeśli $f \in C^r([0,1])$ i $\|f^{(r)}\|_{C([0,1])} \leqslant M$ to maksymalny błąd interpolacji

$$\max_{0 \leqslant x \leqslant 1} |f(x) - \overline{w}_f(x)| \leqslant \frac{2\,M}{r!} \left(\frac{1}{k}\right)^r.$$

W szczególności, błąd zbiega do zera tak szybko jak k^{-r} gdy $k \to +\infty$.

• Niech $x \in \left[\frac{i-1}{k}, \frac{i}{k}\right]$ i niech $x_{i,j}$ dla $0 \leqslant j \leqslant j$ będą węzłami wykorzystywanymi przez $\overline{w}_f$ w tym przedziale. Przyjmując, bez straty ogólności, że $x \neq x_{i,0}$, mamy

$$\begin{aligned} f(x) - \overline{w}_f(x) &= (x - x_{i,0})(x - x_{i,1}) \cdots (x - x_{i,r}) f[x_{i,0}, x_{i,1}, \ldots, x_{i,r}, x] \\ &= (x - x_{i,0})(x - x_{i,1}) \cdots (x - x_{i,r}) \frac{f[x_{i,1}, \ldots, x_{i,r}, x] - f[x_{i,0}, , \ldots, x_{i,r}]}{x - x_{i,0}} \\ &= (x - x_{i,1}) \cdots (x - x_{i,r}) \Big(f[x_{i,1}, \ldots, x_{i,r}, x] - f[x_{i,0}, , \ldots, x_{i,r}] \Big), \end{aligned}$$

a stąd

$$|f(x) - \overline{w}_f(x)| \leqslant \left(\frac{1}{k}\right)^r \left(\frac{|f^{(r)}(\xi_1)|}{r!} + \frac{|f^{(r)}(\xi_2)|}{r!} \right) \leqslant \frac{2\,M}{r!} \left(\frac{1}{k}\right)^r.$$

□

(63) Pokaż, że dla $k \geqslant 1$ mamy

$$T_k(x) = \frac{(x + \sqrt{x^2 - 1})^k + (x - \sqrt{x^2 - 1})^k}{2},$$

gdzie T_k jest k-tym wielomianem Czebyszewa.

• Pokażemy najpierw, że funkcja po prawej stronie, którą oznaczymy przez $P(x)$, jest wielomianem. Rzeczywiście, mamy

$$P(x) = \frac{1}{2} \sum_{i=0}^{k} \binom{k}{i} x^{k-i} \Big((x^2 - 1)^{i/2} + (-1)^i (x^2 - 1)^{i/2} \Big) = \sum_{i=0}^{\lfloor k/2 \rfloor} \binom{k}{2i} x^{k-2i} (x^2 - 1)^i,$$

czyli $P \in \Pi_n$. Dalej, podstawiając $x = \cos t$ dla $t \in [0, \pi]$ dostajemy dla $|x| \leqslant 1$, że

$$\begin{aligned} P(x) &= \frac{1}{2} \Big((\cos t + \sqrt{\cos^2 t - 1})^k + (\cos t - \sqrt{\cos^2 t - 1})^k \Big) \\ &= \frac{1}{2} \Big((\cos t + i \sin t)^k + (\cos t - i \sin t)^k \Big) \\ &= \frac{1}{2} \Big((\cos kt + i \sin kt) + (\cos kt - i \sin kt) \Big) \\ &= \cos kt = T_k(x). \end{aligned}$$

Wielomiany P i T_k są więc tożsame na odcinku $[-1, 1]$, a stąd $P = T_k$.
□

(64) Pokaż, że wielomiany Czebyszewa pierwszego rodzaju $\{T_k\}_{k \geqslant 0}$ tworzą układ ortogonalny w przestrzeni $\mathcal{L}_{2,\rho}(-1, 1)$, gdzie waga $\rho(x) = (1 - x^2)^{-1/2}$, tzn. iloczyn skalarny

$$\langle T_i, T_j \rangle := \int_{-1}^{1} \frac{T_i(x) T_j(x)}{\sqrt{1 - x^2}}\, dx = \begin{cases} \pi & i = j = 0, \\ \pi/2 & i = j \neq 0, \\ 0 & i \neq j. \end{cases}$$

• Po zamianie zmiennych $x = -\cos t$, $t \in [0,\pi]$, mamy $T_k(x) = \cos kt$. Stąd i ze wzoru na sumę cosinusów dostajemy

$$\begin{aligned}\int_{-1}^{1} \frac{T_i(x)T_j(x)}{\sqrt{1-x^2}}\,dx &= \int_0^{\pi} \frac{\cos it\,\cos jt}{\sqrt{1-\cos^2 t}}\sin t\,\mathrm{d}t = \int_0^{\pi} \cos it\,\cos jt\,\mathrm{d}t \\ &= \frac{1}{2}\Big(\int_0^{\pi}\cos(i+j)t\,\mathrm{d}t + \int_0^{\pi}\cos(i-j)t\,\mathrm{d}t\Big).\end{aligned}$$

Teraz wystarczy zastosować równość $\int_0^{\pi}\cos 0\,\mathrm{d}t = \pi$ oraz

$$\int_0^{\pi}\cos kt\,\mathrm{d}t = \frac{1}{k}\sin kt\Big|_0^{\pi} = \frac{1}{k}\sin \pi t = 0 \qquad \text{dla} \quad k \neq 0.$$

□

(65) Pokaż, że wielomiany Czebyszewa drugiego rodzaju $\{U_k\}_{k\geqslant 0}$ tworzą układ ortogonalny w przestrzeni $\mathcal{L}_{2,\rho}(-1,1)$, gdzie waga $\rho(x) = (1-x^2)^{1/2}$, tzn. iloczyn skalarny

$$\langle U_i, U_j\rangle := \int_{-1}^{1} U_i(x)U_j(x)\sqrt{1-x^2}\,dx = \begin{cases} \pi/2 & i = j, \\ 0 & i \neq j. \end{cases}$$

• Po zamianie zmiennych $x = -\cos t$, $t \in [0,\pi]$, mamy $U_k(x) = \frac{\sin(k+1)t}{\sin t}$. Stąd i ze wzoru na sumę sinusów dostajemy

$$\begin{aligned}\int_{-1}^{1} U_i(x)U_j(x)\sqrt{1-x^2}\,dx &= \int_0^{\pi}\frac{\sin(i+1)t}{\sin t}\,\frac{\sin(j+1)t}{\sin t}\,\sqrt{1-\cos^2 t}\sin t\,\mathrm{d}t \\ &= -\frac{1}{2}\Big(\int_0^{\pi}\cos(i+j+2)t\,\mathrm{d}t - \int_0^{\pi}\cos(i-j)t\,\mathrm{d}t\Big).\end{aligned}$$

Dalej postępujemy jak w rozwiązaniu zadania (64).
□

(66) Zaproponuj algorytm obliczania wartości $p(x)$ wielomianu p danego przez współczynniki $c_0, c_1, \ldots, c_n$ jego rozwinięcia w bazie Czebyszewa, tzn.

$$p(x) = \sum_{k=0}^{n} c_k T_k(x).$$

Algorytm ma działać w czasie proporcjonalnym do n.

• Niech $d_{n+2} = 0$, $d_{n+1} = 0$, oraz $d_k = c_k + 2xd_{k+1} - d_{k+2}$ dla $k = n, n-1, \ldots, 1, 0$. Wtedy

$$\begin{aligned}p(x) &= \sum_{k=0}^{n} c_k T_k(x) = \sum_{k=0}^{n}(d_k - 2xd_{k+1} + d_{k+2})\,T_k(x) \\ &= d_0T_0(x) + d_1\Big(T_1(x) - 2xT_0(x)\Big) + \sum_{k=2}^{n} d_k\Big(T_k(x) - 2xT_{k-1}(x) + T_{k-2}(x)\Big) \\ &= d_0 - d_1x,\end{aligned}$$

gdzie użyliśmy formuły rekurencyjnej dla wielomianów Czebyszewa. Stąd algorytm:

$d_{n+2} := 0;\ d_{n+1} := 0;$
for $k := n$ **downto** 0 **do**
$\quad d_k := c_k + 2\,x\,d_{k+1} - d_{k+2};$
$p := d_0 - d_1 x.$
□

(67) Niech dane będą współczynniki c_i wielomianu w w bazie Czebyszewa, $w(x) = \sum_{k=0}^{n} c_k T_k(x)$. Podaj możliwie tani algorytm obliczający współczynniki a_i tego wielomianu w bazie potęgowej, czyli w postaci $w(x) = \sum_{i=0}^{n} a_i x^i$.

(68) Niech w będzie wielomianem, którego rozwinięciem w bazie Czebyszewa jest $w(x) = \sum_{k=0}^{n} c_k T_k(x)$. Pokaż, że wielomian

$$w_1(x) = \sum_{k=0}^{n-1} c_k T_k(x)$$

najlepiej przybliża w w normie Czebyszewa na przedziale $[-1, 1]$ spośród wszystkich wielomianów stopnia co najwyżej $n-1$, oraz

$$\|w - w_1\|_C = |c_n|.$$

Czy $w_2(x) = \sum_{k=0}^{n-2} c_k T_k(x)$ najlepiej przybliża w w tej samej normie w przestrzeni Π_{n-2}?

• Ponieważ różnica $w(x) - w_1(x) = c_n T_n(x)$ ma $(n+1)$-elementowy alternans w punktach $z_j = -\cos\left(\frac{j\pi}{n}\right)$ dla $0 \leqslant j \leqslant n$, na podstawie twierdzenia o alternansie wielomian w_1 jest optymalny. Ponadto

$$\|w - w_1\|_C = |c_n|\,\|T_n\|_C = |c_n|,$$

bo $\|T_n\|_C = 1$.

Wielomian w_2 nie jest w ogólności optymalny dla w. Na przykład, dla $w(x) = T_2(x) + T_1(x) = 2x^2 + x - 1$ optymalnym nie jest wielomian zerowy, bo wykresem w jest parabola oraz $w(-1) = 0$, $w(1) = 2$, a stąd dla zerowej aproksymacji trzypunktowy alternans nie istnieje.
□

(69) Załóżmy, że wielomian $w(x) = \sum_{k=0}^{n} c_k T_k(x)$ aproksymujemy na przedziale $[-1, 1]$ wielomianem $w_i(x) = \sum_{k=0}^{n-i} c_k T_k(x)$, gdzie $0 \leqslant i \leqslant n$. Uzasadnij, że

$$\|w - w_i\|_C \leqslant \sum_{k=n-i+1}^{n} |c_k|.$$

Kiedy w tej nierówności mamy równość?

• Oczywiście,

$$\|w - w_i\|_C = \Big\| \sum_{k=n-i+1}^{n} c_k T_k \Big\|_C \leqslant \sum_{k=n-i+1}^{n} |c_k|\,\|T_k\|_C = \sum_{k=n-i+1}^{n} |c_k|.$$

W nierówności mamy równość wtedy i tylko wtedy gdy istnieje $z \in [-1, 1]$ takie, że dla wszystkich $n - i + 1 \leqslant k \leqslant n$ mamy $c_k T_k(z) = |c_k|$, albo dla wszystkich takich k mamy $c_k T_k(z) = -|c_k|$. To zaś jest równoważne temu, że z jest punktem ekstremalnym dla wszystkich T_k takich, że $c_k \neq 0$ oraz dodatkowo dla takich k jest

$T_k(z) = \operatorname{sgn} c_k$. Na przykład, jeśli wszystkie c_k są nieujemne, albo wszystkie c_k są niedodatnie to $z = 1$, a jeśli c_k naprzemiennie są nieujemne i niedodatnie to $z = -1$.
□

(70) Niech dany będzie przedział skończony $[a, b]$ nie zawierający zera, oraz liczba c. Niech

$$\tilde{\Pi}_k \;=\; \{\, w \in \Pi_k :\; w(0) = c \,\}.$$

Pokaż, że w klasie $\tilde{\Pi}_k$ najmniejszą normę Czebyszewa na $[a, b]$ ma wielomian

$$w(x) \;=\; c\,\frac{T_k\left(\frac{2x-(a+b)}{b-a}\right)}{T_k\left(\frac{a+b}{a-b}\right)} \qquad \text{oraz} \qquad \|w\|_C \;=\; \frac{|c|}{\left|T_k\left(\frac{a+b}{a-b}\right)\right|}.$$

Jakie jest rozwiązanie, gdy $0 \in [a, b]$?

(71) Niech P będzie zbiorem wszystkich wielomianów rzeczywistych w stopnia co najwyżej n spełniających $\sup_{|x|\leqslant 1} |w(x)| \leqslant 1$. Wykaż, że

$$\max_{w\in P} w(2) \;=\; T_n(2),$$

gdzie T_n jest n-tym wielomianem Czebyszewa.

• Ponieważ $T_n \in P$ to rzeczona maksymalna wartość wynosi co najmniej $T_n(2)$ i jest dodatnia. Załóżmy, że istnieje $w \in P$ taki, że $w(2) > T_n(2)$. Możemy założyć bez straty ogólności, że $\|w\|_{C([-1,1])} < 1$, bo w przeciwnym przypadku wzięlibyśmy wielomian $w_1 = a\,w$, gdzie $T_n(2)/w(2) < a < 1$. Wtedy wielomian $p := w - T_n$ jest niezerowy i przyjmuje naprzemiennie wartości dodatnie i ujemne w 2 i w kolejnych ekstremach wielomianu T_n na odcinku $[-1, 1]$. Ponieważ T_n ma $n + 1$ punktów ekstremalnych, p zeruje się w co najmniej n różnych punktach w $(-1, 1)$ oraz dodatkowo w jakimś punkcie przedziału $(1, 2)$. To jest niemożliwe, bo $\deg p \leqslant n$.
□

(72) Wykaż, że spośród wielomianów p stopnia co najwyżej n $(n \geqslant 1)$ spełniających $p'(1) = A$, najmniejszą normę jednostajną na $[-1, 1]$ ma wielomian AT_n/n^2, gdzie T_n jest n-tym wielomianem Czebyszewa.

• Najpierw pokażemy, że $T_n'(1) = n^2$. Stosując indukcję po n mamy $T_1'(1) = 1$, $T_2'(1) = [(2x^2 - 1)']_{x=1} = 4 = 2^2$. Dla $n \geqslant 3$ mamy

$$T_n'(x) = (2xT_{n-1}(x) - T_{n-2}(x))' = 2T_n(x) + 2xT_{n-1}'(x) - T_{n-2}'(x),$$

a stąd i z założenia indukcyjnego

$$T_n'(1) = 2 + 2(n-1)^2 - (n-2)^2 = n^2.$$

Wielomian $w_n := AT_n/n^2$ spełnia więc warunki zadania. Przypuśćmy, że istnieje wielomian $p \in \Pi_n$ taki, że $p'(1) = A$ i $\|p\|_C < \|w_n\|_C$. Z własności ekstremalnych wielomianów Czebyszewa wynika, że wtedy $p - w_n$ ma n różnych zer w $(-1, 1)$. Stąd $(p - w_n)'$ ma $n - 1$ zer w $(-1, 1)$ i dodatkowe zero w 1. Ponieważ $\deg(p - w_n)' \leqslant n - 1$ to $p = w_n$, co przeczy założeniu, że $\|p\|_C < \|w_n\|_C$.
□

(73) Wyznacz w bazie potęgowej wielomian $w \in \Pi_3$ najlepiej aproksymujący funkcję $f(x) = 128\,x^4$ w przestrzeni $C([0,1])$.

• Ponieważ $f(x) - w(x) = 128\,x^4 + \cdots$, to $f - w$ minimalizuje normę jednostajną na $[0,1]$ wśrod wielomianów stopnia 4 o współczynniku wiodącym 128. Wiadomo, że takim wielomianem jest

$$\begin{aligned} v(x) &= 128\,\frac{1}{2^7}\,T_4(2x-1) = 8(2x-1)^4 - 8(2x-1)^2 + 1 \\ &= 128\,x^4 - 256\,x^3 + 160\,x^2 - 32\,x + 1, \end{aligned}$$

a stąd

$$w(x) = f(x) - v(x) = 256\,x^3 - 160\,x^2 + 32\,x - 1.$$

□

(74) Wykaż, że jeśli funkcja $f \in C([-1,1])$ jest parzysta, tzn. $f(-x) = f(x)$, to jej wielomian optymalny w Π_n jest parzysty, a jeśli f jest nieparzysta, tzn. $f(-x) = -f(x)$, to jej wielomian optymalny w Π_n jest nieparzysty.

• Załóżmy, że f jest parzysta i w jest optymalny dla f. Wtedy wielomian $w_1(x) = w(-x)$ też jest optymalny, bo dla każdego $x \in [-1,1]$ mamy

$$f(x) - w_1(x) = f(-x) - w(-x),$$

a stąd $\|f - w_1\|_C = \|f - w\|_C$. Ponieważ wielomian optymalny jest wyznaczony jednoznacznie to $w_1 = w$, czyli w jest parzysty.

Jeśli zaś f jest nieparzysta i w jest optymalny dla f to wielomian $w_1(x) = -w(-x)$ jest też optymalny dla f, bo

$$f(x) - w_1(x) = -f(-x) + w(-x),$$

a stąd znów $\|f - w_1\|_C = \|f - w\|_C$. Z jednoznaczności wielomianu optymalnego dostajemy $w_1 = w$, czyli w jet nieparzysty.
□

(75) Funkcję $f(x) = |x|$ aproksymujemy wielomianem interpolacyjnym stopnia pierwszego opartym na węzłach x_0, x_1. Dla których węzłów błąd takiej aproksymacji w normie jednostajnej $C([-1,2])$ jest najmniejszy?

• Wielomian stopnia 1 optymalny dla f w normie $C([-1,2])$ jest jednocześnie wielomianem interpolacyjnym stopnia 1 dla f, opartym na pewnych węzłach x_0^*, x_1^*, cf. zadanie (86). Węzły te możemy znaleźć korzystając z twierdzenia o alternansie. Ponieważ f jest wypukła to optymalny wielomian jest postaci $w^*(x) = ax + b$, gdzie

$$a = \frac{f(2) - f(-1)}{2 - (-1)} = \frac{1}{3},$$

a b jest takie, że $f(2) - w^*(2) = w^*(0) - f(0)$, czyli $b = \frac{2}{3}$. Aby znaleźć x_0^*, x_1^*, wystarczy teraz rozwiązać równanie $\frac{1}{3}(x+2) = |x|$. Dostajemy $x_0^* = -\frac{1}{2}$, $x_1^* = 1$. Dodatkowo, minimalny błąd interpolacji/aproksymacji wynosi $\frac{2}{3}$.
□

(76) Czy $V = \operatorname{span}(1, x^2, x^3)$ jest przestrzenią Haara w
(i) $C([0,1])$,
(ii) $C([-1,1])$?

• Zauważmy, że $\dim V = 3$. Jeśli $v \in V$ zeruje się w trzech punktach $x_0 < x_1 < x_2$ dziedziny to $v(x) = a(x-t_0)(x-t_1)(x-t_2)$, albo

$$v(x) = x^3 - x^2(t_0+t_1+t_2) + x(t_0t_1+t_0t_2+t_1t_2) - t_0t_1t_2.$$

Warunek $v \in V$ jest równoważny temu, że współczynnik przy x się zeruje, czyli

$$t_0t_1 + t_0t_2 + t_1t_2 = 0.$$

W przypadku (i) jest to niemożliwe, bo powyższa suma jest nie mniejsza od $x_1x_2 > 0$, czyli V jest przestrzenią Haara. W przypadku (ii), V nie jest przestrzenią Haara, bo możemy wziąć, np. $(t_0,t_1,t_2) = \big(-\frac{1}{3}, \frac{1}{2}, 1\big)$.
□

(77) Niech a_i dla $1 \leqslant i \leqslant n$ będą punktami nie należącymi do danego przedziału $[a,b]$. Wykaż, że

$$V \;=\; \operatorname{span}\left\{\frac{1}{x-a_1}, \frac{1}{x-a_2}, \ldots, \frac{1}{x-x_n}\right\},$$

jest przestrzenią Haara w $C([a,b])$.

(78) Pokaż, że przestrzeń funkcji trygonometrycznych stopnia n, zdefiniowana jako

$$\mathcal{V}_{2n+1} = \operatorname{span}\big(1, \cos t, \sin t, \ldots, \cos nt, \sin nt\big),$$

jest przestrzenią Haara w $C([a,b])$, o ile $0 < b-a < 2\pi$.

• Wobec równości $\mathrm{e}^{\mathrm{i}\phi} = \cos\phi + \mathrm{i}\sin\phi$ (gdzie $\mathrm{i} = \sqrt{-1}$), każda nietrywialna funkcja $t \mapsto h(t) := \sum_{k=0}^{n} a_k \cos kt + b_k \sin kt \in \mathcal{V}_{2n+1}$ może być zapisana w postaci

$$h(t) = \sum_{|k|\leqslant n} c_k \mathrm{e}^{\mathrm{i}kt} = z^{-n} \sum_{k=0}^{2n} c_{k-n} z^k,$$

gdzie $z = \mathrm{e}^{\mathrm{i}t}$ oraz $c_0 = a_0$, $c_{\pm k} = \frac{1}{2}(a_k \mp \mathrm{i}b_k)$, $1 \leqslant k \leqslant n$. Jeśli $c_{\pm k} = 0$ dla wszystkich k to także $a_k = b_k = 0$, co oznacza, że $\dim \mathcal{V}_{2n+1} = 2n+1$. Ponadto, h znaka w nie więcej niż $2n$ punktach przedziału $[a,b]$, ponieważ każdy wielomian po prawej stronie równości ma co najwyżej $2n$ zer, a funkcja $t \mapsto \mathrm{e}^{\mathrm{i}t}$ jest różnowartościowa dla t z przedziału o długości mniejszej niż 2π.
□

(79) Pokaż, że

$$\widehat{\mathcal{V}}_{n+1} = \operatorname{span}(1, \cos t, \cos 2t, \ldots, \cos nt)$$

jest przestrzenią Haara w $C([0,\pi])$.

• Każda nietrywialna funkcja $t \mapsto h(t) := \sum_{k=0}^{n} a_k \cos kt \in \widehat{\mathcal{V}}_{n+1}$ może być zapisana w postaci $h(t) = w(x) := \sum_{k=0}^{n} a_k T_k(x)$ gdzie $x = \cos t$, $t \in [0,\pi]$, a T_k jest k-tym wielomianem Czebyszewa. Ponieważ wielomian w ma co najwyżej n zer, a zamiana zmiennych jest różnowartościowa, funkcja h też ma co najwyżej n zer w $[0,\pi]$.
□

(80) Pokaż, że

$$\widetilde{\mathcal{V}}_n = \operatorname{span}(\sin t,\, \sin 2t,\, \ldots,\, \sin nt)$$

jest przestrzenią Haara w $C([\epsilon, \pi - \epsilon])$ dla każdego $\epsilon \in (0, \pi/2)$.

• Wystarczy zauważyć, że jeśli funkcja $t \mapsto h(t) := \sum_{k=1}^{n} b_k \sin kt$ ma n zer $\epsilon \leqslant t_1 < \ldots < t_n \leqslant \pi - \epsilon$ to ma $2n+1$ zer w $[-\pi + \epsilon, \pi - \epsilon]$; mianowicie 0 i $\pm t_j$, $1 \leqslant j \leqslant n$. To zaś w świetle zadania (78) jest niemożliwe.
□

(81) Znajdź wielomian $w \in \Pi_1$ w bazie potęgowej najlepiej aproksymujący funkcję $\sqrt{x}$
(i) w normie jednostajnej $C([0,1])$,
(ii) w normie średniokwadratowej $\mathcal{L}_2(0,1)$.

• Oznaczmy $f(x) = \sqrt{x}$. W (i) skorzystamy z twierdzenia o alternansie. Ponieważ $\dim \Pi_1 = 2$, alternans liczy 3 punkty. Ponadto f jest funkcją wklęsłą na $[0,1]$. Proste rozważania geometryczne prowadzą do wniosku, że alternans składa się z punktów $x_0 = 0$ i $x_2 = 1$, oraz punktu $x_1 \in (0,1)$ takiego, że styczna w_1 do wykresu funkcji w tym punkcie jest równoległa do wielomianu w_2 stopnia 1 interpolującego f w punktach 0 i 1. Wtedy element optymalny wynosi $w^* = \frac{1}{2}(w_1 + w_2)$.

Z powyższych rozważań wynika, że x_1 spełnia równanie $f'(x_1) = f[0,1] = 1$, czyli $x_1 = \frac{1}{4}$. Ponieważ $f(x_1) = \frac{1}{2}$, mamy $w_1(x) = x + \frac{1}{4}$. Mamy też $w_2(x) = x$, a stąd $w^* = x + \frac{1}{8}$. Ponadto, błąd aproksymacji wynosi $\frac{1}{8}$.

W (ii), łatwo widać, że wielomiany $w_0(x) = 1$ i $w_1(x) = x - \frac{1}{2}$ są prostopadłe w $\mathcal{L}_2(0,1)$. Dlatego element optymalny, będący jednocześnie rzutem prostopadłym f na podprzestrzeń Π_1, wynosi

$$w^*(x) = \frac{\langle f, w_0\rangle}{\langle w_0, w_0\rangle} + \frac{\langle f, w_1\rangle}{\langle w_1, w_1\rangle}\Big(x - \frac{1}{2}\Big).$$

Mamy $\langle w_0, w_0\rangle = \int_0^1 1^2\,\mathrm{d}x = 1$, $\langle w_1, w_1\rangle = \int_0^1 (x-\frac{1}{2})^2 \mathrm{d}x = \frac{1}{12}$, $\langle f, w_0\rangle = \int_0^1 \sqrt{x}\,\mathrm{d}x = \frac{2}{3}$, $\langle f, w_1\rangle = \int_0^1 \sqrt{x}(x - \frac{1}{2})\,\mathrm{d}x = \frac{1}{15}$, a stąd

$$w^*(x) = \frac{2}{3} + \frac{4}{5}\Big(x - \frac{1}{2}\Big) = \frac{4}{5}\,x + \frac{4}{15}.$$

□

(82) Wyznacz w postaci potęgowej wielomian $w_3 \in \Pi_3$ najlepiej aproksymujący funkcję $f(x) = |x|$ w przestrzeni $C([-1,1])$.

• Ponieważ f jest parzysta to wielomian optymalny też jest parzysty, czyli postaci $w_3(x) = ax^2 + b$. Po zamianie zmiennych $z = x^2$ zadanie sprowadza się do najlepszej aproksymacji funkcji $\sqrt{z}$ przez wielomiany postaci $v(z) = az + b$ na odcinku $[0,1]$. Rozwiązanie zadania (81) daje optymalny wielomian $v(z) = z + \frac{1}{8}$, czyli $w_3(x) = x^2 + \frac{1}{8}$.
□

(83) Znajdź współczynniki a_0, a_1, a_2 wielomianu trygonometrycznego

$$g(t) = a_0 + a_1 \cos t + a_2 \sin t,$$

który najlepiej aproksymuje funkcję $f(t) = |\sin(t/2)|$ w przestrzeni $C([-\pi, \pi])$.

• Ponieważ funkcja f jest parzysta, wystarczy ograniczyć się do aproksymacji parzystych w $C([0,\pi])$. Rzeczywiście, jeśli g jest optymalny dla f to $h(t)=\frac{1}{2}(g(t)+g(-t))$ jest w klasie, jest parzysty, a także optymalny. Dla dowolnego t mamy bowiem

$$\begin{aligned}|f(t)-h(t)| &= \tfrac{1}{2}\Big((f(t)-g(t))+(f(-t)-g(-t))\Big)\\ &\leqslant \tfrac{1}{2}\Big(|f(t)-g(t)|+|f(-t)-g(-t)|\Big)\leqslant \|f-g\|_C.\end{aligned}$$

(Por. z zadaniem (74)). Skoro g jest parzysta tylko wtedy gdy $a_2=0$, zadanie jest równoważne aproksymacji w podprzestrzeni rozpiętej na 1 i $\cos t$.

Dokonując zamiany zmiennych $x=\cos t$ i wykorzystując tożsamość

$$\left|\sin\left(\frac{t}{2}\right)\right|=\sqrt{\frac{1-\cos t}{2}},$$

sprowadzamy zadanie do znalezienia a_0,a_1 takich, że wielomian $w(x)=a_0+a_1x$ najlepiej aproksymuje funkcję $\sqrt{\frac{1-x}{2}}$ w $C([-1,1])$. Aby je rozwiązać, wykorzystamy rozwiązanie zadania (81)(i). Łatwo zauważyć, że skoro wielomian $w(y)=y+\frac{1}{8}$ jest optymalny w Π_1 dla $\sqrt{y}$ to wielomian

$$w^*(x)=w\Big(\tfrac{1}{2}(1-x)\Big)=\tfrac{1}{2}(1-x)+\tfrac{1}{8}=-\tfrac{1}{2}x+\tfrac{5}{8}$$

jest optymalny dla $\sqrt{\frac{1-x}{2}}$. Stąd $a_0=\frac{5}{8}$, $a_1=-\frac{1}{2}$, $a_2=0$.
□

(84) Wykaż, że n-ty wielomian optymalny $w^*_{f,n}\in\Pi_n$ w aproksymacji jednostajnej dla funkcji $f\in C([a,b])$ jest jednoznacznie wyznaczony.

• Jeśli istnieją dwa elementy optymalne $w_1,w_2\in\Pi_n$ dla funkcji f to $w:=\frac{1}{2}(w_1+w_2)$ jest też optymalny dla f, bowiem dla każdego x mamy

$$\begin{aligned}|f(x)-w(x)| &= \left|\tfrac{1}{2}(f-w_1)(x)+\tfrac{1}{2}(f-w_2)(x)\right| \leqslant \tfrac{1}{2}\Big(\|f-w_1\|_C+\|f-w_2\|_C\Big)\\ &= \tfrac{1}{2}(E_n(f)+E_n(f)) = E_n(f).\end{aligned}$$

Niech $a\leqslant x_0<\cdots<x_n<x_{n+1}\leqslant b$ będzie alternansem dla $f-w$. Wstawiając powyżej $x=x_i$ dostajemy, że $|(f-w_1)(x_i)+(f-w_2)(x_i)|=2E_n(f)$, a stąd

$$(f-w_1)(x_i)=(f-w_2)(x_i)=\pm E_n(f),\qquad 0\leqslant i\leqslant n+1.$$

Wielomian $w_1-w_2\in\Pi_n$ ma więc $n+2$ zera, czyli $w_1=w_2$.
□

(85) Niech $x_0<x_1<\cdots<x_n$, gdzie $n\geqslant 1$. Niech $f\in C^1([x_0,x_n])$ będzie funkcją, która w kolejnych punktach x_i przyjmuje na przemian wartości nieujemne i niedodatnie. Pokaż, że wtedy f ma zera o łącznej krotności co najmniej n.

• Dowód przeprowadzimy przez indukcję po n. Dla $n=1$ fakt jest oczywisty. Niech $n=2$. Jeśli $f(x_1)\neq 0$ to f ma zero w $[x_0,x_1)$ oraz zero w $(x_1,x_2]$. Jeśli zaś $f(x_1)=0$ i jest to jedyne zero w $(x_0,x_2]$ to f nie zmienia znaku w tym przedziale, a to oznacza, że $f'(x_1)=0$, czyli x_1 jest zerem dwukrotnym.

Niech $n\geqslant 3$. Jeśli $f(x_{n-1})\neq 0$ to f ma z założenia indukcyjnego zera o łącznej krotności co najmniej $n-1$ w przedziale $[x_0,x_{n-1})$ oraz dodatkowe zero w $(x_{n-1},x_n]$. Jeśli zaś $f(x_{n-1})=0$ to mamy, podobnie jak w przypadku $n=2$, zera o łącznej

krotności co najmniej 2 w przedziale $(x_{n-2}, x_n]$, czyli, na podstawie założenia indukcyjnego dla przedziału $[x_0, x_{n-2}]$, zera o łącznej krotności $(n-2)+2 = n$ w całym przedziale $[x_0, x_n]$.
□

(86) Wykaż, że dla każdej funkcji $f \in C([a,b])$ można wskazać węzły $a < t_0 < \cdots < t_n < b$ takie, że wielomian interpolacyjny funkcji f oparty na tych węzłach jest jednocześnie jej n-tym wielomianem optymalnym.

• Jeśli $w^*_{f,n}$ jest n-tym wielomianem optymalnym dla f to dla różnica $f - w^*_{f,n}$ istnieje $(n+1)$-punktowy alternans $a \leqslant x_0 < \cdots < x_{n+1} \leqslant b$. To zaś oznacza, że w każdym przedziale otwartym (x_i, x_{i+1}) dla $0 \leqslant i \leqslant n$ istnieje t_i takie, że $(f - w^*_{f,n})(t_i) = 0$. Stąd $w^*_{f,n}(t_i) = f(t_i)$ czyli $w^*_{f,n}$ jest wielomianem w Π_n interpolującym f w punktach t_i dla $0 \leqslant i \leqslant n$. (Należy podkreślić, że rzeczone węzły interpolacyjne zmieniają się wraz ze zmianą f.)
□

(87) Uzasadnij, że stała Lebesgue'a

$$\Lambda_n := \inf_{a \leqslant x_0 < \cdots < x_n \leqslant b} \Lambda(x_0, \ldots, x_n),$$

gdzie $\Lambda(x_0, x_1, \ldots, x_n) := \max_{a \leqslant x \leqslant b} \sum_{j=0}^n |l_j(x)|$, a l_j są wielomianami Lagrange'a dla węzłów $x_0, \ldots, x_n$, nie zależy od wyboru przedziału $[a,b]$.

• Rozpatrzmy dowolny przedział $[a,b]$. Po zamianie zmiennych $x = a + t(b-a)$ i podstawieniu $x_k = a + t_k(b-a)$ dostajemy

$$\sum_{j=0}^n \prod_{j \neq k=0}^n \frac{x - x_k}{x_j - x_k} = \sum_{j=0}^n \prod_{j \neq k=0}^n \frac{t - t_k}{t_j - t_k}.$$

Ponieważ przekształcenie $t \mapsto x = a + t(b-a)$ jest bijekcją przedziału $[0,1]$ na przedział $[a,b]$, stała Legesgue'a dla $[a,b]$ jest równa stałej Lebesgue'a dla $[0,1]$.
□

(88) Niech $f(x) = x^2$. Jak duże n należy wziąć, aby wielomian Bernsteina $B_n f$ aproksymował f z błędem mniejszym niż 10^{-8} w normie jednostajnej na $[0,1]$?

• Dla funkcji $f(x) = x^2$ wielomian Bernsteina stopnia n wynosi:

$$\begin{aligned}
B_n(x) &= \sum_{k=0}^n \left(\frac{k}{n}\right)^2 \binom{n}{k} x^k (1-x)^{n-k} = \sum_{k=1}^n \binom{n-1}{k-1} x^k (1-x)^{n-k} \\
&= \sum_{k=1}^n \left(\frac{n-1}{n}\,\frac{k-1}{n-1} + \frac{1}{n}\right) \binom{n-1}{k-1} x^k (1-x)^{n-k} \\
&= \frac{n-1}{n}\, x^2 \sum_{k=2}^n \binom{n-2}{k-2} x^{k-2} (1-x)^{n-k} + \frac{x}{n} \\
&= \frac{n-1}{n}\, x^2 + \frac{x}{n}.
\end{aligned}$$

Stąd $f(x) - B(x) = \frac{1}{n} x(1-x)$ i $\|f - B\|_{C([0,1])} = \frac{1}{4n}$, co jest mniejsze od 10^{-8} dla $n > \frac{1}{4} \times 10^8$.
□

(89) Niech $B_n f$ będzie n-tym wielomianem Bernsteina dla funkcji f. Jak szybko błąd aproksymacji jednostajnej $\|f - B_n f\|_{C([0,1])}$ zbiega do zera dla funkcji $f(x) = x^3$?

(90) Wykaż, że jeśli $f \in C^1([0,1])$ to dla każdego $\epsilon > 0$ istnieje wielomian p taki, że $\|f - p\| \leqslant \epsilon$ i $\|f' - p'\| \leqslant \epsilon$, przy czym norma jest jednostajna na $[0,1]$.

• Skoro f' jest ciągła to istnieje wielomian q taki, że $\|f' - q\| \leqslant \epsilon$. Biorąc $p(x) = \int q(x)\,\mathrm{d}x$ mamy $\|f' - p'\| \leqslant \epsilon$ oraz

$$|f(x) - p(x)| \;=\; \left|\int_0^x (f-p)'(t)\,\mathrm{d}t\right| \;\leqslant\; \int_0^x |(f'-q)(t)|\,\mathrm{d}t \;\leqslant\; \epsilon\, x \;\leqslant\; \epsilon,$$

czyli $\|f - p\| \leqslant \epsilon$.
□

(91) Niech $h > 0$ i $c \in \mathbb{R}$. Wyznacz współczynniki kubicznej funkcji sklejanej s opartej na pięciu węzłach $-2h, -h, 0, h, 2h$ i spełniającej dodatkowo następujące warunki interpolacyjne:

$$s(0) = c, \qquad s^{(k)}(\pm 2h) = 0, \quad k = 0, 1, 2.$$

(92) Dla $z \in \mathbb{R}$, niech $z_+ = \max(0, z)$. Wykaż, że każdą naturalną kubiczną funkcję sklejaną s opartą na węzłach $x_0 < x_1 < \cdots < x_n$ można przedstawić w postaci

$$s(x) \;=\; w(x) \;+\; \sum_{k=0}^{n} a_k (x - x_k)_+^3,$$

gdzie $a_k \in \mathbb{R}$, a w jest wielomianem stopnia co najwyżej 1. Ponadto,

$$\sum_{k=0}^{n} a_k x_k^i = 0 \qquad \text{dla} \quad i = 0, 1.$$

(93) Pokaż, że każdą funkcję sklejaną s rzędu r opartą na węzłach $x_0 < \cdots < x_n$ można przedstawić w postaci

$$s(x) \;=\; w(x) \;+\; \sum_{k=0}^{n} a_k (x - x_k)_+^{2r-1},$$

gdzie a_j są pewnymi współczynnikami rzeczywistymi, a $w \in \Pi_{2r-1}$. Ponadto, jeśli s jest naturalna, to $w \in \Pi_{r-1}$, a współczynniki a_j spełniają równania

$$\sum_{k=0}^{n} a_k x_k^i \;=\; 0 \quad \text{dla} \quad 0 \leqslant i \leqslant r-1.$$

• $w \in \Pi_{2r-1}$ jest oczywiście wielomianem równym s na $(-\infty, x_0)$. Dla $x \in [x_0, x_1]$ zapiszemy $s - w$ w rozwinięciu Taylora

$$s(x) - w(x) = \sum_{i=0}^{2r-1} c_i \frac{(x - x_0)^i}{i!}.$$

Z warunków gładkości w x_0 mamy $(s-w)^{(k)}(x_0) = 0$ dla $0 \leqslant k \leqslant 2r-2$, a stąd $s(x) - w(x) = c_{2r-1}\frac{(x-x_0)^{2r-1}}{(2r-1)!}$ i

$$s(x) = w(x) + c_{2r-1}\frac{(x - x_0)_+^{2r-1}}{(2r-1)!} \qquad \text{dla} \quad -\infty < x \leqslant x_1,$$

czyli $a_0 = \frac{c_{2r-1}}{(2r-1)!}$. To samo rozumowanie można przeprowadzić dalej dla kolejnych przedziałów $[x_{j-1}, x_j]$ i $[x_n, +\infty)$.

Jeśli s jest naturalną funkcją sklejaną to oczywiście $w \in \Pi_{r-1}$, bo $s \in \Pi_{r-1}$ na $(-\infty, x_0]$. Pozostałe warunki uzyskujemy stąd, że $s \in \Pi_{r-1}$ także na $[x_n, +\infty)$. Rzeczywiście, mamy $s^{(r)}(x_n) = 0$, czyli

$$(2r-1)(2r-2)\cdots(r+1)r\sum_{k=0}^{n} a_k(x-x_k)^{r-1} = 0.$$

Równoważnie,

$$\begin{aligned}\sum_{k=0}^{n} a_k(x-x_k)^{r-1} &= \sum_{k=0}^{n} a_k \sum_{i=0}^{r-1}\binom{r-1}{i}(-1)^i x_k^i\, x^{r-1-i} \\ &= \sum_{i=0}^{r-1} x^{r-1-i}\binom{r-1}{i}(-1)^i\left(\sum_{k=0}^{n} a_k x_k^i\right) \;=\; 0.\end{aligned}$$

Przyrównując współczynniki przy x^i do zera dla $i = 0, 1, \ldots r-1$ dostajemy tezę.
□

(94) Niech $s : \mathbb{R} \to \mathbb{R}$ będzie funkcją sklejaną pierwszego rzędu (tzn. funkcją ciągłą i kawałkami wielomianem stopnia $\leqslant 1$) opartą na węzłach $x_0 < x_1 < \cdots < x_n$. Wykaż, że s można jednoznacznie przedstawić w postaci

$$s(x) = a + bx + \sum_{j=0}^{n} c_j|x - x_j|$$

dla pewnych a, b, c_j, $0 \leqslant j \leqslant n$. Jeśli ponadto s jest naturalna to $b = \sum_{j=0}^{n} c_j = 0$

• Oczywiście, każda funkcja rzeczonej postaci jest funkcją sklejaną pierwszego rzędu z węzłami x_j. Pokażemy, że układ funkcji 1, x, $|x - x_j|$, $0 \leqslant j \leqslant n$, jest liniowo niezależny. Rzeczywiście, niech $x \mapsto a + bx + \sum_{j=0}^{n} c_j|x - x_j|$ będzie funkcją zerową. Ponieważ $c_j \neq 0$ implikuje brak różniczkowalności w x_j, mamy $c_j = 0$ dla wszystkich j, a to oznacza, że także $a = b = 0$.

Funkcja s jest jednoznacznie wyznaczona przez swoje wartości w x_j dla $0 \leqslant j \leqslant n$ oraz wartości w $x_0 - 1$ i $x_n + 1$. Odpowiada to układowi $n+3$ równań liniowych z $n+3$ niewiadomymi a, b, c_j. Układ ten ma jednoznaczne rozwiązanie bo, jak pokazaliśmy wcześniej, zero jest jedynym rozwiązaniem układu jednorodnego.

Jeśli ponadto s jest naturalna to $s'(x_0^-) = 0 = s'(x_n^+)$. To odpowiada równaniom $b - \sum_{j=0}^{n} c_j = 0$ i $b + \sum_{j=0}^{n} c_j = 0$, czyli $b = 0 = \sum_{j=0}^{n} c_j$.
□

(95) Niech

$$Q_1^I(f) = \frac{b-a}{2}\left(f\Big(\frac{2a+b}{3}\Big) + f\Big(\frac{a+2b}{3}\Big)\right).$$

Pokaż nierówność

$$\sup_{f \in F_M^1([a,b])} |S(f) - Q_1^I(f)| \;\leqslant\; \frac{11}{324}M(b-a)^3 \qquad\qquad \left(\frac{11}{324} = 0.033951\right).$$

• Weźmy najpierw, dla uproszczenia, $[a,b]=[-1,1]$. Wtedy $Q_1^I(f)=f\Big(\frac{-1}{3}\Big)+f\Big(\frac{1}{3}\Big)$ i łatwo sprawdzić, że jest to kwadratura interpolacyjna. Stąd błąd

$$\begin{aligned}\big|S(f)-Q_1^I(f)\big| &= \left|\int_{-1}^{1}\big(x^2-\tfrac{1}{9}\big)f\big[-\tfrac{1}{3},\tfrac{1}{3},x\big]\,\mathrm{d}x\right| \leqslant \frac{1}{2}M\int_{-1}^{1}\big|x^2-\tfrac{1}{9}\big|\,\mathrm{d}x\\ &= \frac{22}{81}\,M \;=\; \frac{11}{324}\,2^3M,\end{aligned}$$

co odpowiada prawej stronie żądanej równości. Dla dowolnego przedziału wystarczy dokonać 'tradycyjnej' zamiany zmiennych $x=\frac{1}{2}(a+b)+\frac{1}{2}t(b-a)$, $t\in[-1,1]$.
□

(96) Niech $\bar{T}_n(f)$ będzie złożoną kwadraturą trapezów dla aproksymacji całki $S(f)=\int_a^b f(x)\,\mathrm{d}x$, opartą na równomiernym podziale odcinka $[a,b]$ na n pododcinków. Niech

$$\bar{T}_n'(f) \;=\; \bar{T}_n(f)-\frac{(b-a)^2}{12n^2}\Big(f'(b)-f'(a)\Big).$$

Wykaż, nie korzystając bezpośrednio z furmuły Eulera-Maclaurina, że jeśli $f\in C^4([a,b])$ to błąd kwadratury $|S(f)-\bar{T}_n'(f)|$ zbiega do zera co najmniej tak szybko jak n^{-4} gdy $n\to\infty$.

• Rozpatrzymy $n=1$ i kwadraturę

$$T'(f)=T(f)-\frac{(b-a)^2}{12}\Big(f'(b)-f'(a)\Big).$$

Jest to kwadratura oparta na dwukrotnych węzłach a i b. Jeśli rząd zbieżności kwadratury ma się zwiększyć z k^{-2} do k^{-4} to być może T' jest kwadraturą interpolacyjną Hermite'a. Sprawdzamy bezpośrednio, że rzeczywiście tak jest, np.przez sprawdzenie, że T' jest dokładna dla wielomianów 1, x, $(x-a)^2$, oraz $(x-a)\Big(x-\frac{1}{2}(a+b)\Big)(x-b)$.

Ze wzoru na błąd kwadratury interpolacyjnej dostajemy

$$\begin{aligned}S(f)-T'(f) &= \int_a^b(x-a)^2(x-b)^2f[a,a,b,b,x]\,\mathrm{d}x\\ &= \frac{f^{(4)}(\xi)}{4!}\int_a^b(x-a)^2(x-b)^2\,\mathrm{d}x \;=\; \frac{(b-a)^5}{720}f^{(4)}(\xi).\end{aligned}$$

Zauważmy teraz, że

$$\bar{T}_n'(f)=\sum_{i=1}^{n}\left(\frac{b-a}{2\,n}\Big(f(x_{i-1})+f(x_i)\Big)-\frac{(b-a)^2}{12\,n^2}\Big(f'(x_i)-f'(x_{i-1})\Big)\right),$$

czyli T_n' jest sumą kwadratur T' na kolejnych przedziałach $[x_{i-1},x_i]$, każdy o długości $\frac{1}{n}(b-a)$. Stąd

$$S(f)-T_n'(f)=\sum_{i=1}^{n}\frac{1}{720}\left(\frac{b-a}{n}\right)^5f^{(4)}(\xi_i) \;=\; \frac{(b-a)^5}{720}f^{(4)}(\xi)\,n^{-4},$$

co należało pokazać
□

(97) Wykaż, że jeśli $f \in C^4([a,b])$ i $\int_a^b f^{(4)}(x)\,\mathrm{d}x \neq 0$, to dla złożonej kwadratury parabol $\bar{P}_k$ mamy

$$\lim_{k\to\infty} \Big(S(f) - \bar{P}_k(f)\Big)\, k^4 \;=\; -\frac{(b-a)^4}{2880}\bigg(\int_a^b f^{(4)}(x)\,\mathrm{d}x\bigg).$$

• Błąd kwadratury $\bar{P}_k$ wyraża się dokładnym wzorem

$$S(f) - \bar{P}_k(f) \;=\; -\frac{1}{2880}\left(\frac{b-a}{k}\right)^5 \sum_{i=1}^{k} f^{(4)}(\xi_i) \;=\; -\frac{(b-a)^4}{2880}k^{-4}\left(\sum_{i=1}^{k}\frac{b-a}{k}f^{(4)}(\xi_i)\right).$$

Teraz wystarczy zauważyć, że suma po prawej stronie jest sumą Riemanna, a więc dąży do całki z $f^{(4)}$ gdy $k \to +\infty$.
□

(98) Przeprowadzając ortogonalizację Grama-Schmidta bazy potęgowej $\{1, x, x^2, x^3\}$ znajdź wielomiany ortogonalne Legendre'a stopnia $0, 1, 2, 3$, tzn. wielomiany ortogonalne na przedziale $[-1, 1]$ z wagą $\rho = 1$. Następnie wskaż zera tych wielomianów, czyli węzły odpowiednich kwadratur Gaussa-Legendre'a.

• Łatwo widać, że dwa pierwsze wielomiany są ortogonalne w $\mathcal{L}_2(-1,1)$, tzn. $L_0(x) = 1$ i $L_1(x) = x$. Przeprowadzając kolejno ortogonalizację wielomianów x^2 i x^3 ze względu na iloczyn skalarny $\langle f, g\rangle = \int_{-1}^{1} f(x)g(x)\,\mathrm{d}x$ i wykorzystując fakt, że

$$\int_{-1}^{1} x^i x^j\,\mathrm{d}x = \begin{cases} 0, & i+j \text{ - nieparzyste,} \\ \frac{2}{i+j+1}, & i+j \text{ - parzyste,} \end{cases}$$

dostajemy:

$$L_2(x) \;=\; x^2 - \frac{\langle x^2, x\rangle}{\langle x, x\rangle}\,x - \frac{\langle x^2, 1\rangle}{\langle 1, 1\rangle} \;=\; x^2 - \frac{1}{3},$$

$$L_3(x) \;=\; x^3 - \frac{\langle x^3, x^2 - \frac{1}{3}\rangle}{\langle x^2 - \frac{1}{3}, x^2 - \frac{1}{3}\rangle}(x^2 - \tfrac{1}{3}) - \frac{\langle x^3, x\rangle}{\langle x, x\rangle}\,x - \frac{\langle x^3, 1\rangle}{\langle 1, 1\rangle} \;=\; x^3 - \frac{3}{5}\,x.$$

Stąd odpowiednie pierwiastki wynoszą: $\{x_0^{(1)}\} = \{0\}$, $\{x_0^{(2)}, x_1^{(2)}\} = \{-\frac{1}{3}\sqrt{3},\, \frac{1}{3}\sqrt{3}\}$, $\{x_0^{(3)}, x_1^{(3)}, x_2^{(3)}\} = \{-\frac{1}{5}\sqrt{15},\, 0,\, \frac{1}{5}\sqrt{15}\}$.
□

(99) Załóżmy, że dane są liczby β_k i γ_k definiujące ciąg wielomianów ortogonalnych $\{p_k\}_{k\geqslant 0}$ przez formułę trójczłonową,

$$\begin{aligned} p_0(x) \;&=\; 1, \qquad p_1(x) \;=\; (x - \beta_1), \\ p_k(x) \;&=\; (x - \beta_k)p_{k-1}(x) \;-\; \gamma_k p_{k-2}(x), \qquad k \geqslant 2. \end{aligned}$$

Zaproponuj ekonomiczny (tzn. o koszcie proporcjonalnym do n) algorytm obliczania wartości w punkcie x wielomianu w danego poprzez jego współczynniki c_k rozwinięcia w bazie $\{p_k\}$, tzn. $w(x) = \sum_{k=0}^{n} c_k p_k(x)$.

• Niech $d_{n+2} = 0$, $d_{n+1} = 0$, oraz

$$d_k = c_k + (x - \beta_{k+1})d_{k+1} - \gamma_{k+2}d_{k+2} \quad \text{dla} \quad k = n, n-1, \ldots, 1, 0.$$

Wtedy, na podstawie formuły trójczłonowej, mamy

$$\begin{aligned} p(x) &= \sum_{k=0}^{n} c_k p_k(x) = \sum_{k=0}^{n} \Big(d_k - (x-\beta_{k+1})d_{k+1} + \gamma_{k+2} d_{k+2} \Big) p_k(x) \\ &= d_0 p_0(x) + d_1 \Big(p_1(x) - (x-\beta_1) p_0(x) \Big) \\ &\quad + \sum_{k=2}^{n} d_k \Big(p_k(x) - (x-\beta_k) p_{k-1}(x) + \gamma_k p_{k-2}(x) \Big) = d_0. \end{aligned}$$

Stąd algorytm:

$d_{n+2} := 0;\ d_{n+1} := 0;$
for $k := n$ **downto** 0 **do**
$\quad d_k := c_k + (x - \beta_{k+1})\, d_{k+1} - \gamma_{k+2} d_{k+2};$
$p := d_0 - d_1 x.$

□

(100) Uzasadnij, że: (a) jeśli kwadratura jest dokładna dla dowolnych $n+1$ wielomianów tworzących bazę w Π_n, to jest ona rzędu co najmniej $n+1$, oraz (b) jeśli kwadratura jest rzędu $n+1$ to jest ona niedokładna dla każdego wielomianu stopnia dokładnie $n+1$.

• Teza wynika z liniowości zarówno całki jak i kwadratury ze względu na f.

(a) Niech $p_0, p_1, \ldots, p_n$ będzie bazą w Π_n. Jeśli $\mathrm{rz}(Q)(p_k) = S_\rho(p_k)$ dla $0 \leqslant k \leqslant n$ to dla dowolnego wielomianu $w = \sum_{k=0}^{n} c_k p_k$ mamy

$$S_\rho(w) = \sum_{k=0}^{n} c_k S_\rho(p_k) = \sum_{k=0}^{n} c_k Q(p_k) = Q(w),$$

czyli $\mathrm{rz}(Q) \geqslant n+1$.

(b) Niech $w^*(x) = c^* x^{n+1} + \cdots$, gdzie $c \neq 0$, będzie takim wielomianem, że $S_\rho(w^*) \neq Q(w^*)$. Jeśli $w(x) = c x^{n+1} + \cdots$, gdzie $c \neq 0$, to $w = \frac{c}{c^*} w^* + v$, gdzie $v \in \Pi_{n-1}$ i

$$S_\rho(w) = \frac{c}{c^*} S_\rho(w^*) + S_\rho(v) = \frac{c}{c^*} S_\rho(w^*) + Q(v) \neq \frac{c}{c^*} Q(w^*) + Q(v) = Q(w).$$

□

(101) Znajdź węzeł $a \in [0,1)$ oraz współczynniki α i β tak, aby kwadratura

$$Q(f) = \alpha f(a) + \beta f(1)$$

przybliżająca całkę $\int_0^1 f(x)\,\mathrm{d}x$ miała największy rząd.

(102) Niech $d \in [a,b]$ będzie ustalone. Ile wynosi (w zależności od d) maksymalny rząd kwadratury postaci

$$Q(f) = \alpha f(d) + \gamma f(c), \quad \text{gdzie} \quad c \in [a,b],\ \alpha, \gamma \in \mathbb{R},$$

przybliżającej całkę $\int_a^b f(x)\,\mathrm{d}x$?

• Dla uproszczenia, możemy założyć, że $[a,b] = [-1,1]$. Ponieważ współczynniki α i γ są uwolnione, maksymalny rząd będzie miała kwadratura interpolacyjna oparta na węzłach d i c, przy czym rząd ten będzie wynosić na pewno co najmniej 2 i co najwyżej

4. Błąd takiej kwadratury to $\int_{-1}^{1}(x-c)(x-d)f[c,d,x]\,\mathrm{d}x$. Warunkiem koniecznym i wystarczającym na to, aby kwadratura była rzędu co najmniej 3 jest więc $|c| \leqslant 1$ oraz

$$0 \;=\; \int_{-1}^{1}(x-c)(x-d)\,\mathrm{d}x \;=\; \int_{-1}^{1} x^2 - x(c+d) + cd\,\mathrm{d}x \;=\; 2\Big(\tfrac{1}{3}+cd\Big),$$

czyli $c = \frac{-1}{3d}$ i $|d| \geqslant \frac{1}{3}$. Dalej już liczyć nie musimy, bo wiemy, że jedyną kwadraturą, która osiąga maksymalny rząd równy 4 jest kwadratura Gaussa-Legendre'a, którą dostajemy gdy $d = -\frac{1}{3}\sqrt{3}$ i $c = \frac{1}{3}\sqrt{3}$ (albo odwrotnie). Podsumowując mamy, że:
- dla $|d| < \frac{1}{3}$ maksymalny rząd wynosi 2 i jest przyjmowany przez kwadraturę interpolacyjną opartą na d i dowolnym $c \in [-1,1]$,
- dla $\frac{1}{3} \leqslant |d| \neq \frac{1}{3}\sqrt{3}$ maksymalny rząd wynosi 3 i jest przyjmowany przez kwadraturę interpolacyjną opartą na węzłach d i $c = \frac{-1}{3d}$,
- dla $|d| = \frac{1}{3}\sqrt{3}$ maksymalny rząd wynosi 4 i jest przyjmowany przez kwadraturę interpolacyjną (Gaussa-Legendre'a) opartą na d i $c = -d$.
□

(103) Ile wynosi maksymalny rząd kwadratury opartej na dwóch węzłach: jednokrotnym $x_0 = a$ i n-krotnym $x_1 \in [a,b]$, dla aproksymacji całki $S(f) = \int_a^b f(x)(x-a)\,\mathrm{d}x$? Czy kwadratura o maksymalnym rzędzie jest wyznaczona jednoznacznie?

• Maksymalny rząd osiąga kwadratura interpolacyjna i rząd ten wynosi co najmniej $n+1$, bo taka jest suma krotności węzłów. Jeśli n jest parzyste to dla dowolnego x_1 kwadratura nie jest dokładna dla wielomianu $(x-a)(x-x_1)^n$ stopnia $n+1$ (bo kwadratura zwraca zero, a całka jest dodatnia), czyli maksymalny rząd wynosi $n+1$ i jest osiągalny dla każdego węzła x_1.

Załóżmy, że n jest nieparzyste. Dla danego x_1, rozpatrzmy znów wielomian $w(x) = (x-a)(x-x_1)^n$. Jeśli $x_1 = a$ to $S(w) > 0$, a jeśli $x_1 = b$ to $S(w) < 0$. Z ciągłości S ze względu na x_1 mamy, że dla pewnego x_1^* jest $S(w) = 0$ i dla takiego x_1 kwadratura ma rząd co najmniej $n+2$. Jest to jednocześnie maksymalny rząd, bo kwadratura nie jest dokładna dla wielomianu $(x-a)(x-x_1)^{n+1}$ stopnia $n+2$.

W przypadku nieparzystego n, kwadratura o maksymalnym rzędzie jest wyznaczona jednoznacznie, bo x_1^* jest wtedy wyznaczony jednoznacznie. Wynika to z faktu, że $S(w)$ maleje gdy x_1 rośnie. Rzeczywiście, oznaczając $F(x,c) = \int_a^b (x-a)^2(x-c)^n \mathrm{d}x$ mamy

$$\frac{\partial F(x,c)}{\partial c} \;=\; -n\int_a^b (x-a)^2(x-c)^{n-1}\mathrm{d}x \;<\; 0,$$

bo, wobec parzystości $n-1$, funkcja pod całką jest dodatnia (poza $x=a$ i $x=c$.)
□

(104) Ile wynosi maksymalny rząd kwadratury opartej na czterech węzłach,

$$a = x_0 \leqslant x_1 \leqslant x_2 \leqslant x_3 = b,$$

dla aproksymacji całki $S(f) = \int_a^b f(x)\,\mathrm{d}x$?

• Możemy założyć bez straty ogólności, że $[a,b] = [-1,1]$. Dla dowolnych węzłów x_1 i x_2, niech $w^*(x) = (x^2-1)(x-x_1)^2(x-x_2)^2$. Wtedy całka z w^* jest ujemna, ale kwadratura zwraca zero. Stąd maksymalny rząd jest nie większy niż $\deg w^* = 6$.

Pokażemy, że maksymalny rząd wynosi rzeczywiście 6. W tym celu, rozpatrzmy kwadratury interpolacyjne Q_3^I oparte na węzłach ± 1 i $\pm a$, gdzie $0 \leqslant a \leqslant 1$. Każda taka kwadratura, jako interpolacyjna oparta na 4 węzłach, jest dokładna dla wielomianów z Π_3. Jest też dokładna dla wielomianu $v(x) = x(x^2-1)(x^2-a^2)$ stopnia 5, bo $S(v) = Q_3^I(v) = 0$. Pozostaje pokazać istnienie $a \in [0,1]$ takiego, że kwadratura jest dokładna dla jakiegoś wielomianu stopnia 4, np. dla $u(x) = (x^2-1)(x^2-a^2)$. Ponieważ $Q_3^I(u) = 0$, musi być $S(u) = 0$. Zauważmy, że dla $a = 1$ mamy $S(u) > 0$, a dla $a = 0$ mamy $S(u) < 0$. Ponieważ $S(u)$ zależy w sposób ciągły istnienie a takiego, że $S(u) = 0$ jest zapewnione.

Aby wyliczyć a (co formalnie nie jest już konieczne dla rozwiązania zadania), musimy rozwiązać równanie

$$0 \;=\; \int_{-1}^{1} (x^2-1)(x^2-a^2)\,\mathrm{d}x \;=\; \tfrac{4}{3}\Big(a^2 - \tfrac{1}{5}\Big),$$

z którego dostajemy $a = \frac{1}{5}\sqrt{5} = 0.4472...$

(Oczywiście, teza zadania wynika również z ogólniejszej tezy zadania (105).)

□

(105) Rozpatrzmy całkowanie z wagą $S_\rho(f) = \int_a^b f(x)\rho(x)\,\mathrm{d}x$, gdzie $-\infty < a < b < +\infty$. Ile wynosi maksymalny rząd kwadratury opartej na $n+1$ węzłach zawierających końce przedziału całkowania,

$$a = x_0 \leqslant x_1 \leqslant \cdots \leqslant x_{n-1} \leqslant x_n = b,$$

dla aproksymacji całki $S_\rho(f)$?

• Dla ustalonych węzłów x_i, największy rząd ma kwadratura interopolacyjna, nazwijmy ją Q_n^I, na tych węzłach oparta. Wtedy, dla wielomianu

$$w^*(x) = (x-a)(x-x_1)^2(x-x_2)^2\cdots(x-x_{n-1})^2(x-b),$$

mamy $Q_n^I(w^*) = 0$ i $S_\rho(w^*) < 0$, a ponieważ $w^* \in \Pi_{2n}$ to $\mathrm{rz}(Q_n^I) \leqslant 2n$.

Z drugiej strony, zdefiniujmy wagę

$$\rho_1(x) = -(x-a)(x-b)\rho(x)$$

i niech x_i^* dla $1 \leqslant i \leqslant n-1$ będą zerami $(n-1)$-szego wielomianu ortogonalnego w przestrzeni $\mathcal{L}_{2,\rho_1}(a,b)$. Niech $f \in \Pi_{2n-1}$ i niech $w_f \in \Pi_n$ będzie wielomianem interpolującym f w węzłach a, b i x_i^*, $1 \leqslant i \leqslant n-1$. Wtedy

$$f(x) - w_f(x) = (x-a)(x-x_1^*)\cdots(x-x_{n-1}^*)(x-b)g(x),$$

gdzie $g \in \Pi_{n-2}$. Stąd

$$\begin{aligned} S_\rho(f) - Q_n^I(f) \;&=\; \int_a^b (f-w_f)(x)\rho(x)\,\mathrm{d}x \\ &=\; -\int_a^b (x-x_1^*)\cdots(x-x_{n-1}^*)g(x)\rho_1(x)\,\mathrm{d}x \;=\; 0, \end{aligned}$$

bowiem wielomian $(x-x_1^*)\cdots(x-x_{n-1}^*)$ jest ortogonalny do g w $\mathcal{L}_{2,\rho_1}(a,b)$,

Pokazaliśmy więc, że maksymalny rząd wynosi $2n$ i jest przyjmowany przez kwadraturę interpolacyjną opartą na a, b i na zerach $(n-1)$-szego wielomianu ortogonalnego w przestrzeni $\mathcal{L}_{2,\rho_1}(a,b)$.

□

(106) Ile wynosi rząd kwadratury opartej na $n+1$ węzłach, przy czym każdy węzeł jest dwukrotny?

• Oczywiście, maksymalny rząd wynosi co najmniej $2n+2$ i jest przyjmowany przez kwadraturę interpolacyjną. Dla dwukrotnych węzłów $x_0 < x_1 < \cdots < x_n$ błąd takiej kwadratury dla funkcji $f \in C^{2n+2}([a,b])$ wynosi

$$\begin{aligned} S_\rho(f) - Q_n(f) &= \int_a^b (x-x_0)^2 \cdots (x-x_n)^2 \rho(x) f[x_0, x_0, x_1, x_1, \ldots, x_n, x_n, x] \,\mathrm{d}x \\ &= \frac{f^{(2n+2)}(\xi)}{(2n+2)!} \int_a^b (x-x_0)^2 \cdots (x-x_n)^2 \rho(x) \,\mathrm{d}x, \end{aligned}$$

jest więc dodatni dla $f(x) = x^{2n+2}$. Stąd, dla dowolnego wyboru $n+1$ dwukrotnych węzłów rząd kwadratury wynosi $2n+2$.
□

(107) Czy istnieje kwadratura postaci $Q(f) = a_0 f(x_0) + a_1 f(x_1)$ dla całki

$$f \mapsto \int_0^{+\infty} f(x) e^{-x^2/2} \,\mathrm{d}x,$$

która jest dokładna dla wszystkich wielomianów stopnia $\leqslant 3$? Jeśli tak to wskaż x_0, x_1 i a_0, a_1.

(108) Rozpatrzmy ciąg wielomianów ortogonalnych $\{p_n\}_{n=0}^{\infty}$ w przestrzeni $\mathcal{L}_{2,\rho}(-1,1)$. Wykaż, że jeśli waga ρ jest parzysta, tzn. $\rho(x) = \rho(-x)$, to p_{2n} są wielomianami parzystymi, a p_{2n+1} są wielomianami nieparzystymi dla $n = 0, 1, 2, \ldots$.

• Załóżmy bez straty ogólności, że $p_n(x) = x^n + \cdots$. Pokażemy, że wielomiany

$$q_n(x) = (-1)^n p_n(-x) = x^n + \cdots \qquad \text{dla} \quad n = 0, 1, 2, ...,$$

również tworzą ciąg wielomianów ortogonalnych w $\mathcal{L}_{2,\rho}(-1,1)$. Rzeczywiście, dla $i \neq j$ mamy

$$\begin{aligned} \langle q_i, q_j \rangle_{\mathcal{L}_{2,\rho}} &= \int_{-1}^{1} q_i(x) q_j(x) \rho(x) \mathrm{d}x = \int_{-1}^{1} (-1)^{i+j} p_i(-x) p_j(-x) \rho(-x) \,\mathrm{d}x \\ &= (-1)^{i+j+1} \int_{-1}^{1} p_i(x) p_j(x) \rho(x) \,\mathrm{d}x = 0. \end{aligned}$$

Z drugiej strony, z jednoznaczności ciągu wielomianów ortogonalnych mamy $q_n = p_n$, czyli $p_n(-x) = (-1)^n p_n(x)$ dla wszystkich n, co kończy dowód.
□

(109) Niech x_j dla $0 \leqslant j \leqslant n$ będą zerami $(n+1)$-szego wielomianu ortogonalnego w przestrzeni $\mathcal{L}_{2,\rho}(a,b)$. Niech dalej $l_j \in \Pi_n$ będą odpowiadającymi tym węzłom wielomianami Lagrange'a. Wykaż, że dla $f, g \in \Pi_n$ iloczyn skalarny w $\mathcal{L}_{2,\rho}(a,b)$ można wyrazić jako

$$\langle f, g \rangle_{\mathcal{L}_{2,\rho}} := \int_a^b f(x) g(x) \rho(x) \,\mathrm{d}x = \sum_{j=0}^{n} w_j f(x_j) g(x_j),$$

gdzie $w_j = \|l_j\|^2_{\mathcal{L}_{2,\rho}} = \langle l_j, l_j \rangle_{\mathcal{L}_{2,\rho}}$.

• Zapiszmy wielomiany Lagrange'a w postaci $l_j(x) = c_j \prod_{j\neq k=0}^{n}(x - x_k)$, gdzie $c_j^{-1} = \prod_{j\neq k=0}^{n}(x_j - x_k)$. Wtedy dla $i \neq j$ mamy

$$l_i(x)l_j(x) = c_i c_j \left(\prod_{k=0}^{n}(x - x_k)\right)\left(\prod_{i,j\neq s=0}^{n}(x - x_s)\right).$$

Pierwszy iloczyn po prawej stronie tej równości to, z dokładnością do czynnika liczbowego, $(n+1)$-szy wielomian ortogonalny, a drugi czynnik jest wielomianem stopnia $n-1$. Wielomiany te są więc prostopadłe w $\mathcal{L}_{2,\rho}$, a to znaczy, że $\langle l_i, l_j\rangle_{\mathcal{L}_{r,\rho}} = 0$. Rozwijając wielomiany $f, g \in \Pi_n$ w bazie Lagrange'a dostajemy

$$\begin{aligned}\langle f, g\rangle_{\mathcal{L}_{2,\rho}} &= \int_a^b \left(\sum_{i=0}^{n} f(x_i)l_i(x)\right)\left(\sum_{j=0}^{n} g(x_j)l_j(x)\right)\rho(x)\,\mathrm{d}x \\ &= \sum_{i,j=0}^{n} f(x_i)g(x_j)\int_a^b l_i(x)l_j(x)\rho(x)\,\mathrm{d}x = \sum_{i,j=0}^{n} f(x_i)g(x_j)\langle l_i, l_j\rangle_{\mathcal{L}_{2,\rho}} \\ &= \sum_{j=0}^{n} f(x_j)g(x_j)\langle l_j, l_j\rangle_{\mathcal{L}_{2,\rho}},\end{aligned}$$

co należało pokazać.
□

(110) Wykaż, że całka

$$\int_a^b (x - x_0)^2(x - x_1)^2 \cdots (x - x_n)^2\rho(x)\,\mathrm{d}x,$$

gdzie ρ jest pewną wagą, jest najmniejsza wtedy i tylko wtedy gdy za $x_0, \ldots, x_n$ weźmiemy zera $(n+1)$-szego wielomianu ortogonalnego w przestrzeni $\mathcal{L}_{2,\rho}(a, b)$.

• Niech

$$f(x) = (x - x_0)\cdots(x - x_n) \in \Pi_{n+1}$$

i niech $p_{n+1}(x) = (x - x_0^*)\cdots(x - x_n^*)$ będzie jednoznacznie wyznaczonym $(n+1)$-szym wielomianem ortogonalnym w przestrzeni $\mathcal{L}_{2,\rho}(a, b)$. Wtedy $f = p_{n+1} + v$, dla pewnego wielomianu $v \in \Pi_n$. Oznaczając przez $\langle\cdot,\cdot\rangle$ i $\|\cdot\|$ iloczyn skalarny i normę w $\mathcal{L}_{2,\rho}(a, b)$, mamy $\langle p_{n+1}, v\rangle = 0$ oraz

$$\begin{aligned}&\int_a^b (x - x_0)^2(x - x_1)^2 \cdots (x - x_n)^2\rho(x)\,\mathrm{d}x \\ &\quad = \|f\|^2 = \|p_{n+1} + v\|^2 = \|p_{n+1}\|^2 + \|v\|^2 \geqslant \|p_{n+1}\|^2 \\ &\quad = \int_a^b (x - x_0^*)^2(x - x_1^*)^2 \cdots (x - x_n^*)^2\rho(x)\,\mathrm{d}x.\end{aligned}$$

□

(111) Niech $f \in \Pi_{n+1}$ i niech w_f będzie wielomianem najlepiej aproksymującym f w ważonej normie średniokwadratowej $\mathcal{L}_{2,\rho}(a, b)$, wśród wielomianów $w \in \Pi_n$. Wykaż, że wtedy wielomian $f - w_f$ ma $n+1$ różnych zer. Wskaż te zera.

• Jeśli $f \in \Pi_n$ to $w_f = f$ i teza jest oczywista. Załóżmy więc, że f jest stopnia dokładnie $n+1$. Wtedy $f - w_f$ jest prostopadły w sensie $\mathcal{L}_{2,\rho}(a, b)$ do Π_n. Wobec

tego, że $f - w_f$ jest wielomianem stopnia dokładnie $n+1$, wielomian ten jest $(n+1)$-szym wielomianem ortogonalnym w $\mathcal{L}_{2,\rho}(a,b)$. Stąd w_f interpoluje f w węzłach-zerach wielomianu ortogonalnego stopnia $n+1$.
□

(112) Niech w_f będzie wielomianem w Π_n najlepiej aproksymującym funkcję $f \in C([a,b])$ w normie $\mathcal{L}_{2,\rho}(a,b)$. Wykaż, że wtedy różnica $f - w_f$ ma co najmniej $n+1$ różnych zer w (a,b). Innymi słowy, można dobrać punkty $a < x_0 < \cdots < x_n < b$ tak, że $w_f(x_i) = f(x_i)$, $0 \leqslant i \leqslant n$.

(113) Uzasadnij, że kwadratura prostokątów jest kwadraturą Gaussa-Legendre'a, natomiast żadna z kwadratur Newtona-Cotesa Q_n^{NC} dla $n \geqslant 1$ nie jest kwadraturą Gaussa dla jakiejkolwiek wagi ρ.

• Rząd kwadratury prostokątów wynosi 2 i jest taki sam jak dla kwadratury Gaussa-Legendre'a. Teza wynika więc z jednoznaczności kwadratur Gaussa. Z kolei, żadna kwadratura Newtona-Cotesa nie jest kwadrataturą Gaussa, bo te pierwsze korzystają z obu końcach przedziału całkowania, a końce te nie są zerami żadnego wielomianu ortogonalnego.
□

(114) Dla wielomianów Legendre'a L_n prawdziwa jest równość

$$\|L_n\|_{\mathcal{L}_2(-1,1)} = \int_{-1}^{1} L_n^2(x)\,\mathrm{d}x = \frac{2}{2n+1}.$$

Niech x_j dla $0 \leqslant j \leqslant n$ będą zerami $(n+1)$-szego wielomianu Legendre'a. Pokaż, że

$$\int_{-1}^{1} (x-x_0)^2(x-x_1)^2\cdots(x-x_n)^2\,\mathrm{d}x \;=\; \frac{2^{2n+3}}{2n+3}\Big(\frac{1\cdot 2\cdots n\cdot(n+1)}{(n+2)(n+3)\cdots(2n+2)}\Big)^2.$$

(115) Niech $Q(f) = \sum_{i=0}^{n} a_i f(u_i)$ będzie kwadraturą o rzędzie $\mathrm{rz}(Q) = r$ dla aproksymacji całki $S(f) = \int_0^1 f(u)\,\mathrm{d}u$. Niech K będzie jądrem Peano tej kwadratury, tzn.

$$S(f) - Q_n(f) = \int_0^1 K(t) f^{(r)}(t)\,\mathrm{d}t.$$

Wykaż, że kwadratura

$$Q_{a,b}(f) = (b-a)\sum_{i=0}^{n} a_i f(x_i), \qquad \text{gdzie} \quad x_i = a + u_i(b-a),$$

aproksymująca całkę $S_{a,b}(f) = \int_a^b f(x)\,\mathrm{d}x$, ma także rząd $\mathrm{rz}(Q_{a,b}) = r$, a jej jądro Peano wynosi

$$K_{a,b}(t) = (b-a)^r K\Big(\frac{t-a}{b-a}\Big).$$

• Niech $x = a + u(b-a)$, $u \in [0,1]$. Dla dowolnego wielomianu w niech

$$v(u) = w(a + u(b-a)) = w(x).$$

Wtedy v jest także wielomianem, przy czym $\deg v = \deg w$. Mamy też

$$S_{a,b}(w) \;=\; \int_a^b w(x)\,\mathrm{d}x \;=\; (b-a)\int_0^1 v(u)\,\mathrm{d}u \;=\; (b-a)S(v),$$

$$Q_{a,b}(w) = (b-a)\sum_{i=0}^{n} a_i w(x_i) \;=\; (b-a)\sum_{i=0}^{n} a_i v(u_i) \;=\; (b-a)Q(v),$$

a stąd $S_{a,b}(w) - Q_{a,b}(w) = (b-a)\Big(S(v) - Q(v)\Big)$. Ponieważ przyporządkowanie v do w jest wzajemnie jednoznaczne, $\mathrm{rz}(Q_{a,b}) = \mathrm{rz}(Q)$.

Jądro Peano dla $Q_{a,b}$ wynosi

$$K_{a,b}(t) \;=\; \int_a^b \frac{(x-t)_+^{r-1}}{(r-1)!}\mathrm{d}x - (b-a)\sum_{i=0}^{n} a_i \frac{(x_i-t)_+^{r-1}}{(r-1)!}.$$

Zamieniając zmienne na $x = a + u(b-a)$, $t = a + s(b-a)$, gdzie $s, u \in [0,1]$, mamy

$$\begin{aligned} K_{a,b}(t) &= (b-a)\int_0^1 \frac{(b-a)^{r-1}(u-s)_+^{r-1}}{(r-1)!}\mathrm{d}u - (b-a)\sum_{i=0}^{n} a_i \frac{(b-a)^{r-1}(u_i-s)_+^{r-1}}{(r-1)!} \\ &= (b-a)^r \left(\int_0^1 \frac{(u-s)_+^{r-1}}{(r-1)!}\mathrm{d}u - \sum_{i=0}^{n} a_i \frac{(u_i-s)_+^{r-1}}{(r-1)!}\right) \\ &= (b-a)^r K(s) \;=\; (b-a)^r K\left(\frac{t-a}{b-a}\right). \end{aligned}$$

□

(116) Niech $K_{n,s}$ będzie jądrem Peano kwadratury Q_n o rzędzie $\mathrm{rz}(Q_n) \geqslant s+1$, aproksymującej całkę $S_\rho(f) = \int_a^b f(x)\rho(x)\,\mathrm{d}x$, tzn.

$$S_\rho(f) - Q_n(f) = \int_a^b K_{n,s}(t) f^{(s)}(t)\,\mathrm{d}t, \qquad f \in C^s([a,b]).$$

Wykaż, że $\int_a^b K_{n,s}(t)\,\mathrm{d}t = 0$.

• Weźmy wielomian $w_s(x) = x^s/s!$, dla którego $w_s^{(s)} = 1$. Ponieważ kwadratura jest dokładna dla wielomianów stopnia s, mamy

$$0 \;=\; S_\rho(w_s) - Q_n(w_s) \;=\; \int_a^b K_{n,s}(t) w_s^{(s)}(t)\,\mathrm{d}t \;=\; \int_a^b K_{n,s}(t)\,\mathrm{d}t.$$

□

(117) Niech $K_{n,s}$ będzie jądrem Peano kwadratury Q_n o rzędzie $r = \mathrm{rz}(Q_n) \geqslant s+1$, aproksymującej całkę $S(f) = \int_a^b f(x)\,\mathrm{d}x$, tzn.

$$S_\rho(f) - Q_n(f) = \int_a^b K_{n,s}(t) f^{(s)}(t)\,\mathrm{d}t, \qquad f \in C^s([a,b]).$$

Wykaż, że funkcji $K_{n,s}$ jest prostopadła w $\mathcal{L}_2(a,b)$ do przestrzeni Π_{r-s-1}, tzn.

$$\int_a^b K_{n,s}(t) v(t)\,\mathrm{d}t = 0 \quad \text{dla wszystkich} \quad v \in \Pi_{r-s-1}.$$

(118) Wykaż, że jądra Peano kwadratur prostokątów R, trapezów T i parabol P dla aproksymacji całki $\int_0^1 f(x)\,\mathrm{d}x$ są symetryczne względem $\frac{1}{2}$ i dla $t \in \left[0, \frac{1}{2}\right]$ wynoszą odpowiednio

$$K_{0,2}^R(t) = \tfrac{1}{2}t^2, \qquad K_{1,2}^T(t) = -\tfrac{1}{2}t(1-t), \qquad K_{2,4}^P(t) = -\tfrac{1}{24}t^3\Big(\tfrac{2}{3} - t\Big).$$

Wywnioskuj stąd dokładne formuły na błędy tych kwadratur .

(119) Niech

$$Q_1^I(f) = \frac{b-a}{2}\left(f\Big(\frac{2a+b}{3}\Big) + f\Big(\frac{a+2b}{3}\Big)\right).$$

Pokaż równość

$$\sup_{f\in F_M^1([a,b])} |S(f) - Q_1^I(f)| \;=\; \frac{2}{72}M(b-a)^3 \qquad\qquad \Big(\frac{2}{72} = 0.027778\Big).$$

• Wystarczy rozpatrzyć $[a,b] = [-1,1]$ i pokazać, że błąd najgorszy wynosi $\frac{2}{9}M$. Dla $[-1,1]$ mamy $Q_1^I(f) = f\Big(\frac{-1}{3}\Big) + f\Big(\frac{1}{3}\Big)$. Kwadratura Q_1^I jest interpolacyjna i ma rząd 2, bo nie całkuje dokładnie wielomianu $x \mapsto x^2 - \frac{1}{9}$. Policzmy jądro Peano $K_{2,2}$ dla tej kwadratury. Mamy

$$K_{2,2}(t) = \int_{-1}^{1}(x-t)_+ \mathrm{d}x - \Big(\tfrac{1}{3}-t\Big)_+ - \Big(-\tfrac{1}{3}-t\Big)_+ = \begin{cases} \frac{1}{2}\Big(t+1\Big)^2, & -1 \leqslant t \leqslant -\frac{1}{3}, \\ \frac{1}{2}\Big(t^2+\frac{1}{3}\Big), & -\frac{1}{3} \leqslant t \leqslant \frac{1}{3}, \\ \frac{1}{2}\Big(t-1\Big)^2, & \frac{1}{3} \leqslant t \leqslant 1. \end{cases}$$

Ponieważ jądro jest nieujemne, dla $f \in C_M^2([-1,1])$ mamy z twierdzenia o wartości średniej, że

$$S(f) - Q_1^I(f) = \int_{-1}^{1} K_{2,2}(t) f^{(2)}(t)\,\mathrm{d}t = f^{(2)}(\xi)\int_{-1}^{1} K_{2,2}(t)\,\mathrm{d}t = \frac{2}{9}\,f^{(2)}(\xi).$$

Stąd błąd jest ograniczony z góry przez $\frac{2}{9}M$ i osiągany dla $f(x) = x^2$.
□

(120) Niech Q_1^I będzie kwadraturą interpolacyjną dla całki $S(f) = \int_a^b f(x)\,\mathrm{d}x$, opartą na odpowiednio przeskalowanych zerach wielomianu Czebyszewa U_2. Pokaż, że

$$\sup_{f\in C_M^2([a,b])} |S(f) - Q_1^I(f)| \;=\; \frac{1}{96}\,(b-a)^3 M \qquad\qquad \Big(\frac{1}{96} = 0.010870\Big).$$

• Wystarczy rozpatrzyć $[a,b] = [-1,1]$ i pokazać, że wtedy błąd najgorszy kwadratury wynosi $\frac{1}{12}M$. Zerami wielomianu U_2 są $\pm\frac{1}{2}$. Jądro Peano kwadratury wynosi

$$K_{2,2}(t) = \int_{-1}^{1}(x-t)_+ \mathrm{d}x - \Big(\tfrac{1}{2}-t\Big)_+ - \Big(-\tfrac{1}{2}-t\Big)_+ = \begin{cases} \frac{1}{2}\Big(t+1\Big)^2, & -1 \leqslant t \leqslant -\frac{1}{2}, \\ \frac{1}{2}\,t^2, & -\frac{1}{2} \leqslant t \leqslant \frac{1}{2}, \\ \frac{1}{2}\Big(t-1\Big)^2, & \frac{1}{2} \leqslant t \leqslant 1. \end{cases}$$

Ponieważ jądro jest nieujemne, dla $f \in C_M^2[-1,1])$ mamy z twierdzenia o wartości średniej, że

$$S(f) - Q_1^I(f) = \int_{-1}^{1} K_{2,2}(t) f^{(2)}(t)\,\mathrm{d}t = f^{(2)}(\xi)\int_{-1}^{1} K_{2,2}(t)\,\mathrm{d}t = \frac{1}{12}\,f^{(2)}(\xi).$$

Stąd błąd jest ograniczony z góry przez $\frac{1}{12}M$ i osiągany dla $f(x) = x^2$.
□

(121) Niech Q_a^I będzie kwadraturą interpolacyjną opartą na dwóch jednokrotnych węzłach $\pm a$, gdzie $0 < a \leqslant 1$, dla aproksymacji całki $S(f) = \int_{-1}^{1} f(x)\,\mathrm{d}x$. Wykaż, że maksymalny błąd w klasie $C_1^2([-1,1])$ wynosi

$$\sup_{f\in C_1^2([-1,1])} |S(f) - Q_a^I(f)| \;=\; \begin{cases} \left(\frac{1}{3} - a^2\right), & 0 < a \leqslant \frac{1}{2}, \\ \left(\frac{1}{3} - a^2\right) + \frac{4}{3}\left(2a-1\right)^{\frac{3}{2}}, & \frac{1}{2} < a \leqslant 1. \end{cases}$$

Czy maksymalny błąd jest najmniejszy gdy kwadratura oparta jest na zerach wielomianu Czebyszewa U_2 albo wielomianu Legendre'a P_2?

• Jądro Peano dla tej kwadratury wynosi

$$K_{2,2}(t) = \int_{-1}^{1} (x-t)_+ \mathrm{d}x - \left(a-t\right)_+ - \left(-a-t\right)_+ = \begin{cases} \frac{1}{2}\left(t+1\right)^2, & -1 \leqslant t \leqslant -a, \\ \frac{1}{2}\left(t^2+1\right) - a, & -a \leqslant t \leqslant a, \\ \frac{1}{2}\left(t-1\right)^2, & a \leqslant t \leqslant 1. \end{cases}$$

Wiemy z ogólnych rozważań, że

$$\sup_{f\in C_1^2([-1,1])} |S(f) - Q_a^I(f)| \;=\; \int_a^b |K_{2,2}(t)|\,\mathrm{d}t.$$

Aby dostać żądane formuły, wystarczy więc policzyć całkę z $|K_{2,2}|$.

Minimalizując maksymalny błąd po $a \in (0,1]$ dostajemy $a^* = 4 - 2\sqrt{3} \approx 0.5359$. Stąd kwadratura minimalizująca błąd najgorszy nie korzysta ani z zer wielomianu U_2 (które wynoszą ± 0.5), ani z zer wielomianu P_2 (które wynoszą $\pm\frac{1}{2}\sqrt{3} \approx 0.57735$). □

(122) Wykaż, że jeśli $|a| \leqslant \frac{1}{2}$ to

$$\sup\left\{\int_{-1}^{1} f(x)\,\mathrm{d}x :\; f \in C_1^2([-1,1]),\, f(\pm a) = 0\right\} \;=\; \int_{-1}^{1} \frac{x^2 - a^2}{2}\,\mathrm{d}x \;=\; \frac{1}{3} - a^2.$$

• Całkę $\int_{-1}^{1} f(x)\,\mathrm{d}x$ dla f zerującej się w $\pm a$ interpretujemy jako błąd kwadratury interpolacyjnej opartej na dwóch węzłach $\pm a$, dla aproksymacji tej właśnie całki. W tym przypadku jądro Peano wynosi

$$K_{2,2}(t) = \int_{-1}^{1} (x-t)_+ \mathrm{d}x - \left(a-t\right)_+ - \left(-a-t\right)_+ = \begin{cases} \frac{1}{2}\left(t+1\right)^2, & -1 \leqslant t \leqslant -a, \\ \frac{1}{2}\left(t^2+1\right) - a, & -a \leqslant t \leqslant a, \\ \frac{1}{2}\left(t-1\right)^2, & a \leqslant t \leqslant 1. \end{cases}$$

Jądro jest węc nieujemne i dlatego, z twierdzenia o wartości średniej,

$$\int_{-1}^{1} f(x)\,\mathrm{d}x \;=\; \int_{-1}^{1} K_{2,2}(t) f^{(2)}(t)\,\mathrm{d}t \;=\; f^{(2)}(\xi) \int_{-1}^{1} K_{2,2}(t)\,\mathrm{d}t \;\leqslant\; \left(\frac{1}{3} - a^2\right) M.$$

Równość jest osiągana dla $f(x) = \frac{1}{2}(x^2 - a^2)$.
□

(123) Wykaż, że w klasie funkcji $f \in C^1([-1,1])$ takich, że $f\left(\pm\frac{3}{4}\right) = 0$ i f' spełnia warunek Lipschitza ze stałą 1, całka $\int_{-1}^{1} f(x)\,\mathrm{d}x$ jest maksymalizowana przez funkcję

$$f^*(x) = \begin{cases} \frac{1}{2}x^2 + \sqrt{2}x + \left(\frac{3\sqrt{2}}{4} - \frac{9}{32}\right) & -1 \leqslant x \leqslant -\frac{1}{2}\sqrt{2}, \\ -\frac{1}{2}x^2 + \frac{9}{32}, & -\frac{1}{2}\sqrt{2} \leqslant x \leqslant \frac{1}{2}\sqrt{2}, \\ \frac{1}{2}x^2 - \sqrt{2}x + \left(\frac{3\sqrt{2}}{4} - \frac{9}{32}\right), & \frac{1}{2}\sqrt{2} \leqslant x \leqslant 1. \end{cases}$$

(124) Niech Q będzie kwadraturą o najwyższym rzędzie spośród kwadratur opartych na dwóch węzłach: 0 i $a \in [0,1]$, dla aproksymacji całki $S(f) = \int_0^1 f(x)\,\mathrm{d}x$. Wyznacz maksymalny błąd kwadratury,

$$\sup_{f \in C_1^2([0,1])} |S(f) - Q(f)|,$$

gdzie $C_1^2([0,1])$ jest klasą funkcji dwa razy różniczkowalnych w sposób ciągły na $[0,1]$ i takich, że $\|f''\|_{C([0,1])} \leqslant 1$.

• Kwadraturą Q o najwyższym rzędzie, równym 3, jest kwadratura interpolacyjna oparta na węzłach 0 i $\frac{2}{3}$, cf. zadanie (102). Wyznaczmy odpowiednie jądro Peano $K_{2,2}$. Mamy

$$K_{2,2}(t) \;=\; \int_0^1 (x-t)_+ \mathrm{d}x \;-\; \int_0^1 v_t(x)\,\mathrm{d}x,$$

gdzie $v_t \in \Pi_1$ interpoluje funkcję $x \mapsto (x-t)_+$ w punktach 0 i $\frac{2}{3}$. Stąd

$$K_{2,2}(t) \;=\; \tfrac{1}{2}(1-t)^2 - \tfrac{1}{2}\left(1 - \tfrac{3}{2}t\right)_+ \;=\; \begin{cases} \frac{1}{2}t\left(t - \frac{1}{2}\right), & 0 \leqslant t \leqslant \frac{2}{3}, \\ \frac{1}{2}\left(1-t\right)^2, & \frac{2}{3} < t \leqslant 1. \end{cases}$$

(Zauważmy, że $\int_0^1 K_{2,2}(t)\mathrm{d}t = 0$, bo rząd kwadratury jest większy niż 2, cf. zadanie (116).) Wiemy, że

$$\begin{aligned} \sup_{f \in C_1^2([0,1])} |S(f) - Q(f)| \;&=\; \int_0^1 |K_{2,2}(t)|\,\mathrm{d}t \\ &= \left(-\int_0^{\frac{1}{2}} + \int_{\frac{1}{2}}^{\frac{2}{3}}\right) \tfrac{1}{2}t\left(t - \tfrac{1}{2}\right)\mathrm{d}t \;+\; \int_0^1 \tfrac{1}{2}\left(1-t\right)^2 \mathrm{d}t \;=\; \frac{1}{48} \;=\; 0.0208\ldots. \end{aligned}$$

□

(125) Całkę $S(f) = \int_a^b f(x)\,\mathrm{d}x$ przybliżamy złożoną kwadraturą Simpsona (parabol) $\overline{P}_k(f)$ z równomiernym podziałem przedziału $[a,b]$ na k podprzedziałów. Wiadomo, że jeśli $f \in C^4([a,b])$ to błąd $S(f) - P_k(f)$ zbiega do zera co najmniej tak szybko jak k^{-4}, gdy $k \to +\infty$. Co więcej, dla niektórych funkcji, takich jak $f(x) = x^4$, błąd ten zbiega dokładnie jak k^{-4}. Pokaż, że jeśli f ma mniejszą regularność, np. $f \in C^3([a,b])$, to

$$\lim_{k \to +\infty} \left(S(f) - \overline{P}_k(f)\right) k^3 \;=\; 0,$$

czyli błąd zbiega do zera szybciej niż k^{-3}.

(126) Niech F będzie klasą funkcji $f : [0,1] \to \mathbb{R}$ takich, że $f(0) = f(1)$ oraz f jest lipschitzowska ze stałą 1. Całkę $S(f) = \int_0^1 f(x)\,\mathrm{d}x$ dla $f \in F$ aproksymujemy algorytmem używającym jedynie wartości f w n punktach, tzn. przy pomocy formuły

$$A_n(f) \;=\; \Phi\Big(f(x_1), f(x_2), \ldots, f(x_n)\Big),$$

gdzie $0 \leqslant x_1 \leqslant x_2 \leqslant \cdots \leqslant x_n \leqslant 1$, a $\Phi : \mathbb{R}^n \to \mathbb{R}$ jest dowolnym przekształceniem. Jak dobrać punkty x_i oraz przekształcenie Φ aby zminimalizować błąd

$$E_n(\Phi, x_1, \ldots, x_n) \;=\; \sup_{f\in F} \Big|S(f) - \Phi\Big(f(x_1), f(x_2), \ldots, f(x_n)\Big)\Big|,$$

czyli błąd najgorszy aproksymacji całki w klasie F?

• Ponieważ funkcje f są periodyczne, możemy założyć, że

$$x_0 := 0 < x_1 < x_2 < \cdots < x_n = 1.$$

Oczywiście, dla każdego i funkcja f spełnia $|f(x) - f(x_i)| \leqslant |x - x_i|$. Oznaczając

$$f_-(x) = \max_{0\leqslant i\leqslant n} f(x_i) - |x - x_i|, \qquad f_+(x) = \min_{0\leqslant i\leqslant n} f(x_i) + |x - x_i|,$$

mamy $f_-, f_+ \in F$ oraz $f_- \leqslant f \leqslant f_+$. Stąd, dla danej informacji $\vec{y} = (f(x_1), \ldots, f(x_n))$, optymalne Φ "zwraca" $\Phi(\vec{y}) = S\Big(\frac{1}{2}(f_- + f_+)\Big)$, a błąd maksymalny dla danej $\vec{y}$ wynosi $S\Big(\frac{1}{2}(f_+ - f_-)\Big)$. Pokażemy, że

$$\min_{\Phi} E_n(\Phi, x_1, \ldots, x_n) \;=\; \sup\left\{ \int_0^1 h(t)\,\mathrm{d}t :\; h \in F,\; h(x_i) = 0, 1 \leqslant i \leqslant n \right\}$$

Rzeczywiście, nierówność '$\leqslant$' wynika z faktu, że dla $f^* = \frac{1}{2}(f_+ - f_-)$ mamy $f^* \in F$ i $f^*(x_i) = 0$ dla wszystkich i. A z drugiej strony, 'sup' po prawej stronie jest osiągane przez funkcję f_+ dla informacji zerowej.

Z powyższej analizy wynika, że

$$\min_{\Phi} E_n(\Phi, x_1, \ldots, x_n) \;=\; \frac{1}{4}\sum_{i=1}^{n} (x_i - x_{i-1})^2.$$

Minimalizując prawą stronę dostajemy, że optymalne punkty wynoszą

$$x_i^* = \frac{i}{n}, \quad 1 \leqslant i \leqslant n,$$

a dla nich minimalny błąd jest równy $\frac{1}{4n}$. Pozostaje zauważyć, że dla optymalnych punktów najlepsze Φ wyraża się prostym wzorem

$$\Phi^*\Big(f(x_1), f(x_2), \ldots, f(x_n)\Big) \;=\; \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

□

(127) Wykaż, że metoda bisekcji jest optymalna wśród algorytmów ϕ korzystających z (w ogólności wybranych adaptacyjnie) n wartości funkcji, w klasie F funkcji rosnących $f \in C([0,1])$ i takich, że $f(0) < 0 < f(1)$. Dokładniej, dla dowolnego takiego algorytmu ϕ_n błąd najgorszy

$$\sup_{f\in F} |x^*(f) - \phi_n(f)| \geqslant 2^{-(n+1)}.$$

(Tutaj $x^*(f)$ jest zerem funkcji f.)

• Niech ϕ_n będzie danym algorytmem korzystającym z n wartości funkcji w punktach $x_i \in (0,1)$ dla $i = 1, 2, \ldots, n$, przy czym x_i wybierane są zależnie od obliczonych wcześniej wartości $f(x_1), f(x_2), \ldots, f(x_{i-1})$. Rozpatrzmy podklasę $F_0 \subset F$ funkcji f dla których $f(x_i) = \pm 1$ dla wszystkich i. (Oczywiście, F_0 nie jest pusta.) Ponieważ dla dowolnego $x \in (a,b)$ odwzorowanie $F_0 \ni f \mapsto f(x)$ przybiera tylko dwie wartości, zbiór Z wszystkich możliwych punktów, których używa ϕ_n dla klasy F_0 zawiera co najwyżej $2^0 + 2^1 + \cdots + 2^{n-1} = 2^n - 1$ elementów. Stąd istnieje przedział otwarty (c,d) długości $|d-c| \geqslant 2^{-n}$, który nie zawiera żadnego punktu ze zbioru Z. Dla dowolnej $0 < \delta < 2^{-(n+1)}$, skonstruujmy dwie funkcje $f_-, f_+ \in F_0$ takie, że obie przyjmują -1 dla argumentów $x \leqslant c$ i przyjmują $+1$ dla argumentów $x \geqslant d$, ale $x^*(f_-) = c + \delta$ i $x^*(f_+) = d - \delta$. Wtedy $\phi_n(f_-) = \phi_n(f_+) := z$, bo obie funkcje są nierozróżnialne ze względu na informację, a stąd

$$\sup_{f \in F} |x^*(f) - \phi_n(f)| \geqslant \max\{|c + \delta - z|, |d - \delta - z|\} \geqslant \tfrac{1}{2}(d-c) - \delta \geqslant 2^{-(n+1)} - \delta.$$

Wobec tego, że δ może być dowolnie mała, dostajemy tezę.
□

(128) Rozważmy problem z zadania (127), ale ograniczmy się do algorytmów korzystających z wartości f w n ustalonych z góry punktach, czyli ograniczmy się do algorytmów nieadaptacyjnych. Wykaż, że wtedy

$$\inf_{\phi} \sup_{f \in F} |x^*(f) - \phi(f)| = \frac{1}{2(n+1)}.$$

(129) Uzasadnij, że metoda iteracji prostych $x_k = \cos(x_{k-1})$ jest zbieżna do jedynego rozwiązania równnania $x = \cos(x)$, dla dowolnego przybliżenia początkowego $x_0 \in \mathbb{R}$.

(130) Niech x^* będzie punktem stałym odwzorowania $\phi : \mathbb{R} \to \mathbb{R}$, przy czym ϕ jest klasy C^1 w pewnym nietrywialnym otoczeniu x^* oraz $|\phi'(x^*)| < 1$. Wykaż, że wtedy istnieje $\delta > 0$ taka, że dla każdego przybliżenia początkowego x_0 spełniającego $|x_0 - x^*| \leqslant \delta$, iteracje proste $x_k = \phi(x_{k-1})$, $k = 1, 2, \ldots$, zbiegają do x^*. Kiedy ciąg kolejnych przybliżeń x_k zbiega do x^* monotonicznie?

• Wybierzmy $\delta > 0$ tak, aby ϕ była klasy C^1 oraz $|\phi'(x)| \leqslant L < 1$ w przedziale $D := [x^* - \delta, x^* + \delta]$, dla pewnego L. Wtedy $\phi(D) \subset D$ oraz ϕ jest zwężająca w D. Rzeczywiście, jeśli $x \in D$, tzn. $|x - x^*| \leqslant \delta$, to

$$|\phi(x) - x^*| = |\phi(x) - \phi(x^*)| \leqslant L|x - x^*| \leqslant L\delta < \delta,$$

czyli $\phi(x) \in D$. Mamy również, że jeśli $x, y \in D$ to z twierdzenia Rolle'a

$$|\phi(x) - \phi(y)| = |\phi(\xi_k)(x - y)| \leqslant L|x - y|, \qquad \text{gdzie} \quad \xi_k \in D.$$

Zbieżność wynika teraz z twierdzenie Banacha o odwzorowaniach zwężających.

Aby odpowiedzieć na pytanie o monotoniczność, zauważmy, że

$$x_k - x^* = \phi(x_{k-1}) - \phi(x^*) = (x_{k-1} - x^*)\phi'(\xi_k).$$

Stąd, jeśli $x_{k-1} \in D$ i $\phi'(x^*) > 0$ to zbieżność jest monotoniczna; rosnąco gdy $x_0 < x^*$ i malejąco gdy $x_0 > x^*$. Jeśli $\phi(x^*) < 0$ to x_k znajduje się naprzemiennie po lewej i po prawej stronie x^*. Jeśli zaś $\phi'(x^*) = 0$ to charakter zbieżności zależy od dalszych

własności funkcji ϕ w okolicy x^*.
□

(131) Jeśli w danym modelu obliczeniowym liczenie pierwiastka kwadratowego $\sqrt{a}$ dla $a > 0$ nie jest dopuszczalne, to możemy go przybliżyć stosując wzór rekurencyjny

$$x_k \;=\; \frac{1}{2}\left(x_{k-1} + \frac{a}{x_{k-1}}\right),$$

który powstaje przez zastosowanie metody Newtona do równania $f(x) = x^2 - a = 0$. Pokaż, że ciąg x_k jest zbieżny do $\sqrt{a}$ dla dowolnego przybliżenia początkowego $x_0 > 0$. Scharakteryzuj szybkość zbieżności.

• W tym przypadku, błąd metody Newtona wynosi

$$\begin{aligned} x_k - \sqrt{a} \;&=\; x_{k-1} - \sqrt{a} - \frac{x_{k-1}^2 - a}{2x_{k-1}} \;=\; (x_{k-1} - \sqrt{a})\left(1 - \frac{x_{k-1} + \sqrt{a}}{2x_{k-1}}\right) \\ &=\; (x_{k-1} - \sqrt{a})\left(\frac{x_{k-1} - \sqrt{a}}{2x_{k-1}}\right). \end{aligned}$$

Jeśli teraz $0 < x_0 < \sqrt{a}$ to $(x_1 - \sqrt{a}) = \frac{(x_0 - \sqrt{a})^2}{2x_0} > 0$, czyli $x_1 > \sqrt{a}$. Jeśli zaś $x_0 > \sqrt{a}$, czy ogólniej $x_{k-1} > \sqrt{a}$, to $\frac{x_{k-1} - \sqrt{a}}{2x_{k-1}} < \frac{1}{2}$, czyli

$$0 \;<\; x_k - \sqrt{a} \;\leqslant\; \frac{1}{2}(x_{k-1} - \sqrt{a}).$$

Stąd mamy zbieżność do $\sqrt{a}$ dla dowolnego przybliżenia początkowego x_0.

Zauważmy jeszcze, że jeśli $x_{k-1} \gg \sqrt{a}$ to $\frac{x_{k-1} - \sqrt{a}}{2x_{k-1}} \approx \frac{1}{2}$, czyli początkowo zbieżność ma charakter liniowy, a dla $x_{k-1} \approx \sqrt{a}$ mamy

$$x_{k-1} - \sqrt{a} \;\approx\; \frac{(x_{k-1} - \sqrt{a})^2}{2\sqrt{a}},$$

czyli mamy zbieżność kwadratową, typową dla metody Newtona.
□

(132) Zaproponuj metodę iteracyjną obliczania $1/a$ dla dowolnego $a > 0$ nie używającą dzielenia. Jak wybrać przybliżenie początkowe, aby metoda była zbieżna? Jaki jest wykładnik zbieżności?

• Zastosujmy metodę Newtona do równania $f(x) := \frac{1}{x} - a = 0$. Dostajemy wzór rekurencyjny

$$x_k \;=\; x_{k-1} - \frac{1/x - a}{-1/x^2} \;=\; x_{k-1}\,(2 - a x_{k-1}),$$

który nie używa dzielenia. Niech $\phi(x) = x(2 - ax)$. Oczywiście $\phi(\frac{1}{a}) = \frac{1}{a}$. Ponieważ $\phi'(\frac{1}{a}) = 0$ i $\phi''(\frac{1}{a}) = -2a \neq 0$, metoda jest zbieżna (lokalnie) z wykładnikiem 2.

Teraz wyznaczymy obszar zbieżności. Bezpośredni rachunek (albo interpretacja geometryczna) pokazuje, że jeśli $0 < x_0 < \frac{1}{a}$ to $x < \phi(x) < \frac{1}{a}$. To zaś oznacza, że jeśli wystartujemy z $x_0 \in (0, \frac{1}{a})$ to ciąg $\{x_k\}$ kolejnych przybliżeń będzie zbiegał rosnąco do rozwiązania $\frac{1}{a}$. Z drugiej strony, jeśli $x = 0$ to $\phi(x) = 0$, a jeśli $x < 0$ to $\phi(x) < x$, co oznacza, że nie mamy zbieżności do rozwiązania gdy $x_0 \leqslant 0$.

Wykorzystując fakt, że ϕ jest symetryczna względem $\frac{1}{a}$ dostajemy ostatecznie, że zbieżność do rozwiązania $\frac{1}{a}$ zachodzi wtedy i tylko wtedy gdy punkt startowy $x_0 \in (0, \frac{2}{a})$. (Zauważmy, że mamy zbieżność w przedziale $(0, \frac{1}{2a})$, pomimo, że tam $\phi'(x) > 1$, a więc nie jest spełniony warunek dostateczny zbieżności.)
□

(133) Jeśli f' istnieje, jest ciągła i dodatnia w $[a, b]$ oraz $f(a)f(b) < 0$ to istnieje dokładnie jedno zero w (a, b). Wykaż, że to zero można otrzymać w granicy stosując metodę iteracji prostych $x_{n+1} = \phi(x_n)$ do funkcji $\phi(x) = x + \lambda f(x)$, dla pewnej wartości parametru λ.

• Przy podanych założeniach, istnieją $0 < c_0 \leqslant c_1 < +\infty$ takie, że

$$0 \;<\; c_0 \;\leqslant\; f'(x) \;\leqslant\; c_1 \quad \text{dla} \quad a \leqslant x \leqslant b.$$

Weźmy

$$\tfrac{-2}{c_1} \;<\lambda\; <\; 0.$$

Wtedy, z jednej strony mamy

$$\phi'(x) \;=\; 1 + \lambda f'(x) \;\leqslant\; 1 + \lambda c_0 \;<\; 1,$$

a z drugiej

$$\phi'(x) \;\geqslant\; 1 + \lambda c_1 \;>\; 1 + \left(\tfrac{-2}{c_1}\right) c_1 \;=\; -1.$$

Stąd, dla $c = \max\{|1+\lambda c_0|, |1+\lambda c_1|\}$ jest $|\phi'(x)| \leqslant c < 1$, co wystarcza do zbieżności iteracji prostych przy dowolnym przybliżeniu początkowym $x_0 \in [a, b]$, cf. zadanie (130).
□

(134) Rozpatrzmy metodę iteracyjną

$$x_k \;=\; x_{k-1} \;-\; \frac{x_{k-1} - a}{f(x_{k-1}) - f(a)}\, f(x_{k-1})$$

dla funkcji f, które są klasy C^1 w pewnym otoczeniu jej zera x^* i takich, że $f'(x^*) \neq 0$. Pokaż, że metoda ta jest lokalnie zbieżna liniowo o ile a jest dostatecznie blisko x^*. Jak blisko?

• Potraktujemy metodę jako iteracje proste zastosowane do równania $\phi(x) = x$, gdzie

$$\phi(x) = x - \frac{x - a}{f(x) - f(a)}\, f(x).$$

Mamy

$$\phi(x) \;=\; \frac{af(x) - xf(a)}{f(x) - f(a)}$$

i

$$\phi'(x) \;=\; \frac{\Big(af'(x) - f(a)\Big)\Big(f(x) - f(a)\Big) - f'(x)\Big(af(x) - xf(a)\Big)}{\Big(f(x) - f(a)\Big)^2},$$

a stąd

$$\phi'(x^*) \;=\; \frac{-\Big(af'(x^*) - f(a)\Big)f(a) + x^* f(a) f'(x^*)}{f^2(a)} \;=\; 1 - \frac{f'(x^*)}{f(a)}\,(a - x^*).$$

Zbieżność lokalną (co najmniej) liniową mamy wtedy gdy $|\phi(x^*)| < 1$, czyli

$$0 \;<\; \frac{f'(x^*)}{\left(\frac{f(a)-f(x^*)}{a-x^*}\right)} \;<\; 2.$$

To zaś zachodzi dla a dostatecznie bliskich x^*, bo wyrażenie w mianowniku dąży do $f'(x^*)$ gdy $a \to x^*$.
□

(135) Pokaż, że *metoda Steffensena*,

$$x_k \;=\; x_{k-1} \;-\; \frac{f^2(x_{k-1})}{f\Big(x_{k-1} + f(x_{k-1})\Big) - f(x_{k-1})},$$

jest zbieżna lokalnie z wykładnikiem 2, w klasie funkcji f dwukrotnie różniczkowalnych w sposób ciągły i takich, że $f'(x^*) \neq 0$.

• Oznaczmy $e_k = x_k - x^*$, gdzie $f(x^*) = 0$. Rozwijając kilka razy w szereg Taylora funkcje f i f' dostajemy

$$\begin{aligned}
e_k &= e_{k-1} - \frac{f(x_{k-1})}{\frac{f(x_{k-1}+f(x_{k-1}))-f(x_{k-1})}{f(x_{k-1})}} = e_{k-1} - \frac{e_{k-1}f'(x^*) + \frac{1}{2}e_{k-1}^2 f''(\alpha_k)}{f'(x_{k-1}) + \frac{1}{2}f(x_{k-1})f''(\beta_k)} \\
&= e_{k-1}\left(1 - \frac{f'(x^*) + \frac{1}{2}e_{k-1}f''(\alpha_k)}{f'(x^*) + e_{k-1}f''(\gamma_k) + \frac{1}{2}f''(\beta_k)(e_{k-1}f'(x^*) + \frac{1}{2}e_{k-1}^2 f''(\delta_k))}\right) \\
&= e_{k-1}^2\left(\frac{f''(\gamma_k) + \frac{1}{2}f''(\beta_k)(f'(x^*) + \frac{1}{2}e_{k-1}f''(\delta_k)) - \frac{1}{2}f''(\alpha_k)}{f'(x^*) + e_{k-1}f''(\gamma_k) + \frac{1}{2}f''(\beta_k)(e_{k-1}f'(x^*) + \frac{1}{2}e_{k-1}^2 f''(\delta_k))}\right) \\
&\approx e_{k-1}^2 \frac{f''(x^*)(1 + f'(x^*))}{2f'(x^*)},
\end{aligned}$$

gdzie asymptotyczna równość zachodzi gdy x_k dąży do x^*.
□

(136) Rozpatrzmy metodę iteracyjną daną wzorem

$$x_k \;=\; x_{k-1} \;-\; m\,\frac{f(x_{k-1})}{f'(x_{k-1})}.$$

Pokaż, że jeśli f ma m-krotne zero x^* to metoda ta jest lokalnie zbieżna do x^* z wykładnikiem 2.

• Mamy

$$\begin{aligned} e_k &= e_{k-1} - m \frac{e_{k-1}^m \frac{f^{(m)}(x^*)}{m!} + e_{k-1}^{m+1} \frac{f^{(m+1)}(\eta_k^{(1)})}{(m+1)!}}{e_{k-1}^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(2)})}{m!}} \\ &= e_{k-1} \left(1 - \frac{e_{k-1}^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(1)})}{(m+1)(m-1)!}}{e_{k-1}^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(2)})}{m!}} \right) \\ &= e_{k-1} \left(\frac{e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(2)})}{m!} - e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(1)})}{(m+1)(m-1)!}}{e_{k-1}^{m-1} \frac{f^{(m)}(x^*)}{(m-1)!} + e_{k-1}^m \frac{f^{(m+1)}(\eta_k^{(2)})}{m!}} \right) \\ &\approx e_{k-1}^2 \frac{f^{(m+1)}(x^*)}{m(m+1)f^{(m)}(x^*)}. \end{aligned}$$

□

(137) Do znalezienia miejsca zerowego funkcji $f(x) = (x-1)e^x$ zastosowano metodę Newtona. Dla jakich punktów początkowych metoda będzie zbieżna do $x^* = 1$? Czy zbieżność będzie kwadratowa dla dostatecznie bliskiego przybliżenia początkowego $x_0 \neq x^*$?