Paweł Omietański

Wstęp do metod numerycznych

Notatki z wykładu prof. Sędziwego.

Spis treści

Metody numeryczne zajmują się badaniem sposobów rozwiązywania zadań matematycznych przy pomocy działań arytmetycznych.

1 Analiza błędów.

1.1 Reprezentacja liczb rzeczywistych.

Każdą liczbę rzeczywistą $x \in \mathbb{R}$ możemy zapisać w postaci

$$x = \pm \sum_{i=k}^{-\infty} c_i \beta^i = \pm(\beta^k c_k + \beta^{k-1} c_{k-1} + \ldots + c_0 + \beta^{-1} c_{-1} + \ldots),$$

Jedna z kropek to kropka dziesiętna!

gdzie β jest podstawą systemu, oraz $c_i \in \{0, \ldots, \beta - 1\}$. Możemy więc zapisać, że

$$x = \pm c_k c_{k-1} \ldots c_0 . c_{-1} c_{-2} \ldots$$

Ustalmy n i niech $\tilde{x}$ oznacza reprezentację liczby x za pomocą n znaków. Wówczas **błędem bezwzględnym** wartości przybliżanej $\tilde{x}$ nazywamy

$$\Delta x = \tilde{x} - x,$$

natomiast **błędem względnym** tej wartości nazywamy

$$\varepsilon = \frac{\Delta x}{x}.$$

1.1.1 Reprezentacja stałoprzecinkowa.

Liczbę $\tilde{x}$ zapisujemy przy pomocy $n = n_1 + n_2$ znaków, gdzie n_1 oznacza liczbę znaków przed kropką dziesiętną, a n_2 za tą kropką. Jeżeli $n_2 = 0$, to mamy liczbę całkowitą. Liczby całkowite o długości n należą do przedziału $[-\beta^n + 1, \beta^n - 1]$.

Jeśli $a, b \in \mathbb{Z}$ są o długości nie większej niż n takimi, że $a \pm b$, ab są liczbami o długości nie większymi niż n, to operacje dodawania, odejmowania i mnożenia są wykonywane dokładnie.

1.1.2 Reprezentacja zmiennoprzecinkowa.

W reprezentacji tej, liczbę rzeczywistą x zapisujemy w postaci

$$x = s\beta^c \cdot m,$$

gdzie s - znak liczby, β - podstawa systemu, c - cecha, m - mantysa.

Ponieważ bez żadnych dodatkowych założeń każdą liczbę można by było zapisać na nieskończenie wiele sposobów, wprowadza się tzw. **warunek normalizacji**: $m \in [\beta^{-1}, 1)$. Zatem mantysę można zapisać jako szereg

$$m = \sum_{i=1}^{\infty} \alpha_i \beta^{-i} = 0.\alpha_1 \alpha_2 \ldots,$$

gdzie $\alpha_i \in \{0, \ldots, \beta - 1\}$, $\alpha_1 \neq 0$. Liczby α_i nazywamy **liczbami znaczącymi**.

Ustalmy t – liczbę znaków mantysy, oraz $n - t$ – liczbę znaków cechy.

Definicja 1.1. *(Zbiór liczb maszynowych)*

$$A = \{x \in \mathbb{R} |\ x = s\beta^c m_t,\ c \text{ } ma \text{ } n-t \text{ } znaków,\ m_t \text{ } ma \text{ } t \text{ } znaków\}$$

Ponieważ zbiór A jest skończony, możemy więc zdefiniować **odwzorowanie zaokrąglania**

$$\tilde{rd} : \mathbb{R} \ni x = s\beta^c m \mapsto s\beta^c m_t \in A,$$

gdzie liczba znaków c jest nie większa niż $n-t$.

Jeżeli $m = 0.\alpha_1 \ldots \alpha_t \alpha_{t+1} \ldots$, gdzie α_i są liczbami znaczącymi, to

$$m_t = \begin{cases} 0.\alpha_1 \ldots \alpha_t, & 0 \le \alpha_{t+1} \le \frac{\beta}{2} - 1 \\ 0.\alpha_1 \ldots \alpha_t + \beta^{-t}, & \frac{\beta}{2} \le \alpha_{t+1} < \beta \end{cases}$$

Wykażemy, że $|m - m_t| \le \frac{1}{2}\beta^{-t} \Leftrightarrow -\frac{1}{2}\beta^{-t} \le m - m_t \le \frac{1}{2}\beta^{-t}$. Jeżeli $0 \le \alpha_{t+1} \le \frac{\beta}{2} - 1$, to

$$\begin{aligned} m - m_t &= \sum_{i \ge t+1} \alpha_i \beta^{-i} = \alpha_{t+1}\beta^{-(t+1)} + \sum_{i \ge t+2} \alpha_i \beta^{-i} \\ &\le \left(\tfrac{\beta}{2} - 1\right)\beta^{-(t+1)} + (\beta - 1)\sum_{i=t+2}^{\infty} \beta^{-i} \\ &= \tfrac{1}{2}\beta^{-t} - \beta^{-(t+1)} + \sum_{i \ge t+1} \beta^{-i} - \sum_{i \ge t+2} \beta^{-i} = \tfrac{1}{2}\beta^{-t} + \sum_{i \ge t+1} \beta^{-i} - \sum_{i \ge t+1} \beta^{-i} \\ &= \tfrac{1}{2}\beta^{-t}. \end{aligned}$$

Podobnie, gdy $\frac{\beta}{2} \le \alpha_{t+1} < \beta$

$$\begin{aligned} m - m_t &= \sum_{i \ge t+1} \alpha_i \beta^{-i} - \beta^{-t} = \alpha_{t+1}\beta^{-(t+1)} + \sum_{i \ge t+2} \alpha_i \beta^{-i} - \beta^{-t} \\ &\ge -\tfrac{1}{2}\beta^{-t} + \sum_{i \ge t+2} \alpha_i \beta^{-i} \ge -\tfrac{1}{2}\beta^{-t} \end{aligned}$$

Jak widać, powyższa definicja m_t nie jest w pełni poprawna dla podstawy β nieparzystej. Sposobów zaokrąglenia jest wiele i często zależą one od architektury komputera.

Wprowadźmy teraz pojęcie dokładności maszynowej. Zauważmy, że

$$\left|\frac{\tilde{rd}(x) - x}{x}\right| = \left|\frac{\beta^c m_t - \beta^c m}{\beta^c m}\right| = \frac{|m_t - m|}{|m|} \le \frac{\frac{1}{2}\beta^{-t}}{m} \le \frac{1}{2}\frac{\beta^{-t}}{\beta^{-1}} = \frac{1}{2}\beta^{-(t-1)}.$$

Dokładnością maszynową nazywamy $\text{eps} = \frac{1}{2}\beta^{-(t-1)}$. Zatem moduł błędu względnego wartości $\tilde{rd}(x)$ jest ograniczony

$$|\varepsilon| = \left|\frac{\tilde{rd}(x) - x}{x}\right| \le \text{eps}\,.$$

Więc wzór na $\tilde{rd}(x)$ możemy zapisać w postaci

$$\tilde{rd}(x) = x(1 + \varepsilon),\ |\varepsilon| \le \text{eps}\,.$$

Niech $\mathtt{rd} : \mathbb{R} \to A$ będzie takie, że $\forall y \in A : \ |\mathtt{rd}(x) - x| \leq |y - x|$. Wartość $\mathtt{rd}(x)$ nazywamy **aproksymacją liczby x liczbą maszynową**.

Zauważmy, że jeśli $\tilde{\mathtt{rd}}(x) \in A$, to $\tilde{\mathtt{rd}}(x) = \mathtt{rd}(x)$. Nasuwa się więc pytanie, kiedy $\tilde{\mathtt{rd}}(x) \notin A$? Jest tak, jeżeli $c \notin [c_{min}, c_{max}]$. Jeśli $c < c_{min}$, mamy wówczas niedomiar cechy i przyjmujemy, że $\tilde{\mathtt{rd}}(x) = 0$. Błąd względny wynosi wówczas 100%. Gdy $c > c_{max}$, mamy nadmiar cechy i przerywamy obliczenia.

1.2 Operacje zmiennoprzecinkowe

Oznaczmy przez $\square$ dowolne z działań $+$, $-$, $\cdot$, $/$. Jeżeli $x, y \in A$, to niekoniecznie $x \square y \in A$. Operację zmiennoprzecinkową oznaczać będziemy przez $\square^\star$ i dla $x, y \in A$ definiujemy jako

$$x \square^\star y = (x \square y)(1 + \varepsilon), \ |\varepsilon| \leq \mathrm{eps}\,,$$

Najpierw zaokrąglamy, później działamy.

zatem

$$x \square^\star y = \tilde{\mathtt{rd}}(x \square y).$$

Operacje zmiennoprzecinkowe nie spełniają praw łączności i rozdzielności. W celu wykazania tej własności, oznaczmy przez $\mathtt{f}(E)$ wartość wyrażenia arytmetycznego E obliczonej według ustalonego algorytmu w arytmetyce zmiennoprzecinkowej.

Niech $A_1 = (a + b) + c$ i niech $A_2 = a + (b + c)$. Pokażemy, że $\mathtt{f}(A_1) \neq \mathtt{f}(A_2)$. Ustalmy więc $a, b, c \in A$.

$$\begin{aligned}
\mathtt{f}(A_1) &= (a +^\star b) +^\star c = ((a + b)(1 + \varepsilon_1) + c)(1 + \varepsilon_2) \\
&= (a + b + c + (a + b)\varepsilon_1)(1 + \varepsilon_2) \\
&= a + b + c + (a + b + c)\varepsilon_2 + (a + b)(1 + \varepsilon_2)\varepsilon_1 \\
&= (a + b + c)\big(1 + \tfrac{a+b}{a+b+c}(1 + \varepsilon_2)\varepsilon_1 + \varepsilon_2\big) = (a + b + c)(1 + \delta_1), \\
\delta_1 &= \tfrac{a+b}{a+b+c}(1 + \varepsilon_2)\varepsilon_1 + \varepsilon_2, \ |\varepsilon_1|, |\varepsilon_2| \leq \mathrm{eps}\,.
\end{aligned}$$

Podobnie obliczymy, że

$$\begin{aligned}
&\mathtt{f}(A_2) = (a + b + c)(1 + \delta_2), \\
&\delta_2 = \tfrac{b+c}{a+b+c}(1 + \varepsilon_4)\varepsilon_3 + \varepsilon_4, \ |\varepsilon_3|, |\varepsilon_4| \leq \mathrm{eps}\,.
\end{aligned}$$

Jeżeli przyjmiemy teraz, że a jest małe oraz $a \approx -b$, to $\frac{a+b}{a+b+c} \approx 0$. Więc $\delta_1 \approx \varepsilon_2$. Zachodzi również $\frac{b+c}{a+b+c} \approx \frac{b+c}{c} \approx 1$, zatem dla δ_2 mamy wzmocnienie błędu, więc $\delta_1 \neq \delta_2$.

Weźmy teraz $A_3 = a^2 - b^2$ oraz $A_4 = (a+b)(a-b)$.

$$\begin{aligned}
f(A_3) &= (a \cdot^\star a) -^\star (b \cdot^\star b) = a^2(1+\varepsilon_1) -^\star b^2(1+\varepsilon_2) \\
&= [a^2(1+\varepsilon_1) - b^2(1+\varepsilon_2)](1+\varepsilon_3) \\
&= a^2 - b^2 + (a^2-b^2)\varepsilon_3 + (a^2\varepsilon_1 - b^2\varepsilon_2)(1+\varepsilon_3) \\
&= (a^2-b^2)\big[1 + \tfrac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2}(1+\varepsilon_3) + \varepsilon_3\big] = (a^2-b^2)(1+\delta_3), \\
\delta_3 &= \frac{a^2\varepsilon_1 - b^2\varepsilon_2}{a^2-b^2}(1+\varepsilon_3) + \varepsilon_3,\ |\varepsilon_i| \le \mathrm{eps}\,,\ i = 1,2,3. \\
f(A_4) &= (a +^\star b) \cdot^\star (a -^\star b) = (a+b)(1+\varepsilon_4) \cdot^\star (a-b)(1+\varepsilon_5) \\
&= (a+b)(a-b)(1+\varepsilon_4)(1+\varepsilon_5)(1+\varepsilon_6) = (a^2-b^2)(1+\delta_4), \\
\delta_4 &= \varepsilon_4 + \varepsilon_5 + \varepsilon_6 + \ldots + \varepsilon_4\varepsilon_5\varepsilon_6,\ |\varepsilon_j| \le \mathrm{eps}\,,\ j = 4,5,6.
\end{aligned}$$

Morał!

Widać, że δ_4 jest rzędu 0, natomiast δ_3 może być dość duże. Zatem **wybór algorytmu może decydować o wielkości błędu**.

1.3 Uwarunkowanie zadania.

Niech $D \subset \mathbb{R}^m$ będzie zbiorem otwartym i niech $\varphi : D \to \mathbb{R}^n$ będzie funkcją ciągłą.

Definicja 1.2. *Zadanie* $y = \varphi(x)$ *jest* ***dobrze uwarunkowane*** *jeśli niewielkie zmiany danych dają małe zmiany wyników. W przeciwnym przypadku zadanie jest* ***źle uwarunkowane****.*

Zastanówmy się, kiedy dane zadanie jest dobrze uwarunkowane. Rozważmy taką sytuację: $\tilde{x}$ jest wartością przybliżającą x, $\tilde{y} = \varphi(\tilde{x})$, $y = \varphi(x)$. Pytamy, czy zachodzi implikacja

$$\Delta x = \tilde{x} - x \text{ małe } \Rightarrow \Delta y = \tilde{y} - y \text{ małe.}$$

Załóżmy, że $\varphi \in \mathcal{C}^1(D, \mathbb{R}^n)$, wówczas

$$\varphi(\tilde{x}) = \varphi(x) + \big(\tfrac{d\varphi}{dx}\big)(\tilde{x} - x) + \text{ reszta,}$$

gdzie

$$\begin{aligned}
&\varphi = (\varphi_1, \ldots, \varphi_n),\ x = (x_1, \ldots, x_m),\ \Delta x = (\Delta x_1, \ldots, \Delta x_m), \\
&\tfrac{d\varphi}{dx} = \big(\tfrac{\partial \varphi_i}{\partial x_j}\big)_{j=1,\ldots,m,\ i=1,\ldots,n} \\
&\Delta y_i = \varphi_i(\tilde{x}) - \varphi_i(x) = \sum_{j=1}^{m} \tfrac{\partial \varphi_i}{\partial x_j} \Delta x_j
\end{aligned}$$

Jeśli $y_i \neq 0$, to

$$\frac{\Delta y_i}{y_i} = \sum_{j=1}^{m} \frac{\partial \varphi_i}{\partial x_j} \cdot \frac{\Delta x_j}{y_i} = \sum_{j=1}^{m} \frac{\partial \varphi_i}{\partial x_j} \cdot \frac{x_j}{y_i} \cdot \frac{\Delta x_j}{x_j} = \sum_{j=1}^{m} k_{ij} \varepsilon_{x_j},$$

gdzie $k_{ij} = \frac{\partial \varphi_i}{\partial x_j} \cdot \frac{x_j}{y_i}$ nazywamy **współczynnikiem wzmocnienia błędu** ε_{x_j}.

Zatem odpowiedzią na postawione wyżej pytanie jest: **zadanie jest dobrze uwarunkowane, gdy** k_{ij} **są małe.**

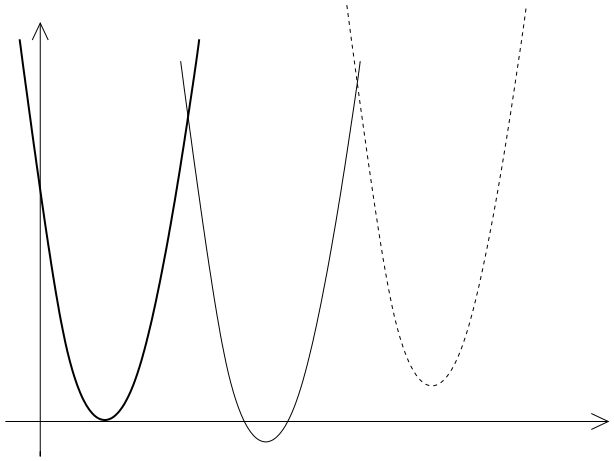
Przykład 1.3. Niech

$$y = \varphi(p, q) = p + \sqrt{p^2 - q},\ q > 0,\ p > 0.$$

Zatem $\varphi(p, q)$ jest rozwiązaniem równania

$$y^2 - 2py + q = 0.$$

Zadanie to jest źle uwarunkowane, jeżeli $p^2 \approx q$. Dla pierwiastków położonych blisko siebie każda zmiana wartości p i q, nawet niewielka, może spowodować brak pierwiastków lub znaczne ich oddalenie.



Rysunek 1:

2 Rozwiązywanie układów równań liniowych.

2.1 Wprowadzenie.

Niech $A = (a_{ij})$ będzie rzeczywistą macierzą wymiaru $n \times n$ i niech $b = (b_i) \in \mathbb{R}^n$. Rozdział ten poświęcimy na szukanie rozwiązania układu równań postaci

(1) $\quad Ax = b,$

równoważny układowi

$$(2) \quad \begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases}$$

Zakładamy, że A jest macierzą nieosobliwą ($\det A \neq 0$). Wprowadźmy oznaczenia

$$A = [a_1, \ldots, a_n],\ a_j \text{ reprezentuje kolumnę macierzy } A, a_j = \begin{pmatrix} a_{1j} \\ \vdots \\ a_{nj} \end{pmatrix},$$

$$A_i = [a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n].$$

Twierdzenie 2.1. *Jeżeli* A *jest macierzą nieosobliwą, to współczynniki* x_i *rozwiązania równania (1) dane są wzorami*

$$x_i = \frac{\det A_i}{\det A},\ i = 1, \ldots, n$$

Zatem, aby wyznaczyć te współczynniki należy obliczyć $n+1$ wyznaczników. Korzystając ze wzorów Cramera

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}\sigma a_{1\sigma(1)} \ldots a_{n\sigma(n)},$$

$\#S_n = n!$ gdzie S_n jest zbiorem permutacji zbioru $\{1, \ldots, n\}$, w celu obliczenia wyznacznika trzeba wykonać $n!(n-1)$ mnożeń. Łącznie dla wyznaczenia x trzeba wykonać

$$\begin{aligned}(n+1)(n-1)n! &= (n-1)(n+1)! \\ &\approx (n-1)\sqrt{2\pi(n+1)}\left(\tfrac{n+1}{e}\right)^{n+1}\left(1+\mathcal{O}\left(\tfrac{1}{n+1}\right)\right) \approx \mathcal{O}(n^{n+2})\end{aligned}$$

mnożeń. Dla $n = 20$ maszyna wykonująca $100\,000$ mnożeń na sekundę potrzebowałaby $3 \cdot 10^8$ lat. Zatem **metoda Cramera jest numerycznie bezużyteczna**.

2.2 Metody dokładne.

Zaczniemy od tzw. metod dokładnych, czyli takich, że po skończonej liczbie kroków otrzymamy rozwiązanie. Podstawową metodą tego typu jest metoda eliminacji Gaussa. Przez odpowiednie przekształcenia sprowadzimy macierz A do postaci trójkątnej. Układ taki będzie już można w łatwy sposób rozwiązać.

Do metod dokładnych należą też metody oparte na faktoryzacji macierzy. W metodach tego typu, dla danej macierzy nieosobliwej A szukać będziemy macierzy B, C takich, że $A = BC$ oraz macierze te dadzą się łatwo odwrócić (układy związane z nimi były łatwe do rozwiązania). Jeżeli znamy już te macierze, to układ (1) zastępujemy dwoma układami

$$(3) \quad \begin{cases} By = b \\ Cx = y \end{cases},$$

które są łatwe do rozwiązania.

2.2.1 Metoda eliminacji Gaussa.

Rozważmy układ (2)

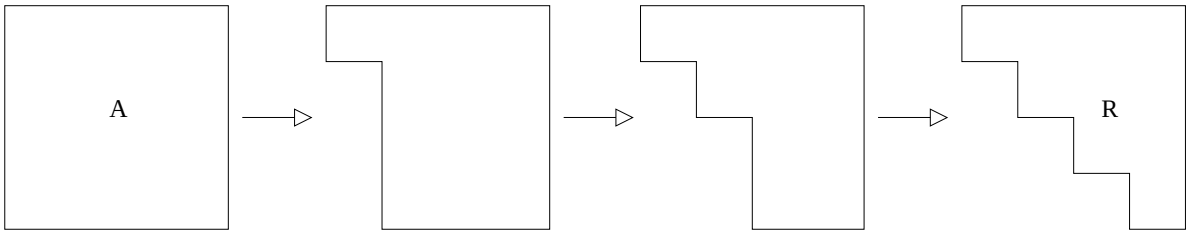
$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ a_{21}x_1 + a_{22}x_2 + \ldots + a_{2n}x_n = b_2 \\ \ldots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n = b_n \end{cases}$$

i załóżmy, że $a_{11} \neq 0$. Mnożymy pierwsze równanie przez $\frac{a_{i1}}{a_{11}}$ $(i = 2, 3, \ldots, n)$ i odejmujemy od i-tego równania, eliminując w ten sposób x_1 z i-tego równania. Dostajemy

więc równoważny układ

$$\begin{cases} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n = b_1 \\ \underbrace{\left(a_{22} - a_{12}\frac{a_{21}}{a_{11}}\right)}_{a_{22}^{(1)}} x_2 + \ldots + \underbrace{\left(a_{2n} - a_{1n}\frac{a_{21}}{a_{11}}\right)}_{a_{2n}^{(1)}} x_n = \underbrace{b_2 - b_1\frac{a_{21}}{a_{11}}}_{b_2^{(1)}} \\ a_{32}^{(1)}x_2 + \ldots + a_{3n}^{(1)}x_n = b_3^{(1)} \\ \ldots \\ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n = b_n^{(1)} \end{cases}$$

Oba układy są równoważne, to znaczy mają to samo rozwiązanie. Po odpowiednich przekształceniach i przy odpowiednich założeniach otrzymamy układ w postaci trójkątnej



Rysunek 2: Kolejne kroki

$$\begin{cases} r_{11}x_1 + r_{12}x_2 + \ldots + r_{1n}x_n = c_1 \\ r_{22}x_2 + \ldots + r_{2n}x_n = c_2 \\ \ldots \\ r_{n-1\,n-1}x_{n-1} + r_{n-1\,n}x_n = c_{n-1} \\ r_{nn}x_n = c_n \end{cases}$$

równoważny układowi (2), który jest o wiele łatwiejszy do rozwiązania.

Algorytm postępowania:

(i) Szukamy r_k takiego, że

$$|a_{r_k k}^{(k-1)}| = \max\{|a_{jk}^{(k-1)}| : j \geq k\}.$$

Zmianie miejsca ulega też prawa strona!

Element ten nazywamy **elementem podstawowym**.

(ii) W macierzy $(A^{(k-1)}, b^{(k-1)})$ przestawiamy miejscami wiersze k i r_k, tak powstałą macierz oznaczamy przez $(\tilde{A}^{(k-1)}, \tilde{b}^{(k-1)})$.

(iii) Obliczamy

$$l_{ik} = \frac{\tilde{a}_{ik}^{(k-1)}}{\tilde{a}_{kk}^{(k-1)}},\ i = k+1, \ldots, n$$

i od i–tego wiersza macierzy $(\tilde{A}^{(k-1)}, \tilde{b}^{(k-1)})$ odejmujemy k–ty wiersz tej macierzy pomnożony przez l_{ik}. W ten sposób eliminujemy x_k z wierszy $k+1, \ldots, n$. Tak otrzymaną macierz oznaczamy przez $(A^{(k)}, b^{(k)})$.

Istnienie elementu podstawowego.

Kroki (i) - (iii) powtarzamy $(n-1)$ krotnie. Otrzymujemy macierz trójkątną górną.

Ale skąd wiadomo, że element podstawowy jest zawsze niezerowy, co umożliwia nam dzielenie w kroku (iii)?

Załóżmy, że tak nie jest. Niech $S_k = \{a_{jk}^{(k-1)} |\ j = k, \ldots, n\}$ będzie zbiorem tych elementów, spośród których wybierzemy element podstawowy w $k-$tym kroku. Niech k_0 oznacza ten krok, w którym wszystkie elementy zbioru S_{k_0} są zerami. Macierz $A^{(k_0-1)}$ jest wówczas postaci

$$A^{(k_0-1)} = \begin{pmatrix} P & Q \\ 0 & S^{(k_0)} \end{pmatrix}$$

gdzie P ma zera pod diagonalną, a pierwsza kolumna macierzy $S^{(k_0)}$ ma same zera. Zatem $\det S^{(k_0)} = 0$. Ale macierz A jest nieosobliwa, zatem

$$0 \neq \det A = \det A^{(k_0-1)} = (\det S^{(k_0)})(\det P) = 0,$$

co dowodzi istnienia niezerowego elementu podstawowego.

Zatem, jeżeli macierz A jest nieosobliwa, to metoda eliminacji Gaussa prowadzi do rozwiązania układu (1).

Kto lepszy, Gauss czy Cramer?

Należy jeszcze odpowiedzieć na pytanie ile czasu nam to zajmie. W $k-$tym kroku wykonujemy $n-k$ dzieleń w celu wyliczenia l_{ik}, oraz $(n-k)(n-k+1)$ mnożeń $n-k$ wierszy z których każdy ma $n-k+1$ kolumn. Dla równania $Rx = b$ wykonujemy

$$1 + 2 + \ldots + n = \tfrac{n(n+1)}{2}$$

Wielkie 'O'... już gdzieś było.

mnożeń i dzieleń. Łącznie, wszystkich operacji tego typu trzeba wykonać

$$\sum_{i=1}^{m} i^2 \ = \ \tfrac{m(m+1)(2m+1)}{6}$$

$$\begin{aligned} M &= \sum_{k=1}^{n-1} [(n-k) + (n-k)(n-k+1)] + \tfrac{n(n+1)}{2} \\ &= 2\sum_{k=1}^{n-1} (n-k) + \sum_{k=1}^{n-1} (n-k)^2 + \tfrac{n(n+1)}{2} = \frac{n}{3}(n^2+3n-1) \\ &= 2\sum_{k=1}^{n-1} k + \sum_{k=1}^{n-1} k^2 + \tfrac{n(n+1)}{2} = \frac{n}{3}(n^2+3n-1) = \mathcal{O}(n^3). \end{aligned}$$

Zatem otrzymaliśmy algorytm znacznie efektywniejszy niż metoda Cramera.

2.2.2 Faktoryzacja LR.

Zauważmy, że w metodzie eliminacji Gaussa w $k-$tej iteracji krok (ii) jest równoważny mnożeniu lewostronnemu macierzy $A^{(k-1)}$ przez macierz permutacji $P_{k r_k}$ (powstałej z macierzy jednostkowej wymaru n przez zamianę ze sobą wierszy r_k i k). Zatem

$$\tilde{A}^{(k-1)} = P_{r_k k} A^{(k-1)}.$$

Krok (*iii*) polega na pomnożeniu lewostronnym macierzy $\tilde{A}^{(k-1)}$ przez macierz L_k postaci

$$L_k = [e_1, \dots, e_{k-1}, l_k, e_{k+1}, \dots, e_n],\ l_k = (0, \dots, 0, 1, -l_{k+1\,k}, \dots, -l_{n\,k})^T,$$

Zatem

$$L_k\tilde{A}^{(k-1)} = A^{(k)},$$
$$\dots$$
$$L_{n-1}P_{n-1,r_{n-1}}L_{n-2}P_{n-2,r_{n-2}}\dots L_1P_{1r_1}A = R,$$

gdzie R jest macierzą trójkątną górną, a L_j macierzami trójkątnymi dolnymi z jedynką na diagonali. Zanim przejdziemy do głównego twierdzenia wprowadźmy definicję macierzy trójkątnej oraz udowodnijmy kilka podstawowych własności.

Definicja 2.2. *Powiemy, że macierz kwadratowa $A = (a_{ij})$ jest trójkątna dolna (górna), gdy $a_{ij} = 0$ dla $i < j$ ($i > j$).*

Lemat 2.3. *Niech A, B będą macierzami kwadratowymi wymiaru n, $A = (a_{ij})$. Wówczas*

(i) Jeżeli macierz A jest trójkątna, to $\det A = a_{11} \cdot \dots \cdot a_{nn}$.

(ii) Jeśli macierze A, B są trójkątne dolne (górne), to AB jest macierzą trójkątną dolną (górną).

(iii) Jeśli macierz A jest trójkątna dolna (górna) i nieosobliwa, to A^{-1} jest macierzą trójkątną dolną (górną).

(iv) Jeżeli macierze A, B są trójkątne (obie dolne lub obie górne) z jedynkami na przekątnej, to macierze AB oraz A^{-1} są trójkątne (odpowiednio dolne lub górne) z jedynkami na przekątnej.

Dowód. Załóżmy, że A, B są macierzami trójkątnymi dolnymi. Dla dowodu pierwszej własności, niech $\sigma \in S_n$ będzie permutacją zbioru $\{1, \dots, n\}$ różną od identyczności. Wówczas istnieją $i, j \in \{1, \dots, n\}$ takie, że $i > \sigma(i)$ oraz $j < \sigma(j)$. Zatem dla każdej permutacji nie będącej identycznością istnieje k taki, że $a_{k\sigma(k)} = 0$. Mamy więc

$$\det A = \sum_{\sigma \in S_n} \operatorname{sgn}\sigma a_{1\sigma(1)} \cdot \dots \cdot a_{n\sigma(n)} = a_{11} \cdot \dots \cdot a_{nn}.$$

Niech teraz $C = AB$. Zatem $C = (c_{ij})$, gdzie

$$c_{ij} = \sum_{s=1}^{n} a_{is}b_{sj}.$$

Ustalmy $i < j$. Ponieważ macierz B jest trójkątna dolna, zatem $b_{sj} = 0$ dla $s < j$. Więc

$$c_{ij} = \sum_{s=j}^{n} a_{is}b_{sj}.$$

Ale A też jest macierzą trójkątną dolną, więc $a_{is} = 0$ dla $i < s$, co daje nam

$$c_{ij} \overset{i<j}{=} 0.$$

W celu wykazania trzeciej własności rozpatrzmy równanie

$$Ax = c,$$

gdzie macierz A jest macierzą trójkątną górną i nieosobliwą. Rozwiązanie tego równania dane jest rekurencją

$$\begin{aligned}
&x_n = \tfrac{1}{a_{nn}} c_n,\\
&x_{n-1} = \tfrac{1}{a_{n-1\,n-1}}(c_{n-1} - a_{n-1\,n}x_n),\\
&\dots\\
&x_i = \tfrac{1}{a_{ii}}\Big(c_i - \sum_{j=i+1}^{n} a_{ij}x_j\Big).
\end{aligned}$$

Widać, że x_i zależy liniowo od $c_i, \dots, c_n$, zatem $x = Zc$ dla pewnej trójkątnej macierzy górnej Z. Ponieważ $Ax = c$, więc $Z = A^{-1}$.

W celu wykazania ostatniej własności zauważmy najpierw, że skoro A jest macierzą trójkątną z jedynkami na diagonali, to jest nieosobliwa ($\det A = 1$). Na mocy własności drugiej i trzeciej wystarczy wykazać, że AB i A^{-1} mają na diagonali jedynki. Ustalmy i,

$$c_{ii} = \sum_{s=1}^{n} a_{is}b_{si} = \sum_{s=1}^{s=i} a_{is}b_{si} = a_{ii}b_{ii} = 1.$$

Niech $A^{-1} = (\tilde{a}_{ij})$ i ustalmy i. Ponieważ

$$A^{-1}A = I \Rightarrow 1 = \sum_{s=1}^{n} \tilde{a}_{is}a_{si} = \tilde{a}_{ii}.$$

□

Te wiadomości wystarczą nam do dowodu głównego twierdzenia w tym dziale.

Twierdzenie 2.4. *(O faktoryzacji macierzy nieosobliwej.)*
Jeżeli macierz A jest nieosobliwa, to istnieją macierze P permutacji, L trójkątna dolna z jedynkami na diagonali oraz R trójkątna górna takie, że

(4) $\quad PA = LR.$

Jeśli nie ma potrzeby wyboru elementu podstawowego, to $P = I$ i wzór (4) ma postać

$$A = LR.$$

Dowód. Na mocy metody eliminacji Gaussa mamy

$$(\star)\ L_{n-1}P_{n-1\,r_{n-1}}L_{n-2}P_{n-2\,r_{n-2}}\dots L_1P_{1r_1}A = R,$$

gdzie macierze $P_{k\,r_k}$ i L_k są jak wcześniej. Ponieważ $P_{ks}P_{ks} = I$ wzór $(\star)$ możemy zapisać w postaci

$$L_{n-1}P_{n-1\,r_{n-1}}\dots L_1P_{1r_1}\underbrace{P_{1r_1}P_{2r_2}\dots P_{n-1\,r_{n-1}}P_{n-1\,r_{n-1}}\dots P_{2j_2}P_{1r_1}}_{I}A = R.$$

Niech $Z = L_{n-1}P_{n-1\,r_{n-1}}\dots L_1P_{2r_2}\dots P_{n-1\,r_{n-1}}$. Twierdzimy, że jest to macierz trójkątna dolna z jedynkami na przekątnej. W tym celu przyjrzyjmy się macierzy $P_{2r_2}L_1P_{2r_2}$. Wiemy, że

$$L_1 = [l_1, e_2, \dots, e_n].$$

Mnożenie z prawej strony przez macierz permutacji P_{2r_2} zamienia między sobą kolumny o numerach 2 i r_2, zatem

$$L_1P_{2r_2} = [l_1, e_{r_2}, \dots, e_2, \dots, e_n].$$

Mnożenie z lewej strony przez macierz permutacji P_{2r_2} zamienia między sobą wiersze o numerach 2 i r_2, zatem

$$P_{2r_2}L_1 = [\tilde{l}_1, e_2, \dots, e_n],$$

gdzie $\tilde{l}_1$ powstał z l_1 przez zamianę miedzy sobą wyrazów o numerach 2 i r_2 (pamiętamy, że $r_2 \geq 2$). Otrzymaliśmy więc macierz trójkątną dolną z jedynkami na przekątnej). Oznaczmy tą macierz jako $L_1^{(1)}$. Mamy teraz

$$Z = L_{n-1}P_{n-1\,r_{n-1}}\dots, P_{3r_3}L_2L_1^{(1)}P_{3r_3}\dots P_{n-1\,r_{n-1}}.$$

Iloczyn $L_2L_1^{(1)}$ jest oczywiście macierzą trójkątną dolną z jedynkami na przekątnej i jest on postaci

$$L_2L_1^{(1)} = [x, l_2, e_3, \dots, e_4],\ \text{gdzie } x \text{ zależy od } \tilde{l}_1, l_2.$$

Powtarzając rozumowanie otrzymujemy, że Z jest macierzą trójkątną dolną z jedynkami na przekątnej.

Zatem wzór $(\star)$ można zapisać w postaci

$$ZPA = R,$$

gdzie $P = P_{n-1\,r_{n-1}} \cdot \dots \cdot P_{2r_2}P_{1r_1}$ jest macierzą permutacji. Możemy teraz zdefiniować $L = Z^{-1}$ (na mocy poprzedniego lematu jest macierzą trójkątną dolną z jedynkami na diagonali), co z połączeniem z ostatnią równością daje nam tezę

$$PA = Z^{-1}R = LR.$$

Jeśli nie potrzeba dokonywać wyboru elementu podstawowego, to

$$\forall k\ P_{k\,j_k} = I,$$

zatem

$$A = LR.$$

□

Zastanówmy się, kiedy nie trzeba wybierać elementu podstawowego. Niech $\mathcal{A}$ oznacza zbiór tych macierzy, że w metodzie eliminacji Gaussa nie trzeba wybierać elementu podstawowego. Na podstawie twierdzenia o faktoryzacji możemy sformułować wniosek.

Wniosek 2.5. *Jeśli* $A \in \mathcal{A}$*, to istnieją macierze macierz* L *trójkątna dolna, oraz macierz* R *trójkątna górna takie, że* $A = LR$*.*

Wykażemy teraz jeden z warunków wystarczających na to, aby dana macierz należała do $\mathcal{A}$.

A=A*>0 **Twierdzenie 2.6.** *Jeśli macierz* A *jest samosprzężona i dodatnio określona to* $A \in \mathcal{A}$*.*

Dowód. Skoro A jest samosprzężona to liczby znajdujące się na diagonali są rzeczywiste, możemy więc napisać, że

$$A = \begin{pmatrix} \alpha & a^* \\ a & A_1 \end{pmatrix},$$

gdzie $\alpha \in \mathbb{R}$, $a \in \mathbb{C}^{n-1}$ oraz $A_1 = A_1{}^*$ jest macierzą kwadratową wymiaru $n-1$. Wykażemy, że $\alpha > 0$.

Macierz A jest dodatnio określona, więc $\forall x : x^*Ax \geq 0$, oraz $x^*Ax = 0 \Leftrightarrow x = 0$([1]). Ustalmy więc

$$x = \begin{pmatrix} \xi \\ y \end{pmatrix} \in \mathbb{C}^n,\ y \in \mathbb{C}^{n-1}, x \neq 0.$$

Wówczas

$$\begin{aligned} x^*Ax &= (\bar{\xi}\ y^*) \begin{pmatrix} \alpha & a^* \\ a & A_1 \end{pmatrix} \begin{pmatrix} \xi \\ y \end{pmatrix} = (\bar{\xi}\ y^*) \begin{pmatrix} \alpha\xi + a^*y \\ a\xi + A_1y \end{pmatrix} \\ &= \bar{\xi}\alpha\xi + \bar{\xi}a^*y + y^*a\xi + y^*A_1y. \end{aligned}$$

Zatem $\alpha > 0$, bo jeśli $y = 0$ (więc $\xi \neq 0$), to

$$x^*Ax = \alpha|\xi|^2 > 0 \Rightarrow \alpha > 0.$$

Możemy zatem napisać, że

$$\begin{aligned} x^*Ax &= \alpha\big(|\xi|^2 + \tfrac{1}{\alpha}\bar{\xi}a^*y + \tfrac{1}{\alpha}\xi y^*a + \tfrac{1}{\alpha^2}|y^*a|^2\big) + y^*A_1y - \tfrac{1}{\alpha}\underbrace{y^*aa^*y}_{|y^*a|^2 = y^*a\overline{y^*a}} \\ &= \alpha\big(\bar{\xi} + \tfrac{1}{\alpha}y^*a\big)\big(\xi + \tfrac{1}{\alpha}a^*y\big) + y^*\big(A_1 - \tfrac{1}{\alpha}aa^*\big)y. \end{aligned}$$

Z dodatniej określoności macierzy A mamy, że macierz $A_1 - \frac{1}{\alpha}aa^*$ też jest macierzą dodatnio określoną. Jest tak, bo jeśli istnieje takie $y_0 \neq 0$, że $y_0{}^*(A_1 - \frac{1}{\alpha}aa^*)y_0 \leq 0$, to dobieramy do niego ξ_0 takie, aby $\bar{\xi}_0 + \frac{1}{\alpha}y_0{}^*a = 0$ i dla

$$x_0 = \begin{pmatrix} \xi_0 \\ y_0 \end{pmatrix}$$

[1] Gdy A jest macierzą rzeczywistą, to $A^* = A^T$ i warunek na dodatnią określoność ma postać

$\forall\ x \neq 0\ x^TAx > 0.$

mamy sprzeczność z dodatnią określonością macierzy A.

W takim razie możemy „wyeliminować pierwszą kolumnę”

$$\begin{pmatrix} 1 & 0 \\ l & I \end{pmatrix} \begin{pmatrix} \alpha & a^* \\ a & A_1 \end{pmatrix} = \begin{pmatrix} \alpha & a^* \\ l\alpha + a & A_1 + la^* \end{pmatrix},$$
$$l\alpha + a = 0 \Leftrightarrow l = -\tfrac{1}{\alpha}a \Rightarrow la^* + A_1 = -\tfrac{1}{\alpha}aa^* + A_1,$$
$$\begin{pmatrix} 1 & 0 \\ l & I \end{pmatrix} \begin{pmatrix} \alpha & a^* \\ a & A_1 \end{pmatrix} = \begin{pmatrix} \alpha & a^* \\ 0 & A_1 - \frac{1}{\alpha}aa^* \end{pmatrix}.$$

Skoro $A_1 - \frac{1}{\alpha}aa^*$ jest dodatnio określona, więc możemy powtórzyć procedurę. Zatem $A \in \mathcal{A}$. □

Wyznaczmy macierze L i R. Niech $A = (a_{ij})$, $L = (l_{ij})$, $R = (r_{ij})$. Skoro $A = LR$, to

$$a_{ij} = \sum_{k=1}^{n} l_{ik} r_{kj}.$$

Wiemy też, że

$$l_{ik} = 0, \text{ dla } i < k.$$
$$r_{kj} = 0, \text{ dla } k > j,$$

więc

$$a_{ij} = \sum_{k=1}^{\min(i,j)} l_{ik} r_{kj}.$$

Macierz L ma jedynki na diagonali, zatem

$$a_{1j} = \sum_{k=1}^{\min(1,j)} l_{1k} r_{kj} = l_{11} r_{1j} \Rightarrow r_{1j} = a_{1j}, \forall\, j = 1, \ldots, n.$$

Podobnie

$$a_{i1} = l_{i1} r_{11} \Rightarrow l_{i1} = \frac{a_{i1}}{r_{11}}, \ \forall\, i = 1, \ldots, n.$$

Znamy więc wzory na pierwszy wiersz macierzy R i pierwszą kolumnę macierzy L. W podobny sposób wyznaczymy kolejne wyrazy. Ogólnie, znając $r_{j\cdot}$ i $l_{\cdot j}$, $j = 1, \ldots, k-1$ mamy

$$a_{kj} = \sum_{s=1}^{k} l_{ks} r_{sj} = \sum_{s=1}^{k-1} l_{ks} r_{sj} + l_{kk} r_{kj} \Rightarrow r_{k\cdot} = a_{k\cdot} - \sum_{s=1}^{k-1} l_{ks} r_{s\cdot}.$$
$$a_{ik} = \sum_{s=1}^{k} l_{is} r_{sk} = \sum_{s=1}^{k-1} l_{is} r_{sj} + l_{ik} r_{kk} \Rightarrow l_{\cdot k} = \tfrac{1}{r_{kk}} \Big(a_{\cdot k} - \sum_{s=1}^{k-1} l_{\cdot s} r_{sk} \Big).$$

Zajmijmy się teraz twierdzeniem, które można potraktować jako wniosek z twierdzenia o faktoryzacji.

Twierdzenie 2.7. *(wniosek z twierdzenia o faktoryzacji[2])*
Załóżmy, że A *jest samosprzężoną macierzą rzeczywistą dodatnio określoną. Wówczas istnieje macierz trójkątna dolna* K *taka, że* $A = KK^T$.

Dowód. Niech $A \in \mathcal{A}$, zatem istnieją macierze L trójkątna dolna i R trójkątna górna takie, że $A = LR$. Zatem $A^T = R^T L^T$. A jest samosprzężona, więc

$$LR = R^T L^T,$$
$$LR(L^T)^{-1} = R^T,$$
$$R(L^T)^{-1} = L^{-1} R^T.$$

Ponieważ macierze R i $(L^T)^{-1}$ są trójkątne górne, a macierze L^{-1} i R^T są trójkątne dolne, więc z ostatniej równości iloczyny te muszą być macierzami diagonalnymi. Niech więc $D^2 = L^{-1}R^T$. Pokażemy, że jest to macierz dodatnio określona. W tym celu ustalmy dowolony $x \neq 0$, mamy

$$x^T D^2 x = x^T R(L^T)^{-1} x.$$

Ponieważ dla dowolnego x istnieje y taki, że $x = L^T y$, więc

$$x^T D^2 x = y^T LRy = y^T Ay > 0,$$

bo A jest dodatnio określona. Więc $D^2 = \operatorname{diag}(d_1^2, \ldots, d_n^2)$. Niech $D = \operatorname{diag}(d_1, \ldots, d_n)$, wówczas

$$R(L^T)^{-1} = D^2 \Rightarrow R = D^2 L^T \overset{A=LR}{\Rightarrow} A = LDDL^T = LD(LD)^T.$$

Za macierz K wystarczy więc przyjąć LD. □

Spróbujmy teraz wyprowadzić efektywny wzór na wyrazy macierzy K. Niech

$$A = (a_{ij}),\ K = (k_{ij}),\ K^T = (\tilde{k}_{ij}).$$
$$A = KK^T \Rightarrow a_{ij} = \sum_{s=1}^{i} k_{is} \cdot \tilde{k}_{sj} = \sum_{s=1}^{i} k_{is} k_{js}.$$

Kolejne wyrazy macierzy K wyznaczać będziemy kolumnami

$$a_{11} = {k_{11}}^2 \Rightarrow k_{11} = \sqrt{a_{11}}.$$
$$a_{1j} = k_{11} k_{j1} \Rightarrow k_{j1} = \tfrac{1}{k_{11}} a_{1j},\ j = 1, \ldots, n.$$

Mamy już pierwszą kolumnę macierzy K, zanim zapiszemy wzór dla dowolnej kolumny, sprawdźmy jak to wygląda dla drugiej kolumny. Znamy już k_{21}, możemy więc wyznaczyć

$$a_{22} = k_{21}k_{21} + k_{22}k_{22} \Rightarrow k_{22} = \sqrt{a_{22} - {k_{21}}^2} = \sqrt{a_{22} - \tfrac{{a_{12}}^2}{a_{11}}},$$

a dzięki temu

$$a_{2j} = k_{21}k_{j1} + k_{22}k_{j2} \Rightarrow k_{j2} = \tfrac{1}{k_{22}}(a_{2j} - k_{21}k_{j1}),\ j = 1, \ldots, n.$$

[2]Postać Choleskiego macierzy symetrycznej dodatnio określonej.

Ogólnie, gdy znamy kolumny $1, \ldots, i-1$ macierzy K, to

$$a_{ii} = \sum_{s=1}^{i} k_{is}^2 \Rightarrow k_{ii} = \sqrt{a_{ii} - {k_{i1}}^2 - \ldots - {k_{i\,i-1}}^2},$$
$$a_{ij} = \sum_{s=1}^{i-1} k_{is}k_{js} + k_{ii}k_{ji} \Rightarrow k_{ji} = \tfrac{1}{k_{ii}}\Big(a_{ij} - \sum_{s=1}^{i-1} k_{is}k_{js}\Big).$$

2.2.3 Faktoryzacja QR

Zajmijmy się teraz innym rozkładem macierzy, tym razem na macierz ortogonalną (Q) i macierz trójkątną górną (R). Rozkład ten ma zastosowanie przy **wyznaczaniu wartości własnych**. Zanim się nim zajmiemy udowodnijmy

Twierdzenie 2.8. *(O ortogonalizacji.)*
Dana niech będzie przestrzeń unitarna $(X, (\cdot|\cdot))$ *oraz ciąg* $\{f_n\}_{n\in\mathbb{N}} \subset X$ *wektorów liniowo niezależnych* [3]*. Wówczas istnieje ciąg* $\{g_n\}_{n\in\mathbb{N}} \subset X$ *taki, że*

(i) $(g_i|g_j) = 0$ *dla* $i \neq j$*;*

Zatem nasz ciąg też jest liniowo niezależny.

(ii) $\forall k \in \mathbb{N} : \operatorname{span}\{f_1, \ldots, f_k\}$ [4] $= \operatorname{span}\{g_1, \ldots, g_k\}$.

Dowód. Zdefiniujmy ciąg $\{g_n\}_{n\in\mathbb{N}}$ wzorem

$$(5) \quad \begin{cases} g_1 = f_1 \\ g_k = f_k - \sum_{s=1}^{k-1} r_{sk}g_s, \ k \geq 2 \end{cases},$$

gdzie r_{ij} $(i < j)$ są pewnymi stałymi, które wyznaczymy tak, aby ciąg $\{g_n\}_n$ spełniał tezę twierdzenia. Ustalmy więc k i załóżmy, że znamy $g_1, \ldots, g_{k-1}$ $(k > 1)$. Wówczas dla dowolnego $i < k$:

$$(g_i|g_k) = \big(g_i\big|f_k - \sum_{s=1}^{k-1} r_{sk}g_s\big) = (g_i|f_k) - \sum_{s=1}^{k-1} r_{sk}(g_i|g_s) = (g_i|f_k) - r_{ik}\|g_i\|^2.$$

Skoro ciąg $\{g_n\}_n$ (a tym samym $g_1, \ldots, g_k$) ma spełniać *(i)*, to dla $i < k$ zachodzić musi równość

$$0 = (g_i|g_k) = (g_i|f_k) - r_{ik}\|g_i\|^2.$$

Więc

$$r_{ik} = \frac{(g_i|f_k)}{\|g_i\|^2}, \ i < k.$$

Należy jeszcze sprawdzić, czy $g_i \neq 0$ dla $i < k$. Zajmiemy się tym później. Pokażmy najpierw, że ciąg (5) jest liniowo niezależny i spełnia *(ii)*. Ustalmy $k > 0$ i niech $f \in \operatorname{span}\{f_1, \ldots, f_k\}$, więc

$$f = \sum_{s=1}^{k} \alpha_s f_s.$$

[3] Tzn. $\forall k \in \mathbb{N} : f_1, \ldots, f_k$ są liniowo niezależne.
[4] $\operatorname{span}\{f_1, \ldots, f_k\} = \{x \in X |\ x = \alpha_1 f_1 + \ldots + \alpha_k f_k, \ \alpha_j \in \mathbb{R}\}$

Z konstrukcji ciągu $\{g_n\}_n$ wiemy, że dla dowolnego i $f_i \in \operatorname{span}\{g_1, \ldots, g_i\}$, czyli

$$f_i = \sum_{t=1}^{i} \beta_t g_t.$$

Zatem

$$\begin{aligned} f &= \sum_{1\le s\le k} \alpha_s \sum_{1\le t\le s} \beta_t g_t = \sum_{s,t} \alpha_s \beta_t g_t [1 \le t \le s \le k] \\ &= \sum_{s,t} \alpha_s \beta_t g_t [1 \le t \le k][t \le s \le k] = \sum_{1\le t\le k} \Big(\Big(\sum_{t\le s\le k} \alpha_s \Big) \beta_t \Big) g_t \in \operatorname{span}\{g_1, \ldots, g_k\}. \end{aligned}$$

Zatem $\operatorname{span}\{f_1, \ldots, f_k\} \subseteq \operatorname{span}\{g_1, \ldots, g_k\}$ dla dowolnego k. Zawieranie w drugą stronę wykazuje się podobnie (korzysta się z tego, że z konstrukcji ciągu $\{g_n\}_n$ wynika, że $g_i \in \operatorname{span}\{f_1, \ldots, f_i\}$ dla dowolnego i).

Pozostaje zatem wykazać, że $g_k \neq 0$ dla dowolnego k. Skoro ciąg $\{f_n\}_n$ jest liniowo niezależny, to $f_1 \neq 0$, a tym samym $g_1 \neq 0$. Pokażmy, że $g_1, \ldots, g_{k-1} \neq 0 \Rightarrow g_k \neq 0$. Dla dowodu nie wprost przypuśćmy, że $g_k = 0$. Wówczas

$$f_k = \sum_{s=1}^{k-1} r_{sk} g_s,$$

stąd

$$f_k \in \operatorname{span}\{g_1, \ldots, g_{k-1}\} = \operatorname{span}\{f_1, \ldots, f_{k-1}\},$$

czyli $f_1, \ldots, f_k$ są liniowo zależne – sprzeczność z założeniem. Tak więc $g_k \neq 0$. □

Przejdźmy teraz do głównego twierdzenia tego tematu.

Twierdzenie 2.9. *(Faktoryzacja QR)*
Niech $A \in \mathbb{R}^{n\times n}$ *będzie macierzą nieosobliwą,* $A = (a_{ij})$. *Wówczas istnieją macierze* $Q = (q_{ij})$ *ortogonalna* [5] *i macierz* $R = (r_{ij})$ *trójkątna górna takie, że*

$$A = QR.$$

Dowód. Niech

$$A = [a_1, \ldots, a_n], \text{ gdzie } a_j \in \mathbb{R}^n \text{ jest } j\text{-tą kolumną macierzy} A.$$

Macierz A jest nieosobliwa, zatem wektory $a_1, \ldots, a_n$ są liniowo niezależne. Na mocy twierdzenia o ortogonalizacji ciąg wektorów

$$(6) \quad \begin{cases} q_1 = a_1 \\ q_k = a_k - \sum_{s=1}^{k-1} \tilde{r}_{sk} q_s, \ k = 2, \ldots, n \end{cases},$$

[5] Macierz Q jest ortogonalna wtedy i tylko wtedy, gdy $Q^T Q$ jest macierzą diagonalną.

jest liniowo niezależny i

(7) $(q_i|q_j) = 0, i \neq j.$

Przyjmijmy

$$r_{ii} = 1,\ i = 1, \ldots, n$$
$$r_{ij} = \tilde{r}_{ij},\ i < j,$$
$$r_{ij} = 0,\ i > j,$$
$$Q = [q_1, \ldots, q_n].$$

Macierz Q jest oczywiście ortogonalna, a R jest trójkątna górna. Wystarczy wykazać, że $A = QR$. Niech $(\tilde{a}_{ij}) = \tilde{A} = QR$, więc Z definicji (6) wektorów q_k mamy

$$\tilde{a}_{ij} = \sum_{s=1}^{n} q_{is} r_{sj} = \sum_{s=1}^{j} q_{is} r_{sj} = \sum_{s=1}^{j-1} q_{is} r_{sj} + q_{ij} r_{jj} = \sum_{s=1}^{j-1} q_{is} \tilde{r}_{sj} + q_{ij} \overset{(6)}{=} a_{ij},$$

zatem $A = \tilde{A}$, co kończy dowód. □

Dzięki temu rozkładowi układ (1) można sprowadzić do układu który łatwo rozwiązać:

$$QRx = b,$$
$$Q^T QRx = Q^T b,$$
$$DRx = Q^T b,$$

gdzie DR jest macierzą trójkątną górną.

2.3 Metody przybliżone (interacyjne).

2.3.1 Macierze. Normy macierzowe.

Zacznijmy od przypomnienia podstawowych wiadomości dotyczących norm w $\mathbb{R}^n$ (wektorowych). Tak więc odwzorowanie $\|\cdot\| : \mathbb{R}^n \to [0, \infty)$ jest normą, gdy

(i) $\|x\| = 0 \Leftrightarrow x = 0$;

(ii) $\forall x \in \mathbb{R}^n \forall \alpha \in \mathbb{R} : \|\alpha x\| = |\alpha| \cdot \|x\|$;

(iii) $\forall x, y \in \mathbb{R}^n : \|x + y\| \leq \|x\| + \|y\|$.

Norma jest ciągła ze względu na warunek trójkąta.

Najczęściej będziemy zajmować się normami postaci

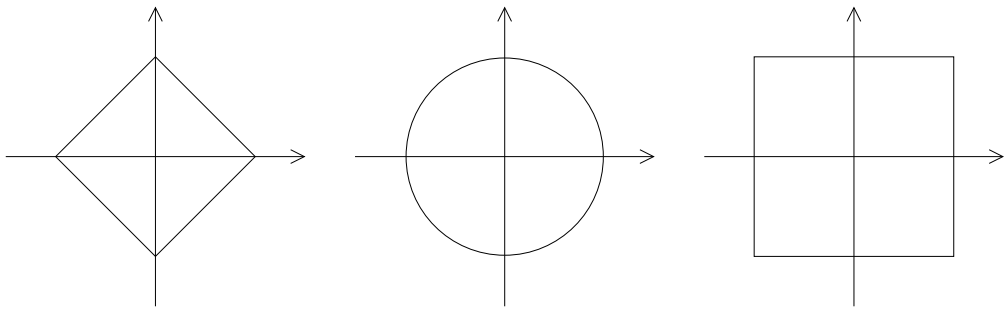
$$\|x\|_p = (\sum_{i=1}^{k} |x_i|^p)^{\frac{1}{p}}, \infty \geq p \geq 1,$$

z czego najważniejsze są

$$\|x\|_1 = \sum_{j=1}^{n} |x_j|,$$

$$\|x\|_2 = (\sum_{j=1}^{n} |x_j|^2)^{1/2} = \sqrt{(x|x)},\ (x|y) = x^* y,$$

$$\|x\|_\infty = \max\{|x_i| :\ i = 1, \ldots, n\}.$$



Rysunek 3: Kule w normach $\|\cdot\|_1,\ \|\cdot\|_2,\ \|\cdot\|_\infty$

Uwaga 2.10. *Wszystkie normy w* $\mathbb{R}^n$ *są równoważne, tzn.*

$$\forall\ \|\cdot\|',\ \|\cdot\|''\ \exists\ \alpha, \beta > 0\ \forall\ x \in \mathbb{R}^n : \alpha\|x\|' \le \|x\|'' \le \beta\|x\|'.$$

Jest to równoważne temu, że

$$a_k \underset{k\to\infty}{\longrightarrow} a\ w\ \|\cdot\|' \Leftrightarrow a_l \underset{l\to\infty}{\longrightarrow} a\ w\ \|\cdot\|'',$$

lub, że odwzorowanie

$$\mathrm{id} : (\mathbb{R}^n, \|\cdot\|') \to (\mathbb{R}^n, \|\cdot\|'')$$

jest ciągłe i odwrotne do niego też jest ciągłe.

Przejdźmy teraz do *normy operatora*. Niech $A = (a_{ij}) \in \mathbb{K}^{n\times n}$ będzie macierzą rzeczywistą (zespoloną) i niech $\|\cdot\|$ będzie dowolną normą wektorową w $\mathbb{K}^n$.

Definicja 2.11. *Normą*[6] *macierzy* A *zgodną z normą wektorową* $\|\cdot\|$ *nazywamy*

$$\|A\| = \max\left\{\frac{\|Ax\|}{\|x\|} : x \neq 0\right\}. \tag{8}$$

[6]Definicję tą naturalnie można uogólnić dla macierzy dowolnych wymiarów. Gdy A jest macierzą $n \times m$, wówczas

$$\|A\| = \max\left\{\frac{\|Ax\|}{\|x\|} : x \neq 0\right\},$$

gdzie $\|\cdot\|$ jest normą zarówno w $\mathbb{R}^n$ jak i w $\mathbb{R}^m$.

Uwaga 2.12. *Normę macierzy możemy zapisać jako*

$$\|A\| = \max\{\|Ax\| : \ \|x\| = 1\}.$$

Wzór (8) określa normę w $\mathbb{K}^{n\times n}$. Oczywiście $\|A\| \geq 0$.

$$\|A\| = 0 \Leftrightarrow \max\{\|Ax\| : \ \|x\| = 1\} = 0 \Leftrightarrow \forall\, x : \|Ax\| = 0 \Leftrightarrow A = 0;$$
$$\frac{\|\lambda Ax\|}{\|x\|} = \frac{|\lambda| \cdot \|Ax\|}{\|x\|} \Rightarrow \|\lambda A\| = |\lambda| \cdot \|A\|;$$
$$\frac{\|(A+B)x\|}{\|x\|} = \frac{\|Ax + Bx\|}{\|x\|} \leq \frac{\|Ax\|}{\|x\|} + \frac{\|Bx\|}{\|x\|} \Rightarrow \|A + B\| \leq \|A\| + \|B\|.$$

Ponadto norma operatorowa spełnia dwa bardzo ważne dla metod numerycznych warunki

(9) $\|Ax\| \leq \|A\| \cdot \|x\|$,
(10) $\|AB\| \leq \|A\| \cdot \|B\|$.

Warunek (9) wynika z

$$\frac{\|Ax\|}{\|x\|} \leq \max\left\{\frac{\|Ax\|}{\|x\|} : x \neq 0\right\} = \|A\|,$$

natomiast warunek (10) z

$$\frac{\|ABx\|}{\|x\|} = \frac{\|A(Bx)\|}{\|x\|} \overset{(9)}{\leq} \|A\|\frac{\|Bx\|}{\|x\|}.$$

Przykład 2.13. Przykład normy w $\mathbb{R}^{n^2}$ która nie spełnia warunku (10)

$$|A| = \max\{|a_{ij}| : i = 1, \ldots, n,\ j = 1, \ldots, n\}.$$

Wystarczy wziąć

$$A = B = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix},$$

wówczas

$$AB = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix},$$

$|AB| = 2,\ |A| = |B| = 1$.

Definicja 2.14. *Niech $A = (a_{ij}) \in \mathbb{K}^{m\times n}$ będzie macierzą. Liczbę zespoloną λ nazywamy wartością własną jeśli istnieje wektor $x \neq 0$ taki, że*

$$Ax = \lambda x.$$

Jeśli tak jest, to x nazywamy wektorem własnym macierzy A skojarzonym z wartością własną λ.

Zbiór wszystkich wartości własnych macierzy A nazywamy widmem (spektrum) i oznaczamy przez $\sigma(A)$.

Przez $\lambda_{max}(A)$ rozumieć będziemy największą wartość własną macierzy A.

Lemat 2.15. *Dla dowolnej macierzy A macierz A^*A jest symetryczna i półdodatnio określona. Ponadto $\lambda_{max}(A^*A) \geq 0$.*

Twierdzenie 2.16. *Niech $A = (a_{ij})$ będzie $n-$wymiarową macierzą rzeczywistą (zespoloną), wówczas*

$$\|A\|_\infty = \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\},$$
$$\|A\|_1 = \max\Big\{\sum_{i=1}^{n} |a_{ij}| : \ j = 1, \ldots, n\Big\},$$
$$\|A\|_2 = \sqrt{\lambda_{max}(A^TA)}.$$

Dowód. Niech $A = (a_{ij})$ będzie macierzą kwadratową (wymiaru n) i niech $x = (x_1, \ldots, x_n)^T$. Wówczas

$$\begin{pmatrix} a_{11} & \ldots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \ldots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nn}x_n \end{pmatrix}$$

Zajmijmy się najpierw $\|A\|_\infty$. Ponieważ dla $i = 1, \ldots, n$

$$\Big|\sum_{j=1}^{n} a_{ij}x_j\Big| \leq \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| \leq \|x\|_\infty \sum_{j=1}^{n} |a_{ij}|,$$

więc

$$\|Ax\|_\infty \leq \|x\|_\infty \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\},$$
$$\|A\|_\infty \leq \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\}.$$

Aby wykazać, że zachodzi nierówność "$\geq$" wystarczy wskazać takie x, aby zachodziła równość. Niech zatem $k = \text{argmax}\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\}$ i niech $y = (y_1, \ldots, y_n)^T$, gdzie

$$y_j = \begin{cases} 1, & a_{kj} \geq 0 \\ -1, & a_{kj} < 0 \end{cases}.$$

Wówczas $\|y\|_\infty = 1$, $a_{kj}y_j = |a_{kj}|$ oraz

$$\Big|\sum_{j=1}^{n} a_{kj}y_j\Big| = \sum_{j=1}^{n} a_{kj}y_j = \sum_{j=1}^{n} |a_{kj}| = \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\} \Rightarrow$$
$$\frac{\|Ay\|_\infty}{\|y\|_\infty} = \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\} \Rightarrow \|A\|_\infty \geq \max\Big\{\sum_{j=1}^{n} |a_{ij}| : \ i = 1, \ldots, n\Big\}.$$

Przejdźmy teraz do $\|A\|_1$. Niech $b = \max\{\sum_{i=1}^{n} |a_{ij}| : \ j = 1, \ldots, n\}$. Zatem

$$\|Ax\|_1 = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} a_{ij}x_j \Big| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} |a_{ij}| \cdot |x_j| = \sum_{j=1}^{n} |x_j| \sum_{i=1}^{n} |a_{ij}| \leq b \sum_{j=1}^{n} |x_j| = b\|x\|_1,$$

$$\forall\, x \neq 0 : \ \frac{\|Ax\|_1}{\|x\|_1} \leq b \Rightarrow \|A\|_1 \leq b.$$

Weźmy $k = \operatorname{argmax}\{\sum_{i=1}^{n} |a_{ij}| : \ j = 1, \ldots, n\}$ i $e_k = (\delta_{ik})^T, \ i = 1, \ldots, n$. Oczywiście $\|e_k\|_1 = 1$ i

$$\|Ae_k\|_1 = \sum_{i=1}^{n} |a_{ik}| = b \Rightarrow \|A\|_1 \geq b,$$

co dowodzi równości $\|A\|_1 = \max\{\sum_{i=1}^{n} |a_{ij}| : j = 1, \ldots, n\}$.

Na koniec chcemy wykazać, że $\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)}$. Na mocy ostatniego lematu wzór ma sens. Z definicji mamy

$$\|Ax\|_2 = ((Ax)^T(Ax))^{1/2} = \sqrt{x^T(A^TA)x}.$$

Macierz A^TA jest symetryczna i półdodatnio określona, zatem z algebry liniowej mamy, że

$$\|x\|_2^2\lambda_{\min}A^TA \leq x^T(A^TA)x \leq \|x\|_2^2\lambda_{\max}(A^TA),$$

zatem

$$\frac{\|Ax\|_2}{\|x\|_2} = \frac{\sqrt{x^T(A^TA)x}}{\|x\|_2} \leq \sqrt{\lambda_{\max}(A^TA)}.$$

Aby wykazać nierówność w drugą stronę weźmy v wektor własny macierzy A^TA skojarzony z wartością własną $\lambda_{\max}$. Wówczas

$$\frac{\|Av\|_2}{\|v\|_2} = \frac{\sqrt{v^T(A^TA)v}}{\|v\|_2} = \frac{\sqrt{\lambda_{\max}(A^TA)}\sqrt{v^Tv}}{\|v\|_2} = \frac{\|v\|_2\sqrt{\lambda_{\max}(A^TA)}}{\|v\|_2} = \sqrt{\lambda_{\max}(A^TA)},$$

a zatem $\|A\|_2 \geq \sqrt{\lambda_{\max}(A^TA)}$ co kończy dowód. □

Wniosek 2.17. *Jeżeli $A = A^T > 0$, to $\|A\|_2 = \lambda_{\max}(A)$.*

Dowód.

$$\|A\|_2 = \sqrt{\lambda_{\max}(A^TA)} = \sqrt{\lambda_{\max}(A^2)} = \sqrt{\lambda_{\max}^2(A)} = \lambda_{\max}(A).$$

□

Zajmiemy się teraz współczynnikiem uwarunkowania zadania rozwiązania układu równań postaci

$$Ax = b,$$

gdzie A jest macierzą nieosobliwą. Niech zatem $\tilde{b} = A\tilde{x}$, gdzie $\tilde{x}$ przybliża x. Wówczas

$$A\Delta x = A(\tilde{x} - x) = A\tilde{x} - Ax = \tilde{b} - b = \Delta b,$$
$$\Delta x = A^{-1}\Delta b,$$
$$\|b\| = \|Ax\| \leq \|A\| \cdot \|x\| \Rightarrow \|x\| \geq \frac{\|b\|}{\|A\|},$$
$$\frac{\|\Delta x\|}{\|x\|} = \frac{\|A^{-1}\Delta b\|}{\|x\|} \leq \frac{\|A^{-1}\| \cdot \|\Delta b\|}{\|x\|} \leq \|A\| \cdot \|A^{-1}\| \frac{\|\Delta b\|}{\|b\|}.$$

Wartość $\|A\| \cdot \|A^{-1}\|$ nazywamy *wskaźnikiem uwarunkowania zadania rozwiązywania układu równań liniowych* i oznaczamy przez $\mathrm{cond}(A)$.

Jeżeli $A^{-1} = (\alpha_{ij})$, to $\alpha_{ij} = \frac{\beta_{ji}}{\det A}$, gdzie $\beta_{ji} = (-1)^{i+j} \det A_{ji}$ (A_{ji} jest to macierz która powstała przez usunięcie z macierzy A j–tego wiersza i i–tej kolumny). Widać więc, że jeśli $\det A \approx 0$, to pewne α_{ij} są duże, a zatem $\mathrm{cond}(A)$ jest duże (bo $\|A^{-1}\|$ jest duże).

Definicja 2.18. *Wielomianem charakterystycznym macierzy A nazywamy*

$$\varphi_A(\lambda) = \det(A - \lambda I).$$

$$\det(A - \lambda I) = \begin{vmatrix} a_{11} - \lambda & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} - \lambda & a_{23} & \dots & a_{2n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} - \lambda \end{vmatrix} =$$
$$(a_{11} - \lambda) \cdot \ldots \cdot (a_{nn} - \lambda) + \text{reszta} = (-\lambda)^n + (-\lambda)^{n-1} \cdot \mathrm{tr}\{A\} + \ldots + \det A + \text{reszta},$$

a zatem widać, że φ_A jest wielomianem stopnia n. Wykazaliśmy więc

Twierdzenie 2.19. *Macierz $A \in \mathbb{C}^{n \times n}$ ma n wartości własnych (licząc z krotnościami).*

Dowód. Definicja wielomianu charakterystycznego + zasadnicze twierdzenie algebry. □

Definicja 2.20. *Macierze A i B są podobne ($A \sim B$) jeśli istnieje nieosobliwa macierz P taka, że*

$$A = PBP^{-1}.$$

Lemat 2.21. *Relacja podobieństwa jest równoważnością. Co więcej, jeśli macierze A i B są podobne, to ich wielomiany charakterystyczne $\varphi_A(\lambda)$ i $\varphi_B(\lambda)$ są sobie równe.*

Dowód.

$$\begin{aligned}\varphi_A(\lambda) &= \det(A - \lambda I) = \det(PBP^{-1} - \lambda I) = \det(P(B - \lambda I)P^{-1}) \\ &= \det P \cdot \varphi_B(\lambda) \cdot \det(P^{-1}) = \varphi_B(\lambda).\end{aligned}$$

□

Wniosek 2.22. *Macierze podobne mają te same wartości własne licząc z krotnościami.*

Twierdzenie 2.23. *(postać kanoniczna Jordana macierzy A.)*
Niech A będzie macierzą kwadratową wymiaru n i niech $\varphi_A(\lambda)$ będzie jej wielomianem charakterystycznym, tzn.

$$\varphi_A(\lambda) = (\lambda - \lambda_1)^{n_1} \cdot \ldots \cdot (\lambda - \lambda_k)^{n_k},\ \sum_{i=1}^{k} n_i = n,\ \lambda_i \neq \lambda_j, i \neq j.$$

Wtedy A jest podobna do macierzy J postaci

$$\begin{pmatrix} J_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & J_k \end{pmatrix},\ J_s = \begin{pmatrix} \lambda_s & \epsilon & 0 & \ldots & 0 \\ \vdots & & \ddots & & \vdots \\ 0 & \ldots & 0 & \lambda_s & \epsilon \\ 0 & \ldots & 0 & 0 & \lambda_s \end{pmatrix},\ \epsilon \in \{0, 1\}.$$

Ponadto, jeśli z_s jest liczbą zer nad przekątną macierzy J_s ($z_s \leq n_s - 1$), to $z_s + 1$ jest liczbą wektorów własnych liniowo niezależnych, skojarzonych z λ_s, np.:

$$\begin{pmatrix} \lambda & 0 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \quad \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 0 \\ 0 & 0 & \lambda \end{pmatrix} \quad \begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}$$

3 wektory *2 wektory* *1 wektor*

W metodach numeryczny na ogół nie interesują nas dokładne wartości wartości własnych, często wystarczy wiedzieć przez co są szacowane. Mówi o tym następujące twierdzenie

Twierdzenie 2.24. *Jeśli A jest rzeczywistą (zespoloną) macierzą kwadratową wymiaru n, to*

$$|\lambda_i(A)| \leq \|A\|,\ i = 1, \ldots, n,$$

gdzie $\|\cdot\|$ jest dowolną normą macierzową zgodną z normą wektorową.

Dowód. Niech λ_i będzie wartością własną macierzy A i niech v_i będzie wektorem własnym odpowiadającym λ_i, zatem

Ostatnia nierówność ze zgodności.

$$|\lambda_i| \cdot \|v_i\| = \|\lambda_i v_i\| = \|Av_i\| \leq \|A\| \cdot \|v_i\|.$$

Dzieląc obustronnie nierówność przez $\|v_i\|$ otrzymujemy tezę. □

Definicja 2.25. *Niech A będzie macierzą kwadratową wymiaru n. Promieniem spektralnym nazywamy*

$$\rho(A) = \max\{|\lambda_i(A)| : \ i = 1, \ldots, n\}.$$

Poprzednie twierdzenie pokazuje, że dla dowolnej normy macierzowej $\|\cdot\|$ zgodnej z normą wektorową mamy

$$\rho(A) \leq \|A\|.$$

Twierdzenie 2.26. *(o wydobywaniu normy.)*
Dana niech będzie macierz $A \in \mathbb{R}^{n\times n}$. Wtedy dla dowolnego $\varepsilon > 0$ istnieje norma $\|\cdot\|'$ wektorowa taka, że

$$\|A\|' \leq \rho(A) + \varepsilon.$$

Dowód. Pokażemy najpierw, że jeśli R jest macierzą nieosobliwą i $\|\cdot\|$ dowolną normą w $\mathbb{R}^n$, to

(11) $\|x\|' = \|Rx\|$ jest normą w $\mathbb{R}^n$.

Oczywiście $\forall x : \|x\|' \geq 0$. Ponieważ macierz R jest nieosobliwa zachodzi też $\|x\|' = 0 \Leftrightarrow x = 0$. Jednorodność i nierówność trójkąta zachodzą, bo $\|\cdot\|$ jest normą

$$\|\alpha x\|' = \|R\alpha x\| = |\alpha| \cdot \|Rx\| = |\alpha| \cdot \|x\|',$$
$$\|x + y\|' = \|R(x + y)\| = \|Rx + Ry\| \leq \|Rx\| + \|Ry\| = \|x\|' + \|y\|'.$$

Teraz, mając (11) wyprowadzimy wzór na szukaną normę.

$$\frac{\|Ax\|'}{\|x\|'} = \frac{\|RAx\|}{\|Rx\|} = \frac{\|RAR^{-1}y\|}{\|y\|}, \ y = Rx$$
$$\|A\|' = \max\left\{\tfrac{\|Ax\|'}{\|x\|'} : x \neq 0\right\} = \max\left\{\tfrac{\|RAR^{-1}y\|}{\|y\|} |\ y \neq 0\right\} = \|RAR^{-1}\|,$$

zatem szukaną normę definiujemy jako

(12) $\|A\|' = \|RAR^{-1}\|, \ \|x\|' = \|Rx\|$.

Dzięki twierdzeniu o postaci kanonicznej Jordana istnieje macierz przejścia P taka, że $A = PJP^{-1}$, gdzie $J = (\alpha_{ij})$. Ustalmy $\varepsilon > 0$ i niech $D = \operatorname{diag}(\varepsilon^0, \ldots, \varepsilon^{n-1})$. Skoro

$$D^{-1}JD = \hat{J} = (\hat{\lambda}_{ij}), \ \hat{\lambda}_{ij} = \alpha_{ij}\varepsilon^{-i+j},$$

to macierz $\hat{J}$ nad przekątną ma „0" lub „ε".

$$J = D\hat{J}D^{-1},$$
$$A = PJP^{-1} = PD\hat{J}D^{-1}P^{-1} = PD\hat{J}(PD)^{-1},$$

więc przyjmując $R = (PD)^{-1}$ mamy, że $A = R^{-1}\hat{J}R$. Niech $\|x\|' = \|Rx\|_\infty$. Na mocy (11) i (12)

$$\|A\|' = \|RAR^{-1}\|_\infty = \|RR^{-1}\hat{J}RR^{-1}\|_\infty = \|\hat{J}\|_\infty \leq \max\{|\lambda_i| : \ i = 1, \ldots, n\} + \varepsilon$$
$$= \rho(A) + \varepsilon,$$

co należało wykazać. □

Twierdzenie 2.27. *Niech A będzie rzeczywistą (zespoloną) macierzą kwadratową wymiaru n. Wówczas*

$$\lim_{k\to\infty} A^k = 0 \Leftrightarrow \rho(A) < 1.$$

Dowód. Załóżmy, że istnieje wartość własna λ macierzy A taka, że $|\lambda| \geq 1$. Niech v będzie wektorem własnym odpowiadającym wartości λ. Zatem

$$Av = \lambda v,$$
$$A^2 v = A \cdot Av = A\lambda v = \lambda Av = \lambda^2 v.$$

Zatem indukcyjnie można wykazać, że

$$A^k v = \lambda^k v.$$

Więc

$$\|\lambda^k v\| = \|A^k v\| \leq \|A^k\| \cdot \|v\|,$$
$$\|\lambda^k v\| = |\lambda|^k \|v\| \geq \|v\|,$$
$$\|v\| \leq \|A^k\| \cdot \|v\|,$$
$$1 \leq \|A^k\| \underset{k\to\infty}{\to} 0,$$

sprzeczność.

Załóżmy teraz, że $\rho(A) < 1$. Niech $\varepsilon > 0$ będzie taki, że $q = \rho(A) + \varepsilon < 1$. Na mocy twierdzenia o wydobywaniu normy, dla tak dobranego ε, istnieje $\|\cdot\|$ – norma wektorowa w $\mathbb{R}^n$ taka, że $\|A\| \leq q$. Na mocy submultiplikatywności normy wiemy, że $\|A^k\| \leq \|A\|^k$. Zatem

$$0 \leq \|A^k\| \leq \|A\|^k \leq (\rho(A) + \varepsilon)^k.$$

Z twierdzenia o trzech ciągach $\lim_{k\to\infty} \|A^k\| = 0$, a więc $\lim_{k\to\infty} A^k = 0$, co kończy dowód. □

Wniosek 2.28. *Jeżeli $\rho(A) < 1$, to szereg von Naumanna $\sum_{k=0}^{\infty} A^k$ jest zbieżny i jego suma wynosi*

$$\sum_{k=0}^{\infty} A^k = (I - A)^{-1}.$$

Redukcja wyrazów

Dowód. Niech $S_n = \sum_{k=0}^{n} A^k$, wówczas

$$S_n(I - A) = S_n - S_n A = \sum_{k=0}^{n} A^k - \sum_{k=0}^{n} A^{k+1} = I - A^{n+1}.$$

Skoro $\rho(A) < 1$, to $\lambda_i(I - A) > 0$, więc $I - A$ jest macierzą odwracalną. Na mocy poprzedniego twierdzenia

$$\sum_{k=0}^{\infty} A^k = \lim_{n\to\infty} S_n = \lim_{n\to\infty} (I - A^{n+1})(I - A)^{-1} = (I - A)^{-1}.$$

□

Wniosek 2.29. *Szereg* $\sum_{k=0}^{\infty} a_k A^k$ *jest zbieżny, jeśli* $\rho(A) < r$*, gdzie* r *jest promieniem zbieżności szeregu potęgowego* $f(z) = \sum_{k=0}^{\infty} a_k z^k$.

Dowód. Podobnie jak poprzednio, niech $S_n = \sum_{k=0}^{n} a_k A^k$. Na mocy twierdzenia o wydobywaniu normy (tw. 2.26) znajdziemy $\varepsilon > 0$ oraz normę $\|\cdot\|$ takie, że

$$\|A\| \leq \rho(A) + \varepsilon < r.$$

Wtedy, dla $q = \rho(A) + \varepsilon$, otrzymujemy

$$\|S_n\| \leq \sum_{k=0}^{n} |a_k| \cdot \|A\|^k \leq \sum_{k=0}^{n} |a_k| q^k.$$

Więc $\sum_{k=0}^{\infty} q^k$ jest majorantą szeregu $\sum_{k=0}^{\infty} A^k$, a zatem dostajemy zbieżność dla $\rho(A) < r$. □

Zauważmy, że

$$\sum_{k=0}^{\infty} A^k = f(A),$$

gdzie $f(z) = \sum_{k=0}^{\infty} z^k$. Zatem

$$\sum_{k=0}^{\infty} \frac{(At)^k}{k!} = e^{At}.$$

2.3.2 Podstawowe wiadomości dotyczące metod iteracyjnych.

Zajmijmy się innym podejściem do rozwiązywania układu

(13) $Ax = b,$

gdzie $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ jest macierzą nieosobliwą, $b = (b_i) \in \mathbb{R}^n$, $x = (x_j)$. Szukamy $\bar{x} = A^{-1}b$. Będziemy tworzyć ciąg $\{x_k\} \subset \mathbb{R}^n$ w ten sposób, że

(14) $x_{k+1} = F_k(x_k, x_{k-1}, \ldots, x_{k-s}),\ k = 0, 1, \ldots$

Dane początkowe

(gdzie $x_{-s}, x_{-s+1}, \ldots, x_0$ są znane), oraz

(15) $\lim_{k\to\infty} x_k = \bar{x}.$

Oczywiście wybór F_k jest równoważny wyborowi metody iteracyjnej.

Definicja 2.30. *Metoda (15) jest zbieżna, gdy* $\lim_{k\to\infty} x_k = \bar{x}$ *dla dowolnych danych początkowych* $x_{-s}, x_{-s+1}, \ldots, x_0$.

Terminologia 2.31.

- ▷ *Metodę (15) nazywamy* ***wielokrokową metodą niestacjonarną*** *(a dokładniej* $(s+1)$*-krokową).*
- ▷ *Jeżeli* $F_k \equiv F$*, to metodę nazywamy* ***stacjonarną.***
- ▷ *Jeżeli* $x_{k+1} = F_k(x_k)$*, to metodę nazywamy* ***niestacjonarną jednokrokową.***
- ▷ *Jeżeli* F_k *jest liniowa lub afiniczna, to metodę nazywamy* ***liniową.***

Postać ogólna niestacjonarnej metody liniowej:

$$x_{k+1} = B_k x_k + C_k,\ k = 0, 1, \ldots \tag{16}$$

$x_0 \in \mathbb{R}^n$ nazywamy **przybliżeniem początkowym**. Każda metoda *musi* spełniać

$$\bar{x} = B_k \bar{x} + C_k,\ k = 0, 1, \ldots, \tag{17}$$

to znaczy $\bar{x}$ musi być punktem stałym odwzorowania $x \longmapsto B_k x + C_k$.

Twierdzenie 2.32. *(o zbieżności ogólnej niestacjonarnej metody liniowej.)*
Warunkiem dostatecznym zbieżności iteracji (16) jest istnienie stałej $0 \leq q < 1$ *takiej, że dla dowolnego kroku* k

$$\|B_k\| \leq q,$$

gdzie $\|\cdot\|$ *jest pewną normą macierzową zgodną z normą wektorową.*

Dowód. Niech $e_k = x_k - \bar{x}$ będzie błędem k-tego przybliżenia, wówczas

$$\begin{aligned} e_{k+1} &= x_{k+1} - \bar{x} = B_k x_k + C_k - \bar{x} = B_k x_k + C_k - B_k \bar{x} - C_k \\ &= B_k(x_k - \bar{x}) = B_k e_k = \ldots = B_k B_{k-1} \ldots B_0 e_0. \end{aligned}$$

Zatem, korzystając ze zgodności, otrzymujemy

$$0 \leq \|e_{k+1}\| = \|B_k B_{k-1} \ldots B_0 e_0\| \leq \|B_k\| \cdot \ldots \cdot \|B_0\| \cdot \|e_0\| \leq q^{k+1} \|e_0\|.$$

Stąd, na mocy twierdzenia o trzech ciągach, mamy

$$\lim_{k \to \infty} \|e_k\| = 0,$$

co jest równoważne zbieżności metody (16). □

2.3.3 Metody stacjonarne.

Rozważmy teraz iterację postaci

$$x_{k+1} = Bx_k + C, k = 0, 1, \ldots,\ x_0 \in \mathbb{R}^n. \tag{18}$$

Twierdzenie 2.33. *(warunek dostateczny zbieżności.)*
Jeżeli $\|B\| < 1$ *dla pewnej normy macierzowej zgodnej z normą wektorową , to iteracja (18) jest zbieżna.*

Dowód. Wyniki z poprzedniego twierdzenia dla $B_k = B,\ k \geq 0$. □

Twierdzenie 2.34. *(warunek konieczny i wystarczający zbieżności.)*
x_0 *jest dowolne.* *Iteracja (18) jest zbieżna wtedy i tylko wtedy, gdy* $\rho(B) < 1$.

Dowód. Podobnie jak w dowodzie twierdzenia 2.32 definiujemy $e_k = x_k - \bar{x}$. Zatem $e_{k+1} = B^{k+1}e_0$. W takim razie

$$\lim_{k\to\infty} e_k = 0 \Leftrightarrow \lim_{k\to\infty} B^k = 0 \overset{\text{tw } 2.27}{\Leftrightarrow} \rho(B) < 1.$$

□

2.3.4 Metody Gaussa-Seidla i Jacobiego.

Dana niech będzie macierz $A = (a_{ij})$, oraz jej rozkład $A = L + D + U$, gdzie L jest macierzą trójkątną dolną z zerami na przekątnej, U macierzą trójkątną górną z zerami na przekątnej, $D = \operatorname{diag}(a_{11}, \ldots, a_{nn})$ macierzą diagonalną. Zapiszmy układ $Ax = b$ w postaci skalarnej

$$\begin{cases} a_{11}x_1 + \ldots + a_{1n}x_n = b_1 \\ \ldots \\ a_{n1}x_1 + \ldots + a_{nn}x_n = b_n \end{cases}$$

i załóżmy, że $a_{ii} \neq 0$[7] dla każdego $i = 1, \ldots, n$. Niech $x^{(k)} = (x_1^{(k)}, \ldots, x_n^{(k)})^T$ oznacza k-te przybliżenie.

Metoda Jacobiego korzysta z iteracji

$$x_i^{(k+1)} = -\tfrac{1}{a_{ii}} \sum_{j\neq i}^{n} a_{ij}x_j^{(k)} + \tfrac{b_i}{a_{ii}},\ i = 1, \ldots, n,$$

czyli

$$x^{(k+1)} = -D^{-1}(L + U)x^{(k)} + D^{-1}b. \tag{19}$$

Metoda Gaussa-Seidla różni się nieznacznie i ma postać:

$$x_i^{(k+1)} = -\tfrac{1}{a_{ii}}\Big(\sum_{j<i} a_{ij}x_j^{(k+1)} + \sum_{j>i} a_{ij}x_j^{(k)} \Big) + \tfrac{b_i}{a_{ii}},\ i = 1, \ldots, n,$$
$$(D + L)x^{(k+1)} = -Ux^{(k)} + b,$$

$$x^{(k+1)} = -(D + L)^{-1}Ux^{(k)} + (D + L)^{-1}b. \tag{20}$$

Twierdzenie 2.35. *Jeżeli macierz* $A = (a_{ij})_{i,j=1}^n$ *spełnia jeden z warunków*

(i) $|a_{ii}| > \sum_{j\neq i} |a_{ij}|,\ i = 1, \ldots, n$ *(mocne kryterium sumy wierszy);*

[7] Ze względu na nieosobliwość macierzy A, po odpowiednich permutacjach zawsze można uzyskać taki efekt.

(ii) $|a_{jj}| > \sum_{i \neq j} |a_{ij}|,\ j = 1, \ldots, n$ *(mocne kryterium sumy kolumn),*

to metody Gaussa-Seidla i Jacobiego są zbieżne (w szczególności macierz A *jest nieosobliwa).*

Dowód. Zacznijmy od pokazania, że mocne kryterium sumy wierszy (kolumn) pociąga za sobą nieosobliwość macierzy A (z czego skorzystamy pokazując zbieżność metody Gaussa-Seidla). Dla dowodu nie wprost załóżmy, że zachodzi warunek *(i)* i A jest osobliwa. Zatem istnieje $x \neq 0$ taki, że $Ax = 0$. Załóżmy, że $|x_k| = \|x\|_\infty$, wówczas

$$\sum_{j=1}^{n} a_{kj} x_j = 0,$$

$$a_{kk} x_k = -\sum_{j \neq k} a_{kj} x_j,$$

$$|a_{kk}| \cdot |x_k| \leq \sum_{j \neq k} |a_{kj}| \cdot |x_j|,$$

$$|a_{kk}| \leq \sum_{j \neq k} |a_{kj}| \frac{|x_j|}{|x_k|} \leq \sum_{j \neq k} |a_{kj}|,$$

sprzeczność. Jeśli A spełnia *(ii)*, to A^T spełnia *(i)* i również dochodzimy do sprzeczności, zatem macierz A jest nieosobliwa.

Pokażmy teraz zbieżność metody Jacobiego. Niech $B_J = -D^{-1}(L + U)$, zatem $b_{ij} = -\frac{a_{ij}}{a_{ii}}$ dla $i \neq j$, oraz $b_{ii} = 0$. Na mocy twierdzenia 2.33 wystarczy wykazać, że dla pewnej normy macierzowej $\|\cdot\|$ zgodnej z normą wektorową zachodzi $\|B_J\| < 1$. Jeżeli zachodzi *(i)*, to

$$\sum_{j=1}^{n} |b_{ij}| = \sum_{j \neq i} \frac{|a_{ij}|}{|a_{ii}|} < 1,\ i = 1, \ldots, n \Rightarrow \|B_J\|_\infty < 1.$$

Jeśli natomiast zachodzi *(ii)*, to

$$\sum_{i=1}^{n} |b_{ij}| = \sum_{i \neq j} \frac{|a_{ij}|}{|a_{jj}|} < 1,\ j = 1, \ldots, n \Rightarrow \|B_J\|_1 < 1.$$

Przejdźmy teraz do metody Gaussa-Seidla. Niech $B_{GS} = -(D + L)^{-1}U$, chcemy pokazać, że $\rho(B_{GS}) < 1$, co na mocy twierdzenia 2.34 zakończy dowód. Przypuśćmy, że istnieje wartość własna μ macierzy B_{GS} taka, że $|\mu| \geq 1$. Zatem macierz $B_{GS} - \mu I$ jest osobliwa. Więc osobliwa jest też $-(D + L)^{-1}U - \mu I = -\mu(D + L)^{-1}((D + L) + \frac{1}{\mu}U)$. Wiemy, że $D + L$ jest nieosobliwa, więc osobliwa musi być macierz $(D + L) + \frac{1}{\mu}U$. Ale dzięki *(i)* mamy

$$|a_{ii}| > \sum_{j<i} |a_{ij}| + \sum_{j>i} |a_{ij}| \geq \sum_{j<i} |a_{ij}| + \frac{1}{|\mu|} \sum_{j>i} |a_{ij}|,\ i = 1, \ldots, n,$$

co oznacza, że macierz $(D+L)+\frac{1}{\mu}U$ spełnia mocne kryterium sumy wierszy. Zatem na mocy tego co pokazaliśmy na początku dowodu, $(D+L)+\frac{1}{\mu}U$ jest macierzą nieosobliwą – sprzeczność. □

Definicja 2.36. *Macierz* $A = (a_{ij})$ *nazywamy nieredukowalną, jeśli nie istnieje macierz permutacji* P *taka, że*

$$PAP^{-1} = \begin{pmatrix} B_{11} & B_{12} \\ 0 & B_{22} \end{pmatrix}.$$

Widać więc, że macierz A jest nieredukowalna, jeśli nie istnieje zbiór $J \subsetneq \{1, \ldots, n\}$ taki, że

$$\forall i \in J \; \forall j \notin J : \; a_{ij} = 0.$$

Redukowalność macierzy A jest równoważna temu, że układ $Ax = b$ jest równoważny układowi

$$\begin{cases} B_{11}y_1 + B_{12}y_2 = c_1 \\ \qquad\quad B_{22}y_2 = c_2. \end{cases}$$

Okazuje się, że nieredukowalność macierzy A jest silnie związana z grafem z nią skojarzonym, wprowadźmy więc kilka pojęć dotyczących grafów.

Grafem (zorientowanym, skierowanym) G nazywamy parę (P, V), gdzie $P = \{P_1, \ldots, P_n\}$ jest zbiorem skończonym, a V dwuargumentową relacją w P. Zbiór P nazywamy **zbiorem wierzchołków**, a $V \subset P \times P$ **zbiorem krawędzi**.

Drogą długości k z wierzchołka u do wierzchołka u' w grafie $G = (P, V)$ jest ciąg wierzchołków $< P_0, P_1, \ldots, P_k >$ takich, że $P_0 = u, P_k = u'$ i dla $i = 1, \ldots, k$ krawędź $\overline{P_{i-1}P_i} \in V$.

Graf jest **cyklicznie spójny** jeśli dla dowolnych wierzchołków P_i, P_j istnieje droga z wierzchołka P_i do wierzchołka P_j.

Niech $A = (a_{ij})$ będzie macierzą kwadratową $n \times n$. Grafem skojarzonym z macierzą A nazywamy graf zorientowany $G(A) = (P, V)$ taki, że P jest zbiorem n-elementowym oraz dla dowolnych $i, j \in \{1, \ldots, n\} : \; a_{ij} \neq 0 \Leftrightarrow \overline{P_iP_j} \in V$.

Możemy już zapisać interesującą nas zależność.

Twierdzenie 2.37. *Macierz* $A = (a_{ij})$ *jest nieredukowalna wtedy i tylko wtedy, gdy graf* $G(A)$ *jest cyklicznie spójny.*

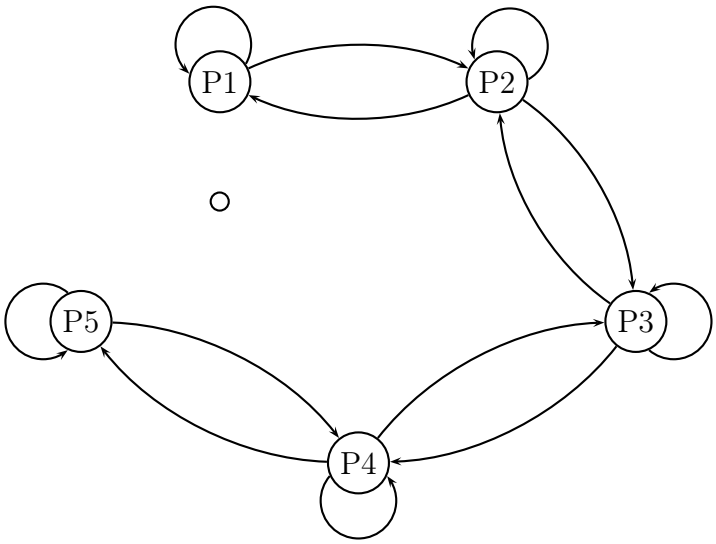
Dowód. Macierz A jest redukowalna wtedy i tylko wtedy, gdy istnieje zbiór $J \subsetneq \{1, \ldots, n\}$ taki, że $a_{ij} = 0$ dla $i \in J, j \notin J$. To jest zaś równoważne temu, że nie istnieje droga od P_k do P_j dla $k \in J, j \notin J$. □

Następny przykład jest klasycznym przykładem macierzy nieredukowalnej. Wykazuje się to stosując poprzednie twierdzenie.

Przykład 2.38. Niech

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

Rysunek 4 przedstawia graf skojarzony z tą macierzą (dla $n = 5$).



Rysunek 4: Graf cyklicznie spójny

Twierdzenie 2.39. *Załóżmy, że macierz* $A = (a_{ij})_{i,j=1}^{n}$ *jest niereduk owalna i spełnia jeden z warunków*

(iii) $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|,\ i = 1, \ldots, n$ *oraz istnieje* i_0 *takie, że mamy ostrą nierówność (słabe kryterium sumy wierszy);*

(iv) $|a_{jj}| \geq \sum_{i \neq j} |a_{ij}|,\ j = 1, \ldots, n$ *oraz istnieje* j_0 *takie, że mamy ostrą nierówność (słabe kryterium sumy kolumn).*

Wtedy metody Gaussa-Seidla i Jacobiego są zbieżne (w szczególności macierz A *jest nieosobliwa).*

Dowód. Pokażemy najpierw, że słabe kryterium sumy wierszy pociąga nieosobliwość macierzy (analogicznie dla słabego kryterium sumy kolumn).

Załóżmy więc, że A spełnia słabe kryterium sumy wierszy. Istnieje więc i_0 taki, że

$$|a_{i_0 i_0}| > \sum_{j \neq i_0} |a_{i_0 j}|.$$

Załóżmy, że istnieje $x \neq 0$ taki, że $Ax = 0$. Niech

$$J = \{k :\ |x_k| \geq |x_i|,\ i = 1, \ldots, n, \text{ oraz } |x_k| > |x_j| \text{ dla pewnego } j\}.$$

Twierdzimy, że $J \neq \emptyset$. Załóżmy, że J jest zbiorem pustym, więc $\forall i, k\ |x_k| = |x_i|$, zatem $\|x\|_\infty = |x_i|, i = 1, \ldots, n$. Mamy więc

$$\sum_{j=1}^{n} a_{i_0 j} x_j = 0,$$

$$a_{i_0 i_0} x_{i_0} = -\sum_{j \neq i_0} a_{i_0 j} x_j,$$

$$|a_{i_0 i_0}| \cdot \|x\|_\infty \leq \|x\|_\infty \sum_{j \neq i_0} |a_{i_0 j}|,$$

$$|a_{i_0 i_0}| \leq \sum_{j \neq i_0} |a_{i_0 j}|,$$

co jest sprzeczne ze słabym kryterium sumy wierszy. Niech więc $k \in J$, wówczas

$$a_{kk}x_k = -\sum_{j \neq k} a_{kj}x_j,$$
$$|a_{kk}| \leq \sum_{j \neq k} |a_{kj}| \frac{|x_j|}{|x_k|}.$$

Z definicji zbioru J widać, że $|x_k| > |x_j|$ wtedy i tylko wtedy, gdy $j \notin J$. Zatem, skoro A spełnia (iii), to w ostatniej nierówności zachodzić musi równość, więc spełniony musi być warunek

$$\forall k \in J, \forall j \notin J : a_{kj} = 0,$$

co jest sprzeczne z nieredukowalnością macierzy A.

Jeżeli macierz A spełnia (iv), to A^T spełnia (iii). Więc A jest nieosobliwa (bo A^T jest nieosobliwa).

Wykażmy teraz, że metoda Jacobiego jest zbieżna, a więc że dla macierzy $B_J = -D^{-1}(L+U)$ mamy $\rho(B_J) < 1$. Dla dowodu nie wprost załóżmy, że $\rho(B_J) \geq 1$. Istnieje więc wartość własna $|\lambda| \geq 1$. Zatem macierz $B_J - \lambda I$ (a tym samym $\lambda I - B_J$) jest osobliwa. Skoro

$$\lambda I - B_J = \lambda I + D^{-1}(L+U) = \lambda D^{-1}\big(D + \tfrac{1}{\lambda}(L+U)\big),$$

oraz macierz D^{-1} jest nieosobliwa, to macierz $D + \frac{1}{\lambda}(L+U)$ jest macirzą osobliwą. Na mocy założenia $A = D + L + U$ jest nieredukowalna, a więc macierz $D + \frac{1}{\lambda}(L+U)$ też jest nieredukowalna. Z drugiej strony, skoro

$$\tfrac{1}{|\lambda|} \leq 1,$$

to jeśli A spełnia (iii) (ewentualnie (iv)), to $D + \frac{1}{\lambda}(L+U)$ też, jest więc (na mocy pierwszej części dowodu) nieosobliwa.

Pozostało wykazać, że metoda Gaussa-Seidla jest zbieżna. Niech więc μ będzie wartością własną macierzy B_{GS} taką, że $|\mu| \geq 1$. Zatem macierz $I - \frac{1}{\mu}B_{GS}$ jest macierzą osobliwą (bo $\mu I - B_{GS}$ jest osobliwa). Ponieważ

$$I - \tfrac{1}{\mu}B_{GS} = I + \tfrac{1}{\mu}(D+L)^{-1}U = (D+L)^{-1}(D + L + \tfrac{1}{\mu}U)$$

i $(D+L)^{-1}$ jest nieosobliwa, zatem $D + L + \frac{1}{\mu}U$ jest macierzą osobliwą. Jest ona nieredukowalna i spełnia słabe kryterium sumy wierszy (kolumn), a więc podobnie jak wcześniej jest nieosobliwa – sprzeczność. Zatem metoda Gaussa-Seidla jest zbieżna. □

2.3.5 Metoda kolejnych nadrelaksacji (SOR - succesive overrelaxation).

Zapoznamy się teraz z kolejną metodą rozwiązywania układu (13). Niech

$$x^{(k)} = \begin{pmatrix} x_1^{(k)} \\ \vdots \\ x_n^{(k)} \end{pmatrix}$$

będzie kolejnym przybliżeniem danym wzorem

$$(21)\quad x_i^{(k+1)} = x_i^{(k)} + \omega(\overline{x}_i^{(k+1)} - x_i^{(k)}),$$

gdzie $\overline{x}_i^{(k+1)}$ jest i-tą współrzędną $k+1$ iteracji (wektora x^{k+1}) otrzymaną metodą Gaussa-Seidla, ω jest parametrem liczbowym. Widać, że jeśli $\omega = 1$, to $x_i^{(k+1)} = \overline{x}_i^{(k+1)}$. Zatem

$$\begin{aligned}
a_{ii}x_i^{(k+1)} &= a_{ii}x_i^{(k)} + \omega\big(-\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} + b_i - a_{ii}x_i^{(k)}\big) \\
&= (1-\omega)a_{ii}x_i^{(k)} + \omega\big(-\sum_{j<i} a_{ij}x_j^{(k+1)} - \sum_{j>i} a_{ij}x_j^{(k)} + b_i\big), \\
(D+\omega L)x^{(k+1)} &= ((1-\omega)D - \omega U)x^{(k)} + \omega b, \\
x^{(k+1)} &= (D+\omega L)^{-1}((1-\omega)D - \omega U)x^{(k)} + (D+\omega L)^{-1}\omega b,
\end{aligned}$$

zatem

$$(22)\quad x^{(k+1)} = B(\omega)x^{(k)} + C,$$

gdzie

$$B(\omega) = (D+\omega L)^{-1}((1-\omega)D - \omega U), C = \omega(D+\omega L)^{-1}b.$$

Zastanówmy się, jak dobrać ω tak, aby metoda ta była zbieżna, a więc żeby $\rho(B(\omega)) < 1$? Częściową odpowiedź daje nam

Twierdzenie 2.40. *(Kahan)*
Dla dowolnej macierzy $A \in \mathbb{C}^{n\times n}$ zachodzi nierówność

$$\rho(B(\omega)) \geq |1-\omega|.$$

Dowód. Niech $\varphi(\lambda)$ będzie wielomianem charakterystycznym macierzy $B(\omega)$. Wówczas, z definicji 2.25 promienia spektralnego

$$|\lambda_1 \cdot \ldots \cdot \lambda_n| \leq (\rho(B(\omega)))^n.$$

Ale z drugiej strony mamy

$$\begin{aligned}
(-1)^n\lambda_1 \cdot \ldots \cdot \lambda_n &= \varphi(0) = \det B(\omega), \\
|\lambda_1 \cdot \ldots \cdot \lambda_n| &= |\det(B(\omega))| = |\det(D+\omega L)^{-1}((1-\omega)D - \omega U)| \\
&= \left|\tfrac{1}{a_{11}\ldots a_{nn}}(1-\omega)^n a_{11}\ldots a_{nn}\right| = |1-\omega|^n,
\end{aligned}$$

co kończy dowód. □

Twierdzenie to daje nam natychmiastowo następujący wniosek

Wniosek 2.41. *Jeżeli metoda SOR jest zbieżna, to $\omega \in (0,2)$.*

Dowód. Z założenia i poprzedniego twierdzenia mamy

$$1 > \rho(B(\omega)) \geq |1 - \omega|,$$

zatem $\omega \in (0, 2)$. □

Twierdzenie 2.42. *(Zbieżność metody SOR.)*
Jeśli $A = A^\star > 0$, to metoda SOR jest zbieżna dla każdego $\omega \in (0, 2)$. W szczególności metoda Gaussa-Seidla jest zbieżna dla każdej macierzy $A = A^\star > 0$.

Dowód. Niech

$$Q = A^{-1}(2(D + \omega L) - \omega A) = 2A^{-1}(D + \omega L) - \omega I.$$

Pokażemy, że

(*i*) Wartości własne macierzy Q leżą w prawej półpłaszczyźnie, tzn. $\Re\lambda_j(Q) > 0 \ \forall j$,

(*ii*) $B(\omega) = (Q + \omega I)^{-1}(Q - \omega I)$,

(*iii*) $\rho(B(\omega)) < 1$.

Niech λ będzie wartością własną macierzy Q i niech $x \neq 0$ będzie wektorem własnym skojarzonym z λ, zatem

$$\begin{aligned} &Qx = \lambda x, \\ &A^{-1}(2(D + \omega L) - \omega A)x = \lambda x, \\ &(2(D + \omega L) - \omega A)x = \lambda A x, \end{aligned}$$

więc

$$x^*(2(D + \omega L) - \omega A)x = \lambda x^* A x. \tag{23}$$

Sprzęgając po hermitowsku obustronnie otrzymamy

$$x^*(2(D + \omega L^*) - \omega A)x = \bar{\lambda} x^* A x. \tag{24}$$

Teraz dodając stronami (23) i (24), a następnie korzystając z tego, że $A = D + L + L^*$ (A samosprzężona, zatem $L^* = U$) mamy

$$\begin{aligned} &x^*(4D + 2\omega(L + L^\star - A))x = (\lambda + \bar{\lambda})x^* A x, \\ &2x^*(2D - \omega D)x = 2\Re\lambda x^* A x, \\ &(2 - \omega)x^* D x = \Re\lambda x^* A x. \end{aligned}$$

Ponieważ A jest dodatnio określona, to $a_{ii} > 0$, zatem $x^* D x > 0$. Z założenia $2-\omega > 0$, więc $\Re\lambda > 0$. Wykazaliśmy więc punkt pierwszy.

Aby wykazać (*ii*) zauważmy, że z definicji macierzy Q mamy

$$\begin{aligned} (Q + \omega I)^{-1}(Q - \omega I) &= (2A^{-1}(D + \omega L))^{-1}(2A^{-1}(D + \omega L) - 2\omega I) \\ &= (D + \omega L)^{-1}(D + \omega L - \omega A) \\ &= (D + \omega L)^{-1}((1 - \omega)D - \omega U) = B(\omega). \end{aligned}$$

Przejedźmy teraz do punktu (*iii*). Niech μ będzie wartością własną macierzy $B(\omega)$. Zatem

$$\exists x \neq 0 : \; B(\omega)x = \mu x.$$

Dzięki (*ii*)

$$\begin{aligned}
&(Q + \omega I)^{-1}(Q - \omega I)x = \mu x,\\
&(Q - \omega I)x = \mu(Q + \omega I)x,\\
&(1 - \mu)Qx = \omega(1 + \mu)x.
\end{aligned}$$

W takim razie ($x \neq 0 \Rightarrow \mu \neq 1$) $\lambda = \omega\frac{1+\mu}{1-\mu}$ jest wartością własną macierzy Q. Oczywiście $\mu = \frac{\lambda-\omega}{\lambda+\omega}$, zatem

$$|\mu|^2 = \mu\overline{\mu} = \frac{(\lambda-\omega)\overline{(\lambda-\omega)}}{(\lambda+\omega)\overline{(\lambda+\omega)}} = \frac{\lambda\overline{\lambda} - \omega\overline{\lambda} - \omega\lambda + \omega^2}{\lambda\overline{\lambda} + \omega\overline{\lambda} + \omega\lambda + \omega^2} = \frac{|\lambda|^2 - 2\omega\Re\lambda + \omega^2}{|\lambda|^2 + 2\omega\Re\lambda + \omega^2} \overset{(i)}{<} 1.$$

Druga część tezy wynika z (22):

$$B(1) = (D + L)^{-1}(-U) = B_{GS}.$$

□

2.3.6 Metoda Richardsona.

Ostatnią metodą iteracyjną rozwiązywania układu równań postaci (1) jaką się zajmiemy jest metoda Richardsona.

Układ (1) jest równoważny temu, że dla dowolnego parametru $\alpha \in \mathbb{R}$

$$x = x - \alpha(Ax - b).$$

W oparciu o tą uwagę zdefiniujmy tzw. **iterację Richardsona**:

$$x_{k+1} = x_k - \alpha(Ax_k - b). \tag{25}$$

W naszych rozważaniach wygodniejsza będzie postać równoważna

$$x_{k+1} = (I - \alpha A)x_k + \alpha b,$$

która jak widać jest postaci (18). Zauważmy, że dla $\alpha \neq 0$

$$\bar{x} = (I - \alpha A)\bar{x} + \alpha b \Leftrightarrow \bar{x} = A^{-1}b.$$

Zatem, jedynym problemem w tej metodzie jest taki dobór α, aby $\rho(B_R(\alpha)) < 1$, gdzie $B_R(\alpha) = I - \alpha A$. Niech μ będzie wartością własną macierzy $B_R(\alpha)$, zatem

$$\begin{aligned}
&\det(B_R(\alpha) - \mu I) = 0,\\
&\det(I - \alpha A - \mu I) = 0,\\
&\det\big[-\alpha\big(A - \tfrac{1-\mu}{\alpha}I\big)\big] = 0,\\
&(-\alpha)^n \det\big(A - \tfrac{1-\mu}{\alpha}I\big) = 0,\\
&\det\big(A - \tfrac{1-\mu}{\alpha}I\big) = 0.
\end{aligned}$$

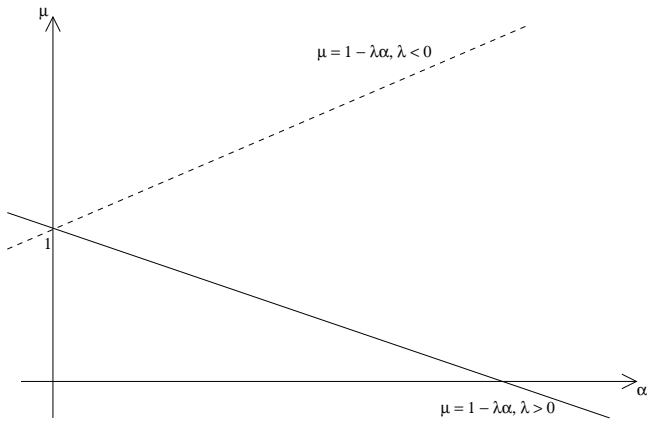
Zatem, μ_i jest wartością własną macierzy $B_R(\alpha)$ wtedy i tylko wtedy, gdy $\lambda_i = \frac{1-\mu_i}{\alpha}$ jest wartością własną macierzy A. Niech więc λ_i $(i = 1, \ldots, n)$ będą wartościami własnymi macierzy A. Wówczas

$$\mu_i = 1 - \alpha\lambda_i$$

są wartościami własnymi macierzy $B_R(\alpha)$. Tak więc warunek $\rho(B_R(\alpha)) < 1$ jest równoważny warunkowi

(26) $\max\{|1 - \alpha\lambda_{max}|, |1 - \alpha\lambda_{min}|\} < 1.$

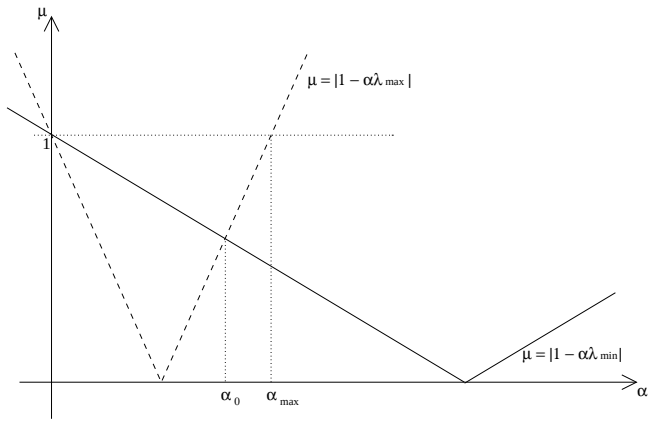
Zatem dla istnienia takiego α potrzeba, aby wszystkie wartości własne macierzy A były tego samego znaku, czyli $A = A^T > 0$ albo $A = A^T < 0$.



Rysunek 5:

Przyjmijmy, że tak jest. Wówczas zbiór z którego możemy wybrać α jest postaci $(0, \alpha_{max})$ (albo $(\alpha_{min}, 0)$). Optymalne α, oznaczone jako α_0, spełnia warunek

$$|1 - \alpha_0\lambda_{max}| = |1 - \alpha_0\lambda_{min}|.$$



Rysunek 6:

Zatem

$$\alpha_0 = \frac{2}{\lambda_{min} + \lambda_{max}}.$$

Wówczas

$$\rho(B_R(\alpha_0)) = 1 - \lambda_{min}\alpha_0 = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

Udowodniliśmy w ten sposób następujące twierdzenie:

Twierdzenie 2.43. *Jeżeli macierz* A *jest symetryczna i dodatnio (ujemnie) określona, to dla*

$$\alpha_0 = \frac{2}{\lambda_{min} + \lambda_{max}}$$

metoda Richardsona jest zbieżna, oraz

$$\rho(B_R(\alpha_0)) = \frac{\lambda_{max} - \lambda_{min}}{\lambda_{max} + \lambda_{min}}.$$

Zajmijmy się wskaźnikiem uwarunkowania dla tej metody. Skoro $A = A^T > 0$, to na mocy wniosku 2.17

$$\|A\|_2 = \lambda_{max}(A),$$

$\lambda(A^{-1}) = 1/\lambda(A)$

zatem

$$\mathrm{cond}(A) = \|A\|_2\|A^{-1}\|_2 = \tfrac{\lambda_{max}}{\lambda_{min}}.$$

Tak więc

$$\rho(B_R(\alpha_0)) = \frac{\frac{\lambda_{max}}{\lambda_{min}} - 1}{\frac{\lambda_{max}}{\lambda_{min}} + 1} = \frac{\mathrm{cond}(A) - 1}{\mathrm{cond}(A) + 1}.$$

Otrzymaliśmy, że jeśli $\mathrm{cond}(A)$ jest duże, to $\rho(B_R(\alpha_0)) \approx 1$.

Bez dowodu

Twierdzenie 2.44. *(Stein-Rosenberg)*
Jeśli macierz $A = (a_{ij})$ *spełnia warunek*

(1) $a_{ii} > 0$ *dla* $i = 1, \ldots, n$,

(2) $a_{ij} \leq 0$ *dla* $i \neq j$,

to zachodzi jeden z wykluczających się warunków

(i) $\rho(B_J) = \rho(B_{GS}) = 0$,

(ii) $0 < \rho(B_{GS}) < \rho(B_J) < 1$,

(iii) $\rho(B_{GS}) = \rho(B_J) = 1$,

(iv) $1 < \rho(B_J) < \rho(B_{GS})$.

2.4 Metody gradientowe.

Ponownie rozważmy układ (1). Przez $\bar{x}$ oznaczmy rozwiązanie tego układu. Metody gradientowe opierają się na obserwacji, że wyznaczenie rozwiązania układu (1) równoważne jest problemowi wyznaczenia

(27) $\min\{\tilde{\varphi}(x) : x \in \mathbb{R}^n\}$,

gdzie

(28) $\tilde{\varphi} : \mathbb{R}^n \ni x \mapsto \|Ax - b\|_R^2 \in \mathbb{R}$,

$$R = R^T > 0,\ \|u\|_R = \sqrt{u^T R u}.$$

Dla funkcji tej zachodzi

$$\lim_{\|x\| \to \infty} \tilde{\varphi}(x) = \infty.$$

Oczywiście $\tilde{\varphi}(x) > 0$ dla $Ax - b \neq 0$, oraz

$$\tilde{\varphi}(x) = 0 \Leftrightarrow b - Ax = 0 \Leftrightarrow x = \bar{x}.$$

2.4.1 Metoda najmniejszych kwadratów.

Przyjmijmy w (28) $R = I$, wówczas

$$\tilde{\varphi}(x) = (Ax - b)^T(Ax - b) = x^T A^T A x - 2x^T A^T b + b^T b.$$

Skoro $\tilde{x}$ realizuje $\min \tilde{\varphi}$ jeśli $\operatorname{grad}\tilde{\varphi}(\tilde{x}) = 0$, to $\tilde{x}$ jest rozwiązaniem problemu (27) wtedy, gdy $A^T A\tilde{x} = A^T b$.

2.4.2 Uogólniona metoda najmniejszych kwadratów.

Niech $A = A^T > 0$. Przyjmijmy $R = A^{-1}$. Wówczas

$$\tilde{\varphi}(x) = x^T A x - 2b^T x + b^T A^{-1} b.$$

Zatem, jeśli

(29) $\varphi(x) = x^T A x - 2b^T x = \tilde{\varphi}(x) - b^T A^{-1} b$,

to $\tilde{\varphi}(x)$ i $\varphi(x)$ osiągają minimum w tym samy punkcie.

Więc, aby znaleźć rozwiązanie układu (1) wystarczy wyznaczyć minimum funkcji $\varphi(x)$, a więc rozważyć dla niej problem (28). Jest to dość ciężkie, więc zamiast tego rozważymy nieskończony ciąg problemów 1-wymiarowych. Ustalmy przybliżenie początkowe $x_0 \in \mathbb{R}^n$. Szukać będziemy

(30) $\lambda_k = \min\{\varphi(x_k + \lambda d_k) : \lambda \in \mathbb{R}\}$,

gdzie $d_k \in \mathbb{R}^n$ jest zadanym wektorem (kierunkiem), a ciąg $\{x_k\}_k$ dany jest wzorem

(31) $x_{k+1} = x_k + \lambda_k d_k$.

Oczywiście

$$\varphi(x_{k+1}) = \varphi(x_k + \lambda_k d_k),$$

należy zatem wyznaczyć λ_k. Liczymy

$$\begin{aligned}\varphi(x_k + \lambda d_k) &= (x_k + \lambda d_k)^T A(x_k + \lambda d_k) - 2b^T(x_k + \lambda d_k) \\ &= \lambda^2 d_k^T A d_k - 2\lambda d_k^T(b - Ax_k) + \varphi(x_k),\end{aligned}$$

zatem, o ile $d_k^T A d_k \neq 0$, $\varphi(x_k + \lambda d_k)$ jest trójmianem kwadratowym zmiennej λ. Wystarczy więc wyznaczyć λ_k dla którego funkcja

(32) $g_k(\lambda) = \lambda^2 d_k^T A d_k - 2\lambda d_k^T r_k$,

gdzie $r_k = b - Ax_k$ – **residuum przybliżenia** x_k, osiąga minimum (które jest ujemne z definicji funkcji). Jest to oczywiście

(33) $\lambda_k = \dfrac{d_k^T r_k}{d_k^T A d_k}$.

Uwaga 2.45. *Dla* $k = 0, 1, \ldots$ *zachodzi*

$$\varphi(x_{k+1}) < \varphi(x_k).$$

Dowód. Ustalmy k i liczmy

$$\varphi(x_{k+1}) = \varphi(x_k + \lambda_k d_k) = g_k(\lambda_k) + \varphi(x_k) \overset{g_k(\lambda_k)<0}{<} \varphi(x_k).$$

□

Otrzymujemy zatem następujący algorytm. Wybieramy $x_0 \in \mathbb{R}^n$ przybliżenie początkowe oraz $\{d_k\}_k \subset \mathbb{R}^n$ ciąg kierunków. Tworzymy ciąg kolejnych przybliżeń:

$$x_{k+1} = x_k + \lambda_k d_k,\ k = 0, 1, \ldots,$$

gdzie

$$\lambda_k = \frac{d_k^T r_k}{d_k^T A d_k},\ r_k = b - Ax_k.$$

Jako rozwiązanie układu (1) bierzemy

(34) $\bar{x} = \lim\limits_{k\to\infty} x_k$.

Pozostaje więc wykazać, że ta granica istnieje. Wprowadźmy drobną zmianę w definicji ciągu kolejnych przybliżeń, a mianowicie niech

(35) $x_{k+1} = x_k + \beta_k \lambda_k d_k,\ k = 0, 1, \ldots,$

gdzie $\beta_k \in (0, 2)$ są zadane. Dostaliśmy zatem metodę bardziej ogólną niż uogólniona metoda najmniejszych kwadratów, mianowicie **uogólnioną metodę najszybszego spadku**.

Twierdzenie 2.46. *(zbieżność uogólnionej metody najszybszego spadku) Załóżmy, że*

(1) $A = A^T > 0$,

(2) istnieją $\delta_1, \delta_2 > 0$ *takie, że*

(2.1) $0 < \delta_1 \leq \beta_k \leq 2 - \delta_1,\ k = 0, 1, \ldots$

(2.2) $\frac{r_k{}^T d_k}{\|r_k\| \cdot \|d_k\|} \geq \delta_2,\ k = 0, 1, \ldots$[8]

wówczas

$$\lim_{k\to\infty} x_k = \bar{x},$$

gdzie $\{x_k\}_k$ *jest określony wzorem (35).*

Dowód. Skoro

$$\varphi(x_{k+1}) = \varphi(x_k + \beta_k\lambda_k d_k) = (\beta_k\lambda_k)^2 d_k^T A d_k - 2\beta_k\lambda_k r_k^T d_k + \varphi(x_k),$$

więc

$$\varphi(x_{k+1}) - \varphi(x_k) = (\beta_k\lambda_k)^2 d_k^T A d_k - 2\beta_k\lambda_k r_k^T d_k,$$

$$\varphi(x_{k+1}) - \varphi(x_k) \overset{(33)}{=} \beta_k^2 \frac{(d_k^T r_k)^2}{(d_k^T A d_k)^2} d_k^T A d_k - 2\beta_k \frac{d_k^T r_k}{d_k^T A d_k} r_k^T d_k,$$

zatem, skoro $d_k^T A d_k \leq \lambda_{max}(A)\|d_k\|^2$ [(9)]

$$\varphi(x_{k+1}) - \varphi(x_k) = \beta_k(\beta_k - 2)\frac{(d_k^T r_k)^2}{d_k^T A d_k},$$

$$\varphi(x_k) - \varphi(x_{k+1}) = \beta_k(2 - \beta_k)\frac{(d_k^T r_k)^2}{d_k^T A d_k} \overset{(2.1)}{\geq} \delta_1^2 \frac{(d_k^T r_k)^2}{d_k^T A d_k} \overset{(2.1)}{\geq} \delta_1^2\delta_2^2 \frac{\|r_k\|^2\|d_k\|^2}{d_k^T A d_k}$$

$$\overset{(2.2)}{\geq} \delta_1^2\delta_2^2 \frac{\|r_k\|^2\|d_k\|^2}{\lambda_{max}(A)\|d_k\|^2} > 0,\ k = 0, 1, \ldots$$

Tak więc ciąg $\{\varphi(x_k)\}_k$ jest silnie malejący oraz ograniczony od dołu, przez $-b^T A^{-1} b$ (co wynika z (29)), jest więc zbieżny. Zatem

$$\lim_{k\to\infty} (\varphi(x_k) - \varphi(x_{k+1})) = 0.$$

Ale

$$\varphi(x_k) - \varphi(x_{k+1}) \geq \frac{\delta_1^2\delta_2^2}{\lambda_{max}(A)}\|r_k\|^2 > 0,$$

więc z twierdzenia o trzech ciągach

$$\lim_{k\to\infty} \|r_k\| = 0,$$

co jest równoważne temu, że

$$\lim_{k\to\infty} x_k = A^{-1} b = \bar{x}.$$

□

[8] Na mocy nierówności Schwarza i tego, że $r_k{}^T d_k \leq \|r_k\| \cdot \|d_k\|$ mamy $\delta_2 \leq 1$.

[9] z algebry liniowej.

Uogólniona metoda najszybszego spadku może być w zależności od wyboru ciągów $\{\beta_k\}_k$ i $\{d_k\}_k$ metodą liniową jak i metodą nieliniową. W następnych przykładach zajmiemy się przypadkami szczególnymi.

Przykład 2.47. Przyjmijmy

$$d_k = r_k, \ \beta_k = 1.$$

Otrzymujemy zatem, że

$$x_{k+1} = x_k + \lambda_k r_k.$$

Oczywiście założenia twierdzenia 2.46 są spełnione, zatem metoda jest zbieżna. Jest ona nieliniowa, bo

$$\lambda_k = \frac{r_k^T r_k}{r_k^T A r_k}$$

jest nieliniową funkcją zmiennej r_k.

Metodę tą nazywamy **metodą najszybszego spadku.**

Przykład 2.48. Niech

$$\beta_k = 1,$$
$$d_k = \begin{cases} e_{k \mod n}, & k \neq ln \\ e_n, & k = ln \end{cases}$$

Wówczas

$$\lambda_k = \frac{d_k^T r_k}{d_k^T A d_k}$$

zależy w sposób liniowy od r_k. Ciąg przybliżeń przyjmuje postać

$$x_{k+1} = x_k + \lambda_k e_j. \tag{36}$$

Policzmy jeszcze

$$\begin{aligned} r_{k+1}^T d_k &= (b - Ax_{k+1})^T d_k = (b - A(x_k + \lambda_k d_k))^T d_k = (b - Ax_k - \lambda_k A d_k)^T d_k \\ &= r_k^T d_k - \lambda_k d_k^T A d_k = 0, \end{aligned}$$

czyli

$$r_{k+1}^T d_k = 0. \tag{37}$$

Ale skoro $d_k = e_j$ to $r_{k+1}^T = 0$. Równość ta zachodzi wtedy i tylko wtedy, gdy j-ta współrzędna $r_{k+1} = b - Ax_{k+1}$ zeruje się. Inaczej mówiąc, j-te równanie układu (1) jest spełnione dokładnie.

Równania (36) i (37) charakteryzują metodę Gaussa-Seidla. Dostajemy zatem, na mocy twierdzenia 2.46, że metoda Gaussa-Seidla jest dla macierzy $A = A^T > 0$ zbieżna.

Przykład 2.49. Jeszcze jeden przykład metody liniowej – **metoda Richardsona.** Niech

$$d_k = r_k,\ \beta_k\lambda_k = \alpha.$$

Wówczas

$$x_{k+1} = (I - \alpha A)x_k + \alpha b.$$

Twierdzenie 2.50. *(oszacowanie zbieżności metody najszybszych spadków)*
Jeśli $\bar{x}$ *jest rozwiązaniem układu (1) dla macierzy* $A = A^T > 0$, $x_0 \in \mathbb{R}^n$ *przybliżenie początkowe,*

$$x_{k+1} = x_k + \lambda_k r_k,$$
$$r_k = b - Ax_k,$$
$$\lambda_k = \frac{r_k^T r_k}{r_k^T A r_k},$$

to istnieje stała K *taka, że*

$$\|x_k - \bar{x}\|_2^2 \le K\Big(\frac{M-m}{M+m}\Big)^{2k},$$

gdzie M, m *takie, że*

$$0 < m \le \lambda_j(A) \le M,\ j = 1, \dots, n.$$

Dowód. Oszacujmy $\varphi(x_k) - \varphi(\bar{x})$.

$$\begin{aligned}\varphi(x_{k+1}) - \varphi(\bar{x}) &= \varphi(x_{k+1}) - \varphi(x_k) + \varphi(x_k) - \varphi(\bar{x}) \\ &= (\varphi(x_k) - \varphi(\bar{x}))(1 - \frac{\varphi(x_k) - \varphi(x_{k+1})}{\varphi(x_k) - \varphi(\bar{x})}).\end{aligned}$$

Na podstawie dowodu poprzedniego twierdzenia, dla $d_k = r_k,\ \beta_k = 1$, mamy

$$\varphi(x_k) - \varphi(x_{k+1}) = \tfrac{(r_k^T r_k)^2}{r_k^T A r_k}.$$

Z (29) otrzymujemy

$$\varphi(\bar{x}) = -b^T A^{-1} b,$$
$$\tilde{\varphi}(x) = \varphi(x) + b^T A^{-1} b = \varphi(x) - \varphi(\bar{x}).$$

Z defnicji $\tilde{\varphi}$ mamy

$$\tilde{\varphi}(x_k) = (Ax_k - b)^T A^{-1}(Ax_k - b) = r_k^T A^{-1} r_k.$$

Zatem

$$\varphi(x_k) - \varphi(\bar{x}) = r_k^T A^{-1} r_k,$$
$$\varphi(x_{k+1}) - \varphi(\bar{x}) = (\varphi(x_k) - \varphi(\bar{x}))\big(1 - \tfrac{(r_k^T r_k)^2}{r_k^T A r_k} \cdot \tfrac{1}{r_k^T A^{-1} r_k}\big).$$

Więc

$$\varphi(x_{k+1}) - \varphi(\bar{x}) \leq (\varphi(x_k) - \varphi(\bar{x}))(1-q), \tag{38}$$

gdzie

$$\begin{aligned} q &= \min\left\{\tfrac{(r_k^T r_k)^2}{(r_k^T A r_k)(r_k^T A^{-1} r_k)} : r_k \neq 0\right\} = \min\left\{\tfrac{(u^T u)^2}{(u^T A u)(u^T A^{-1} u)} : u \neq 0\right\} \\ &= \min\left\{\tfrac{1}{(v^T A v)(v^T A^{-1} v)} : \ \|v\| = 1\right\}. \end{aligned}$$

Niech $\alpha > 0$. Skoro dla dowolnych a, b mamy $ab \leq \frac{1}{4}(a+b)^2$, to

$$(v^T A v)(v^T A^{-1} v) = (v^T(\alpha A)v)(v^T(\alpha A)^{-1} v) \leq \tfrac{1}{4}(v^T((\alpha A) + (\alpha A)^{-1})v)^2. \tag{39}$$

Macierz $\alpha A + (\alpha A)^{-1}$ jest dodatnio określona i ma wartości własne

$$\lambda_i(\alpha A + (\alpha A)^{-1}) = \alpha\lambda_i(A) + (\alpha\lambda_i(A))^{-1}.$$

Niech $\|v\| = 1$, więc

$$\begin{aligned} v^T(\alpha A + (\alpha A)^{-1})v &\leq \max\{\alpha\lambda_i(A) + (\alpha\lambda_i(A))^{-1} |\ i = 1, \ldots, n\} \\ &\leq \max\{\alpha\lambda + (\alpha\lambda)^{-1} |\ \lambda \in [m, M]\}, \end{aligned}$$

czyli

$$v^T(\alpha A + (\alpha A)^{-1})v \leq \max\{f(\lambda) |\ \lambda \in [m, M]\}, \tag{40}$$

gdzie

$$f(\lambda) = \alpha\lambda + \tfrac{1}{\alpha\lambda}.$$

Bo $\alpha > 0$ Oczywiście f jest funkcją wypukłą osiągającą swoje maksimum na brzegu przedziału $[m, M]$. Wybierzmy α tak, aby $f(m) = f(M)$, czyli niech

$$\alpha = \frac{1}{\sqrt{mM}}.$$

Wówczas

$$f(m) = \frac{1}{\sqrt{mM}} m + \sqrt{mM}\frac{1}{m} = \sqrt{\frac{m}{M}} + \sqrt{\frac{M}{m}} = \frac{1}{\sqrt{Mm}}(m+M).$$

Na mocy (39) i (40) otrzymujemy

$$\begin{aligned} 1 - q &= 1 - \min\left\{\tfrac{1}{(v^T A v)(v^T A^{-1} v)} : \ \|v\| = 1\right\} \\ &\leq 1 - \min\left\{\tfrac{1}{\frac{1}{4}(v^T((\alpha A)+(\alpha A)^{-1})v)^2} : \ \|v\| = 1\right\} \\ &\leq 1 - \tfrac{1}{\frac{1}{4}\max^2\{f(\lambda) |\ \lambda\in[m,M]\}} = 1 - \tfrac{4}{f^2(m)} = 1 - \tfrac{4Mm}{(m+M)^2} \\ &= \tfrac{(m-M)^2}{(m+M)^2}. \end{aligned}$$

W (38) dostajemy więc

$$\varphi(x_{k+1}) - \varphi(\bar{x}) \leq (\varphi(x_k) - \varphi(\bar{x}))\Big(\frac{m-M}{m+M}\Big)^2 \leq \dots$$
$$\leq (\varphi(x_0) - \varphi(\bar{x}))\Big(\frac{m-M}{m+M}\Big)^{2(k+1)}$$

Jako szukaną stałą K wystarczy przyjąć $\frac{\varphi(x_0)-\varphi(\bar{x})}{\lambda_{min}(A)}$, bo

$$\varphi(x_k) - \varphi(\bar{x}) = (Ax_k - b)^T A^{-1}(Ax_k - b) = (Ax_k - b)^T(x_k - A^{-1}b)$$
$$= (x_k - \bar{x})^T(Ax_k - b) = (x_k - \bar{x})^T A(x_k - A^{-1}b)$$
$$= (x_k - \bar{x})^T A(x_k - \bar{x}) \geq \lambda_{min}(A)\|x_k - \bar{x}\|_2^2.$$

Wówczas

$$\lambda_{min}(A)\|x_k - \bar{x}\|_2^2 \leq \varphi(x_k) - \varphi(\bar{x}) \leq (\varphi(x_0) - \varphi(\bar{x}))\Big(\frac{m-M}{m+M}\Big)^{2k}.$$

□

Uwaga 2.51. *Skoro* $A = A^T > 0$, *cond*$(A) = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$, *to*

$$\frac{M-m}{M+m} = \frac{\frac{M}{m}-1}{\frac{M}{m}+1} = \frac{cond(A)-1}{cond(A)+1}.$$

Zatem dla metody Richardsona otrzymujemy, że

$$\rho(B(\alpha_0)) = \frac{cond(A)-1}{cond(A)+1}.$$

2.4.3 Metoda gradientów sprzężonych.

Załóżmy, że $A = A^T > 0$. Wybieramy dowolnie $x_0 \in \mathbb{R}^n$ – przybliżenie początkowe i określamy pięć ciągów $\{\alpha_n\}_n, \{\beta_n\}_n \subset \mathbb{R}$, $\{x_n\}_n, \{r_n\}_n, \{p_n\}_n \subset \mathbb{R}^n$ następująco:

$$r_0 = b - Ax_0,\ p_0 = r_0,$$

(41) $\alpha_k = \dfrac{r_k^T p_k}{p_k^T A p_k}$,

(42) $x_{k+1} = x_k + \alpha_k p_k$,

(43) $r_{k+1} = r_k - \alpha_k A p_k$,

(44) $\beta_k = -\dfrac{r_{k+1}^T A p_k}{p_k^T A p_k}$,

(45) $p_{k+1} = r_{k+1} + \beta_k p_k$.

Twierdzenie 2.52. *Jeżeli* A *jest dodatnio określoną macierzą symetryczną, to po skończonej liczbie kroków algorytm metody gradientów sprzężonych daje rozwiązanie dokładne układu (1).*

Dowód. Pokażemy najpierw, że zachodzą następujące równości:

(i) $r_i^T p_j = 0,\ i > j$,

(ii) $r_i^T r_j = 0,\ i \neq j$,

(iii) $p_i^T A p_j = 0,\ i \neq j$, to znaczy kierunki p_i są A-ortogonalne[10].

Dowód tych trzech równości przeprowadzimy indukcyjnie. Pierwszy krok indukcyjny:

$$r_1^T p_0 \overset{(43)}{=} (r_0 - \alpha_0 A p_0)^T p_0 = r_0^T p_0 - \alpha_0 p_0^T A^T p_0 \overset{(41)}{=} r_0^T p_0 - r_0^T p_0 = 0,$$
$$r_1^T r_0 = r_1^T p_0 \overset{(i)}{=} 0,$$
$$p_1^T A p_0 = (r_1 + \beta_0 p_0)^T A p_0 = r_1^T A p_0 + \beta_0 p_0^T A p_0 \overset{(44)}{=} r_1^T A p_0 - r_1^T A p_0 = 0.$$

Załóżmy teraz, że wzory *(i)* – *(iii)* zachodzą dla $0 \leq i, j \leq k$.

Ad*(i)* Z definicji ciągu $\{r_k\}_k$, dla $j \leq k$, mamy

$$r_{k+1}^T p_j \overset{(43)}{=} (r_k - \alpha_k A p_k)^T p_j = r_k^T p_j - \alpha_k p_k^T A p_j.$$

Jeśli $j < k$ to na mocy założenia indukcyjnego $p_k^T A p_j \overset{(iii)}{=} 0$ i $r_k^T p_j \overset{(i)}{=} 0$. Jeśli $j = k$ to na podstawie definicji ciągu $\{\alpha_k\}_k$

$$r_{k+1}^T p_k = r_k^T p_k - \alpha_k p_k^T A p_k \overset{(41)}{=} r_k^T p_k - r_k^T p_k = 0.$$

Ad*(ii)*

$$r_{k+1}^T r_j \overset{(45)}{=} r_{k+1}^T (p_j - \beta_{j-1} p_{j-1}) = r_{k+1}^T p_j - \beta_{j-1} r_{k+1}^T p_{j-1} \overset{(i)}{=} 0.$$

Ad*(iii)* Oczywiście

$$p_{k+1}^T A p_j \overset{(45)}{=} (r_{k+1} + \beta_k p_k)^T A p_j = r_{k+1}^T A p_j + \beta_k p_k^T A p_j.$$

Jeśli $j < k$ oraz $\alpha_j \neq 0$, to

$$p_{k+1}^T A p_j \overset{(iii)}{=} r_{k+1}^T A p_j \overset{(43)}{=} r_{k+1}^T \frac{1}{\alpha_j}(r_j - r_{j+1}) \overset{(ii)}{=} 0.$$

Jeśli $j = k$, to

$$p_{k+1}^T A p_j = r_{k+1}^T A p_k + \beta_k p_k^T A p_k \overset{(44)}{=} r_{k+1}^T A p_k - r_{k+1}^T A p_k = 0.$$

Ponieważ układ $\{r_k\}_k$ jest, na mocy *(ii)*, ortogonalny w $\mathbb{R}^n$, zatem

$\exists k < n : \{r_0, \ldots, r_k\}$ jest liniowo niezależny,

[10] ortogonalne w sensie iloczynu skalarnego $(u|v) = u^T A v$

więc

$$\forall\ k \geq n : \{r_0, \ldots, r_k\} \text{ jest liniowo zależny,}$$

czyli

(46) $\forall\ k \geq n : r_k = 0.$

Wykażemy, że

(47) $r_{k+1} = b - Ax_{k+1}.$

Na mocy (43)

$$r_{k+1} = r_k - \alpha_k A p_k = r_{k-1} - \alpha_k A p_k - \alpha_{k-1} A p_{k-1} = \ldots = r_0 - \sum_{j=0}^{k} \alpha_j A p_j$$

$$= r_0 - \sum_{j=0}^{k} A\alpha_j p_j \overset{(42)}{=} r_0 - \Big(\sum_{j=0}^{k} A x_{j+1} - \sum_{j=0}^{k} A x_j\Big) = b - A x_{k+1}.$$

Zatem, na mocy (46) i (47), dla $k \geq n-1$, x_{k+1} jest rozwiązaniem układu (1). □

3 Wyznaczanie wartości własnych i wektorów własnych macierzy.

Niech $A \in \mathbb{C}^{n\times n}$, $A = (a_{ij})$.

Twierdzenie 3.1. *(Gerszgorin, Schur; lokalizacja wartości własnych)*
Jeśli λ jest wartością własną macierzy A, to

$$\exists i \in \{1, \ldots, n\} :\ |\lambda - a_{ii}| \leq \sum_{j\neq i} |a_{ij}|.$$

Dowód. Niech x będzie wektorem własnym odpowiadającym wartości własnej λ, zatem

$$\lambda x = Ax,$$

czyli

$$\lambda x_i = \sum_{j=1}^{n} a_{ij} x_j,\ i = 1, \ldots, n,$$

$$(\lambda - a_{ii}) x_i = \sum_{j\neq i} a_{ij} x_j,\ i = 1, \ldots, n.$$

Skoro $x \neq 0$, więc

$$\exists i \in \{1, \ldots, n\} :\ x_i \neq 0,$$

zatem

$$\exists i \in \{1, \ldots, n\}: \ |x_i| = \|x\|_\infty \neq 0.$$

Dla takiego i mamy

$$\|x\|_\infty |\lambda - a_{ii}| = |(\lambda - a_{ii})x_i| = \Big| \sum_{j \neq i} a_{ij} x_j \Big| \leq \sum_{j \neq i} |a_{ij}| \cdot |x_j| \leq \|x\|_\infty \sum_{j \neq i} |a_{ij}|,$$

więc

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|.$$

□

Dowód alternatywny. Przypuśćmy dla dowodu nie wprost, że

$$\forall \ i \in \{1, \ldots, n\}: \ |\lambda - a_{ii}| > \sum_{j \neq i}^{n} |a_{ij}|.$$

Wówczas macierz $B = A - \lambda I$ spełnia mocne kryterium sumy wierszy, zatem na mocy twierdzenia 2.35, B jest macierzą nieosobliwą, co prowadzi do sprzeczności, bo λ jest wartością własną macierzy A. □

Wniosek 3.2. *Jeśli λ jest wartością własną macierzy A, to*

$$\exists j \in \{1, \ldots, n\}: \ |\lambda - a_{jj}| \leq \sum_{i \neq j}^{n} |a_{ij}|.$$

Dowód. Wystarczy zastosować poprzednie twierdzenie dla macierzy A^T, bo A i A^T mają te same wartości własne. □

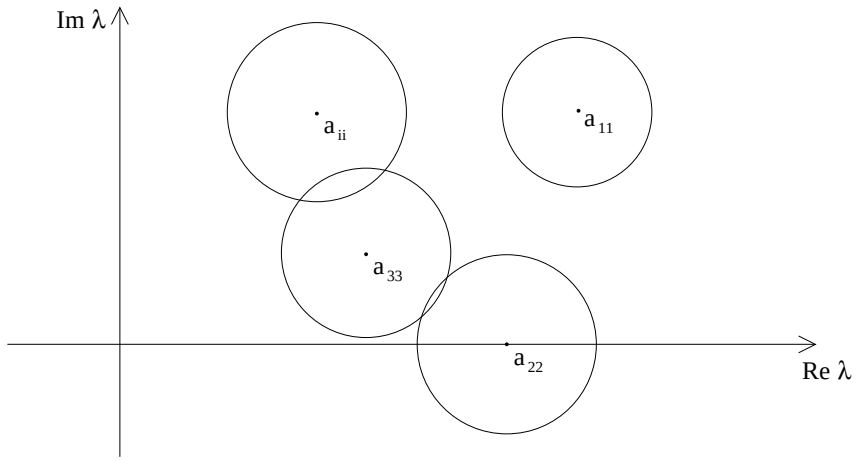
Ustalmy $k \in \{1, \ldots, n\}$. Niech

$$K_1 = \bigcup_{i=1}^{k} \big\{\lambda : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\big\},$$
$$K_2 = \bigcup_{i=k+1}^{n} \big\{\lambda : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\big\},$$

Na mocy twierdzenia 3.1

$$\{\lambda_i(A)\} \subset \bigcup_{i=1}^{n} \big\{\lambda : |\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}|\big\}.$$

Uwaga 3.3. *Jeśli $K_1 \cap K_2 = \emptyset$ oraz K_1 zawiera k wartości własnych macierzy A, to K_2 zawiera $n - k$ wartości własnych macierzy A.*



Rysunek 7: Zbiory K_1 i K_2

Dowód. Niech D będzie macierzą diagonalną, L macierzą trójkątną dolną i U macierzą trójkątną górną takimi, że

$$A = D + L + U.$$

Zdefiniujmy macierz $A_t = (a_{ij}(t))$ dla $t \in [0, 1]$ następująco

$$A_t = D + t(L + U).$$

Zatem

$$a_{ii}(t) = a_{ii},\ a_{ij}(t) = ta_{ij}.$$

Niech $K_1(t), K_2(t)$ będą określone jak wyżej dla macierzy A_t. Zatem, skoro $t \in [0, 1]$, mamy

$$K_i(t) \subseteq K_i,\ i = 1, 2.$$

Skoro $A_0 = D$, więc $\lambda_i(A_0) = a_{ii},\ i = 1, \ldots, n$, oraz

$$K_1(0) = \{a_{11}, \ldots, a_{kk}\},\ K_2(0) = \{a_{k+1,k+1}, \ldots, a_{nn}\}.$$

Mamy też

$$\lambda_i(A_t) \in K_1(t) \subset K_1, i = 1, \ldots, k.$$

Teza twierdzenia wynika z ciągłości względem t funkcji $\lambda_i(A_t)$. □

W rozdziale 2.3.1 spotkaliśmy się z wartościami własnymi i wektorami własnymi macierzy. Teraz zajmiemy się ich wyznaczaniem. Pokazaliśmy już, że λ jest wartością własną macierzy A wtedy i tylko wtedy, gdy jest miejscem zerowym wielomianu charakterystycznego macierzy A, oraz że macierze podobne mają te same wielomiany charakterystyczne. O innej cesze wielomianu charakterystycznego mówi twierdzenie

Twierdzenie 3.4. *(Cayley-Hamiltona)*
Niech

$$\varphi_A(\lambda) = \det(A - \lambda I),$$

będzie wielomianem wielomianem charakterystycznym macierzy A*. Wówczas*

$$\varphi_A(A) = 0.$$

Dowód. Definicja (2.18) wielomianu charakterystycznego. □

Definicja 3.5. *Niech* $x_0 \in \mathbb{R}^n \setminus \{0\}$ *będzie dowolnie ustalony. Ciągiem Kryłowa dla macierzy* A *nazywamy ciąg* $\{x_k\}_k$ *taki, że*

$$x_{k+1} = Ax_k, \ k = 0, 1, \ldots \tag{48}$$

Lemat 3.6. *W ciągu Kryłowa co najwyżej* n *elementów jest liniowo niezależnych.*

Uwaga 3.7. *Jeśli elementy* $x_0, x_1, \ldots, x_{p-1}$ *ciągu Kryłowa są liniowo niezależne, a* $x_0, x_1, \ldots, x_p$ *są liniowo zależne, to*

$\forall k \geq p$ *wektory* $x_0, \ldots, x_{p-1}, x_k$ *są liniowo zależne.*

Dowód. Przeprowadzimy indukcję matematyczną. Krok pierwszy ($k = p$) wynika z założenia. Niech $k > p$. Na mocy założenia indukcyjnego, ciąg $x_0, \ldots, x_{p-1}, x_{k-1}$ jest liniowo zależny, zatem istnieją stałe $\alpha_0, \ldots, \alpha_{p-1}$ takie, że

$$x_{k-1} = \sum_{j=0}^{p-1} \alpha_j x_j.$$

Wówczas

$$\begin{aligned} x_k = Ax_{k-1} = A\sum_{j=0}^{p-1} \alpha_j x_j = A\alpha_{p-1}x_{p-1} + \sum_{j=0}^{p-2} \alpha_j Ax_j = \alpha_{p-1}x_p + \sum_{j=0}^{p-2} \alpha_j x_{j+1} \\ = \alpha_{p-1}\sum_{j=0}^{p-1} \tilde{\alpha}_j x_j + \sum_{j=1}^{p-1} \alpha_{j-1}x_j = \alpha_{p-1}\tilde{\alpha}_0 x_0 + \sum_{j=1}^{p-1} (\alpha_{p-1}\tilde{\alpha}_j + \alpha_{j-1})x_j, \end{aligned}$$

a więc $x_k \in \operatorname{span}\{x_0, \ldots, x_{p-1}\}$. □

3.1 Metody dokładne.

Niech A będzie kwadratową macierzą wymiaru n. Zajmiemy się wyznaczaniem wielomianu charakterystycznego macierzy A lub jego dzielnika. Zacznijmy od zdefiniowania **wielomianu minimalnego** macierzy A. Ustalmy $x_0 \in \mathbb{R}^n$ i weźmy ciąg Kryłowa $\{x_n\}_{n=0}^{\infty}$. Na mocy lematu 3.6 istnieje $p \leq n$ takie, że

$x_0, \ldots, x_{p-1}$ są liniowo niezależne,

$x_0, \ldots, x_p$ są liniowo zależne.

Zatem istnieją $\alpha_0, \ldots, \alpha_{p-1}$ takie, że

$$x_p + \alpha_{p-1}x_{p-1} + \ldots + \alpha_0 x_0 = 0.$$

Z konstrukcji ciągu Kryłowa mamy

$$\begin{aligned}
&A^p x_0 + \alpha_{p-1}A^{p-1}x_0 + \ldots + \alpha_0 I x_0 = 0,\\
&(A^p + \alpha_{p-1}A^{p-1} + \ldots + \alpha_0 I)x_0 = 0,\\
&\psi(A)x_0 = 0,\ \psi(\lambda) = \lambda^p + \alpha_{p-1}\lambda^{p-1} + \ldots + \alpha_0.
\end{aligned}$$

Definicja 3.8. *Niech $A \in \mathbb{R}^{n\times n}$ będzie macierzą i niech $x_0 \in \mathbb{R}^n,\ x_0 \neq 0$. Wielomianem minimalnym dla x_0 nazywamy wielomian $\psi(\lambda) = \sum_{k=0}^{p} \alpha_k \lambda^k$ taki, że*

(i) $\psi(A)x_0 = 0$;

(ii) jeśli $\tilde{\psi}(\lambda)$ spełnia (i) oraz $\deg\tilde{\psi} < \deg\psi$, to $\tilde{\psi} = 0$.

Lemat 3.9. *Jeśli wielomian $\tilde{\psi}(\lambda)$ spełnia warunek (i) definicji 3.8, to ψ dzieli $\tilde{\psi}$.*

Dowód. Załóżmy, że stopień wielomianu $\tilde{\psi}$ jest nie mniejszy niż stopień wielomianu ψ, wówczas

$$\tilde{\psi}(\lambda) = a(\lambda)\psi(\lambda) + r(\lambda),\ \deg r < \deg\psi.$$

Zatem

$$0 = \tilde{\psi}(A)x_0 = a(A)\psi(A)x_0 + r(A)x_0 = r(A)x_0,$$

a więc r spełnia warunek *(ii)* definicji 3.8, zatem $r = 0$. □

Odpowiedzmy teraz na pytanie, jaki jest związek między wielomianami minimalnymi dla pewnego $x_0 \in \mathbb{R}^n$ i wielomianami charakterystycznymi. Ustalmy $x_0 \in \mathbb{R}^n$. Niech φ_A będzie wielomianem charakterystycznym macierzy A, ψ wielomianem minimalnym dla x_0. Na mocy twierdzenia 3.4

$$\varphi_A(A)x_0 = 0.$$

Zatem $\psi(\lambda)$ dzieli $\varphi_A(\lambda)$, więc miejsca zerowe wielomianu ψ są wartościami własnymi macierzy A. Jeśli więc $p = n$, to

$$\psi(\lambda) = \lambda^p + \alpha_{p-1}\lambda^{p-1} + \ldots + \alpha_1\lambda + \alpha_0,$$

gdzie $\alpha_0, \ldots, \alpha_p$ są współczynnikami wielomianu φ_A. Jeśli $p < n$ to $\alpha_0, \ldots, \alpha_p$ są współczynnikami pewnego dzielnika wielomianu φ_A.

Oznaczmy przez $\psi_j(\lambda)$ wielomian minimalny dla e_j.

Definicja 3.10. *Wielomianem minimalnym macierzy A nazywamy najmniejszą wspólną wielokrotność wielomianów $\psi_1, \ldots, \psi_n$*[11].

[11] Jest to wielomian najmniejszego stopnia podzielny przez każdy z wielomianów $\psi_1, \ldots, \psi_n$.

Twierdzenie 3.11. *Jeśli* ψ *jest wielomianem minimalnym macierzy* A*, to* $\deg\psi \leq \deg\varphi_A$.

Dowód. Zauważmy, że jeśli ψ jest wielomianem minimalnym macierzy A, to $\psi(A)e_j = 0$ dla $j = 1, 2, \ldots, n$. Niech $x_0 \in \mathbb{R}^n$. Wówczas

$$x_0 = \sum_{j=1}^{n} \xi_j e_j,$$

$$\psi(A)x_0 = \psi(A)\sum_{j=1}^{n} \xi_j e_j = \sum_{j=1}^{n} \xi_j \psi(A)e_j = 0.$$

Zatem ψ spełnia warunek (i) definicji 3.8. □

Załóżmy że $x_0, \ldots, x_{n-1}$ są liniowo niezależne. Niech $X = [x_0, \ldots, x_{n-1}]$, wówczas

$$AX = A[x_0, \ldots, x_{n-1}] = [Ax_0, \ldots, Ax_{n-1}] = [x_1, \ldots, x_n] =$$

$$[x_0, \ldots, x_{n-1}] \begin{pmatrix} 0 & 0 & \ldots & 0 & -\alpha_0 \\ 1 & 0 & \ldots & 0 & -\alpha_1 \\ 0 & 1 & \ldots & 0 & -\alpha_2 \\ \ldots & & & & \\ 0 & 0 & \ldots & 1 & -\alpha_{n-1} \end{pmatrix} = XF.$$

Macierzy F nazywamy **macierzą Frobeniusa.** Skoro $x_0, \ldots, x_{n-1}$ są liniowo niezależne, to $\det X \neq 0$, więc X jest odwracalna. Zatem macierze A i F są podobne (bo $A = XFX^{-1}$). Macierz F ma dużo zer, więc łatwo jest wyznaczyć jej wielomian charakterystyczny

$$\varphi_F(\lambda) = (-1)^n(\lambda^n + \alpha_{n-1}\lambda^{n-1} + \ldots + \alpha_0).$$

Ze względu na podobieństwo macierzy A i F jest to też wielomian charakterystyczny macierzy A.

3.2 Metody iteracyjne.

3.2.1 Metoda potęgowa.

Załóżmy, że wartości własne $\lambda_1, \ldots, \lambda_n$ macierzy $A \in \mathbb{R}^{n\times n}$ są takie, że

(49) $|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|$.

Ustalmy dowolny $x_0 \in \mathbb{R}^n \setminus \{0\}$ i weźmy ciąg Kryłowa $\{x_k\}_k$. Oczywiście

$$x_k = \begin{pmatrix} x_{1k} \\ \vdots \\ x_{nk} \end{pmatrix}, \; x_{ik} = e_i{}^T \cdot x_k.$$

Twierdzenie 3.12. *Jeśli wartości własne macierzy* A *spełniają (49), to prawie zawsze*[12]

$$\lim_{k\to\infty} \frac{x_{i,k+1}}{x_{ik}} = \lambda_1$$

[12]Określenie „prawie zawsze" oznacza, że można źle wybrać x_0, ale prawdopodobieństwo takiego wyboru jest równe zero.

Dowód. Niech $v_1, \dots, v_n$ będą wektorami własnymi (liniowo niezależnymi) macierzy A odpowiadającymi wartością $\lambda_1, \dots, \lambda_n$ ($Av_j = \lambda_j v_j$). Niech też $x_0 \in \mathbb{R}^n$. Wówczas istnieją $\xi_i, \dots, \xi_n$ takie, że

$$x_0 = \sum_{j=1}^{n} \xi_j v_j.$$

Zatem

$$\begin{aligned} x_k &= A^k x_0 = \sum_{j=1}^{n} \xi_j A^k v_j = \sum_{j=1}^{n} \xi_j \lambda_j^k v_j, \ k = 1, 2, \dots \\ x_{ik} &= e_i{}^T x_k = \sum_{j=1}^{n} \xi_j \lambda_j^k e_i{}^T v_j = \lambda_1^k \big(\xi_1 e_i{}^T v_1 + \sum_{j=2}^{n} \xi_j \big(\tfrac{\lambda_j}{\lambda_1}\big)^k e_i{}^T v_j\big) \\ &= \lambda_1^k (\xi_1 e_i{}^T v_1 + \mathcal{O}(|\lambda_2/\lambda_1|^k)), \end{aligned}$$

zatem

$$\frac{x_{i,k+1}}{x_{ik}} = \frac{\lambda_1^{k+1}(\xi_1 e_i{}^T v_1 + \mathcal{O}(|\lambda_2/\lambda_1|^{k+1}))}{\lambda_1^k(\xi_1 e_i{}^T v_1 + \mathcal{O}(|\lambda_2/\lambda_1|^k))} \underset{k\to\infty}{\to} \lambda_1 \frac{\xi_1 e_i{}^T v_1}{\xi_1 e_i{}^T v_1} = \lambda_1,$$

o ile $\xi_1 e_i{}^T v_1 \neq 0$.

Wystarczy wykazać, że

$$m\{x_0 \in \mathbb{R}^n : \xi_1 e_i{}^T v_1 = 0\} = 0.$$

Skoro $v_1, \dots, v_n \in \mathbb{R}^n$ są liniowo niezależne, to $\mathbb{R}^n = V_1 \oplus V_2$, gdzie

$$\begin{aligned} V_1 &= \operatorname{span}\{v_1\}, \\ V_2 &= \operatorname{span}\{v_2, \dots, v_n\}. \end{aligned}$$

Skoro $\dim V_1 = 1$, to $\dim V_2 = n - 1$, więc $m_n(V_2) = 0$.

Oczywiście

$$\xi_1 e_i^T v_1 = 0 \Leftrightarrow \xi_1 = 0 \text{ lub } e_i^T v_1 = 0.$$

Jeżeli $\xi_1 = 0$, to

$$x_0 = \sum_{j=2}^{n} \xi_j v_j.$$

Podobnie, jeśli $e_i{}^T v_1 = 0$, to

$$\begin{aligned} x_{i0} &= e_i^T x_0 = \sum_{j=1}^{n} \xi_j e_i^T v_j = \xi_1 e_i^T v_1 + \sum_{j=2}^{n} \xi_j e_i^T v_j = \sum_{j=2}^{n} \xi_j e_i^T v_j = e_i^T \sum_{j=2}^{n} \xi_j v_j, \\ x_0 &= \sum_{j=2}^{n} \xi_j v_j. \end{aligned}$$

Zatem, jeżeli $\xi_1 e_i^T v_1 = 0$, to $x_0 \in V_2$, co kończy dowód. □

Widać więc, że można łatwo wyznaczyć wartość własną macierzy A, bo

$$\frac{x_{j,k+1}}{x_{jk}} \approx \lambda_1, \ \forall k \geq k_0, \ j = 1, \ldots, n,$$
$$x_{k+1} \approx \lambda_1 x_k,$$
$$Ax_k \approx \lambda_1 x_k \Rightarrow \lambda_1 \ - \text{ wartość własna.}$$

Uwaga 3.13. *Powyższe twierdzenie zachodzi przy założeniu, że*

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Dowód. Niech v_1 będzie wektorem własnym odpowiadającym λ_1 i niech

$$V_1 = \{\lambda v_1 : \lambda \in \mathbb{R}\}.$$

Wykażemy, że $AV_1 = V_1$. Niech $x \in AV_1$, zatem

$$\exists v \in V_1 : x = Av.$$

Skoro $v \in V_1$, to istnieje $\lambda \in \mathbb{R}$ taka, że $v = \lambda v_1$, a więc

$$\exists \lambda \in \mathbb{R} : x = \lambda A v_1 = \lambda \lambda_1 v_1 \in V_1.$$

Niech teraz $v \in V_1$. Oczywiście

$$\exists \tilde{\lambda} \in \mathbb{R} : v = \tilde{\lambda} v_1.$$

Niech $\lambda = \frac{\tilde{\lambda}}{\lambda_1}$, zatem

$$\exists \lambda \in \mathbb{R} : v = \lambda \lambda_1 v_1 = \lambda A v_1 = A(\lambda v_1) \in AV_1.$$

Niech V_2 będzie taka, że $V_1 \oplus V_2 = \mathbb{R}^n$. Skoro macierz A jest podobna (na mocy twierdzenia o postaci kanonicznej Jordana) do macierzy

$$\begin{pmatrix} \lambda_1 & 0 \\ 0 & B \end{pmatrix},$$

gdzie macierz B ma wartości własne $\lambda_2, \ldots, \lambda_n$, to V_2 możemy tak wybrać, aby $AV_2 = V_2$.

Ustalmy $x_0 \in \mathbb{R}^n$. Istnieje dokładnie jedna para $(u, w) \in V_1 \times V_2$ taka, że $x_0 = u + w$. Wówczas

$$\begin{aligned} x_k &= A^k x_0 = A^k u + A^k w = \xi_1 A^k v_1 + A^k w = \xi_1 \lambda_1^k v_1 + A^k w \\ &= \lambda_1^k \big(\xi_1 v_1 + \big(\tfrac{1}{\lambda_1} A\big)^k w\big), \end{aligned}$$

gdzie $\xi_1 \in \mathbb{R}$ jest taka, że $u = \xi_1 v_1$.

Skoro $w \in V_2$ oraz $AV_2 = V_2$, to $Aw = A|_{V_2} w$. Reprezentowana przez B macierz $A|_{V_2}$ ma wartości własne $\lambda_2, \ldots, \lambda_n$, zatem $\big(\frac{1}{\lambda_1} A\big)\Big|_{V_2}$ ma wartości własne $\frac{\lambda_2}{\lambda_1}, \ldots, \frac{\lambda_n}{\lambda_1}$. Zatem

$$x_k = \lambda_1^k \big(\xi_1 v_1 + \mathcal{O}\big(\big|\tfrac{\lambda_2}{\lambda_1}\big|^k\big)\big)$$

i dalej postępujemy jak w twierdzeniu 3.12. □

Uwaga 3.14. *Metoda jest niedogodna, bo zwykle* $\|x_k\| \underset{k \to \infty}{\to} \infty$ *(o ile* $|\lambda_1| > 1$*).*

3.2.2 Wariant metody potęgowej

Zakładamy, że

$$|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|.$$

Niech $x_0 \in \mathbb{R}^n$ będzie punktem startowym. Rozpatrzmy następujące ciągi:

$$\begin{aligned} z_0 &= \tfrac{x_0}{\|x_0\|}, \\ y_k &= Az_{k-1}, \ k = 1, 2, \ldots \\ z_k &= \tfrac{y_k}{\|y_k\|}. \end{aligned}$$

Twierdzenie 3.15. *Przy powyższych założeniach*

$$\lim_{k\to\infty} \frac{y_{ik}}{z_{ik}} = |\lambda_1|.$$

Dowód. Niech $\{x_k\}_k$ będzie ciągiem Kryłowa. Wówczas

$$\begin{aligned} y_k &= \tfrac{x_k}{\|x_{k-1}\|}, \ k = 1, 2, \ldots, \\ z_k &= \tfrac{x_k}{\|x_k\|}, \ k = 0, 1, \ldots \end{aligned}$$

Wykażemy to indukcyjnie. Na mocy założenia

$$z_0 = \tfrac{x_0}{\|x_0\|},$$

oraz

$$y_1 = Az_0 = \tfrac{Ax_0}{\|x_0\|} = \tfrac{x_1}{\|x_0\|}.$$

Drugi krok indukcyjny. Załóżmy, że znamy już $z_0, \ldots, z_{k-1}$ oraz $y_0, \ldots, y_k$, wówczas

$$\begin{aligned} z_k &= \tfrac{y_k}{\|y_k\|} = \tfrac{x_k}{\|x_{k-1}\|} \cdot \tfrac{\|x_{k-1}\|}{\|x_k\|} = \tfrac{x_k}{\|x_k\|}, \\ y_{k+1} &= Az_k = \tfrac{Ax_k}{\|x_k\|} = \tfrac{x_{k+1}}{\|x_k\|}. \end{aligned}$$

Skoro

$$z_{ik} = \tfrac{y_{ik}}{\|y_k\|},$$

to

$$\tfrac{y_{ik}}{z_{ik}} = \|y_k\| = \tfrac{\|x_k\|}{\|x_{k-1}\|} = \frac{|\lambda_1|^k \cdot \|\xi_1 v_1 + \mathcal{O}(|\frac{\lambda_2}{\lambda_1}|^k)\|}{|\lambda_1|^{k-1} \cdot \|\xi_1 v_1 + \mathcal{O}(|\frac{\lambda_2}{\lambda_1}|^{k-1})\|} \underset{k\to\infty}{\to} |\lambda_1|,$$

o ile $\xi_1 v_1 \neq 0$. □

Twierdzenie 3.16. *Jeśli* $A = A^T$, $|\lambda_1| > |\lambda_2| \geq \ldots \geq |\lambda_n|$ *oraz* $\{x_n\}_n$ *jest ciągiem Kryłowa, to iloraz Rayleigha*

$$\sigma_k = \frac{x_k{}^T A x_k}{x_k{}^T x_k}, \tag{50}$$

jest zbieżny do λ_1.

Dowód. Skoro $A = A^T$, to ma n wektorów własnych liniowo niezależnych, wzajemnie ortogonalnych $v_1, \ldots, v_n$. Więc

$$\begin{aligned}
x_0 &= \sum_{j=1}^{n} \xi_j v_j, \\
x_k &= \sum_{j=1}^{n} \xi_j \lambda_j^k v_j, \\
x_k^T A x_k &= \Big(\sum_{j=1}^{n} \xi_j \lambda_j^k v_j\Big)^T \sum_{j=1}^{n} \xi_j \lambda_j^{k+1} v_j = \sum_{j=1}^{n}\sum_{s=1}^{n} \xi_j \xi_s \lambda_j^k \lambda_s^{k+1} v_j^T v_s \\
&= \sum_{j=1}^{n} \xi_j^2 \lambda_j^{2k+1} = \lambda_1^{2k+1}\Big(\xi_1^2 + \sum_{j=2}^{n} \xi_j^2 (\lambda_j/\lambda_1)^{2k+1}\Big) \\
&= \lambda_1^{2k+1}\big(\xi_1^2 + \mathcal{O}(|\lambda_2/\lambda_1|^{2k+1}\big), \\
x_k^T x_k &= \sum_{j=1}^{n}\sum_{s=1}^{n} \xi_j \xi_s \lambda_j^k \lambda_s^k v_j^T v_s = \lambda_1^{2k}\big(\xi_1^2 + \mathcal{O}(|\lambda_2/\lambda_1|^{2k}\big),
\end{aligned}$$

zatem $\sigma_k \underset{k\to\infty}{\to} \lambda_1$, o ile $\xi_1 \neq 0$. □

Zajmiemy się teraz wyznaczaniem kolejnych wartości własnych macierzy symetrycznych (Hermitowskich). Załóżmy, że znamy już λ_1 i v_1, oraz że $|\lambda_1| > |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$. Niech

$$A_1 = A - \lambda_1 v_1 {v_1}^T.$$

Wówczas

$$\begin{aligned}
A_1 v_1 &= A v_1 - \lambda_1 v_1 ({v_1}^T v_1) = \lambda_1 v_1 - \lambda_1 v_1 = 0, \\
A_1 v_j &= A v_j - \lambda_1 v_1 ({v_1}^T v_j) = A v_j = \lambda_j v_j,
\end{aligned}$$

bo ${v_1}^T v_j = \delta_{1j}$ (A jest ortogonalna). Zatem A_1 ma wartości własne $0, \lambda_2, \ldots, \lambda_n$. Dostajemy w ten sposób kolejną wartość własną macierzy A. Sposób ten nie jest najlepszy, może dawać duże błędy jeśli λ_1 i v_1 zostały słabo przybliżone.

3.2.3 Metoda Householdera.

Znajdźmy inną macierz podobną do macierzy $A = A^\star$. Szukamy takiej macierzy U, że

$$U^* U = I,$$

oraz

$$U^* A U = \begin{pmatrix} \lambda_1 & 0 \\ 0 & B \end{pmatrix} = A_1.$$

Zatem, z podobieństwa macierzy A i $\tilde{B}$, macierz B ma wartości własne $\lambda_2, \ldots, \lambda_n$.

Konstrukcja macierzy U przebiega następująco. Niech $v = (\omega, w)^T$ będzie takie, że $Av = \lambda_1 v$ oraz $\|v\|_2 = 1$. Zatem

$$\omega\bar{\omega} + w^*w = 1. \tag{51}$$

Definiujemy

$$U = \begin{pmatrix} \omega & -w^* \\ w & I - \mu ww^* \end{pmatrix}$$

Chcemy tak dobrać μ, aby $U^*U = I$.

$$\begin{aligned} U^*U &= \begin{pmatrix} \bar{\omega} & w^* \\ -w & I - \bar{\mu}ww^* \end{pmatrix} \begin{pmatrix} \omega & -w^* \\ w & I - \mu ww^* \end{pmatrix} \\ &= \begin{pmatrix} \bar{\omega}\omega + w^*w & -\bar{\omega}w^* + w^* - \mu w^*ww^* \\ \hline -\bar{\omega}w^* + w^* - \mu w^*ww^* & ww^* + (I - \bar{\mu}ww^*)(I - \mu ww^*) \end{pmatrix} \\ &\overset{(51)}{=} \left(\begin{array}{c|c} 1 & P \\ \hline P & Q \end{array}\right). \end{aligned}$$

Skoro

$$P = -\bar{\omega}w^* + w^* - \mu w^*ww^* = (1 - \bar{\omega} - \mu\|w\|_2^2)w^*,$$

to przyjmując $\mu = \frac{1-\bar{\omega}}{\|w\|_2^2}$ dostajemy $P = 0$. Mamy też

$$|\mu|^2 = \mu\bar{\mu} = \tfrac{1-\omega-\bar{\omega}+|\omega|^2}{\|w\|_2^4},\ \mu + \bar{\mu} = \tfrac{2-\omega-\bar{\omega}}{\|w\|_2^2}. \tag{52}$$

Policzmy teraz Q:

$$\begin{aligned} Q &= ww^* + (I - \bar{\mu}ww^*)(I - \mu ww^*) \\ &= ww^* + I - \mu ww^* - \bar{\mu}ww^* + \bar{\mu}\mu w\|w\|_2^2w^* \\ &= I + (1 - (\mu + \bar{\mu}) + |\mu|^2\|w\|_2^2)ww^* \overset{(52)}{=} I + \left(1 - \tfrac{-1+|\omega|^2}{\|w\|_2^2}\right) \\ &= I + \tfrac{\|w\|_2^2-1+|\omega|^2}{\|w\|_2^2}ww^* \overset{(51)}{=} I. \end{aligned}$$

Pozostaje zatem sprawdzić, czy zachodzi drugi warunek. Niech $A_1 = U^*AU$. Skoro $A = A^*$, to ${A_1}^* = A_1$.

$$A_1e_1 = U^*AUe_1 \overset{Ue_1=v}{=} U^*Av = \lambda_1U^*v \overset{U^*v=e_1}{=} \lambda_1e_1,$$

zatem

$$A_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & B \end{pmatrix},$$

więc

$$\lambda_j(B) = \lambda_j(A),\ j = 2, \ldots, n.$$

3.3 Wyznaczanie wszystkich wartości własnych macierzy symetrycznych.

3.3.1 Metoda obrotów Jacobiego.

Niech $A = (a_{ij})$ będzie macierzą symetryczną $n \times n$. Rozważmy normę macierzową

$$N^2(A) = \sum_{i,j=1}^{n} |a_{ij}|^2. \tag{53}$$

Lemat 3.17. *Dla normy macierzowej (53) zachodzi następujący wzór*

$$N^2(A) = \mathrm{tr}(A^T A) \tag{54}$$

Dowód. Niech $C = A^T A$. Wówczas, korzystając z tego, że A jest symetryczna

$$c_{ij} = \sum_{k=1}^{n} a_{ik}^T a_{kj} = \sum_{k=1}^{n} a_{ki} a_{kj},$$
$$\mathrm{tr}C = \sum_{s=1}^{n} c_{ss} = \sum_{s=1}^{n} \sum_{k=1}^{n} a_{ks}^2 = \sum_{k,s=1}^{n} a_{ks}^2.$$

□

Uwaga 3.18. *Powyższy lemat jest prawdziwy dla macierzy prostokątnych takich, że* $A^T A$ *jest określone.*

Zanim przejdziemy do samej metody obrotów Jacobiego wykażmy pewne własności normy macierzowej (53).

Twierdzenie 3.19. *Niech* A *będzie kwadratową macierzą symetryczną,* U, V *dowolnymi macierzami (odpowiednich wymiarów). Wówczas norma (53) spełnia następujące własności*

(i) $N^2(A) = N^2(A^T)$,

(ii) $U^T U = I \Rightarrow N^2(UA) = N^2(A)$,

(iii) $V^T V = I \Rightarrow N^2(AV) = N^2(A)$,

(iv) U, V *ortogonalne* $\Rightarrow N^2(UAV) = N^2(A)$.

Dowód.

ad*(i)* Oczywisty.

ad*(ii)* $N^2(UA) = \mathrm{tr}((UA)^T UA) = \mathrm{tr}(A^T U^T UA) = \mathrm{tr}(A^T A) = N^2(A)$.

ad*(iii)* $N^2(AV) \overset{(i)}{=} N^2(V^T A^T) \overset{(ii)}{=} N^2(A^T) \overset{(i)}{=} N^2(A)$.

ad*(iv)* $N^2(UAV) = \mathrm{tr}(V^T A^T U^T UAV) = \mathrm{tr}((AV)^T AV) = N^2(AV) \overset{(iii)}{=} N^2(A)$.

□

Przejdźmy zatem do samej metody obrotów. Dla macierzy $A = A^T$ chcemy znaleźć taki ciąg $\{A^{(k)}\}_k$, że

$$A^{(0)} = A,$$
$$A^{(k+1)} \sim A^{(k)},$$
$$\lim_{k\to\infty} A^{(k)} = \operatorname{diag}(d_1, \ldots, d_n).$$

Jeśli znajdziemy taki ciąg i jego granicę, to automatycznie znajdziemy wszystkie wartości własne macierzy A, czyli $\lambda_i(A) = d_i$. Oznaczmy $A^{(k)} = (a_{ij}^{(k)})$ i niech

$$A^{(k+1)} = T_{p_k q_k}^{-1} A^{(k)} T_{p_k q_k},$$

gdzie p_k, q_k spełniają wzór

$$|a_{p_k q_k}^{(k)}| = \max\{|a_{ij}^{(k)}| : \ i \neq j\},$$

a macierz T_{pq} jest **macierzą obrotu o kąt** θ w płaszczyźnie (pq), czyli

$$t_{ii} = 1, \ i \in \{1, \ldots, n\} \setminus \{p, q\},$$
$$t_{pp} = t_{qq} = c, \text{ gdzie } c = \cos\theta,$$
$$t_{pq} = -t_{qp} = s, \text{ gdzie } s = \sin\theta,$$
$$t_{ij} = 0, \text{ dla pozostałych } i, j.$$

Kąt θ dobieramy tak, aby $a_{p_k q_k}^{(k+1)} = 0$. Macierz T_{pq} jest macierzą ortogonalną.

Rozważmy przypadek $n = 2$. Niech

$$A = \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix}, T = \begin{pmatrix} c & s \\ -s & c \end{pmatrix}.$$

Wówczas

$$T^T = \begin{pmatrix} c & -s \\ s & c \end{pmatrix}, T^T T = \begin{pmatrix} c^2 + s^2 & 0 \\ 0 & c^2 + s^2 \end{pmatrix} = I,$$
$$B = T^{-1} A T = \begin{pmatrix} c & -s \\ s & c \end{pmatrix} \begin{pmatrix} \alpha & \gamma \\ \gamma & \beta \end{pmatrix} \begin{pmatrix} c & s \\ -s & c \end{pmatrix}$$
$$= \begin{pmatrix} c^2\alpha - 2cs\gamma + s^2\beta & cs(\alpha - \beta) + \gamma(c^2 - s^2) \\ cs(\alpha - \beta) + \gamma(c^2 - s^2) & s^2\alpha + 2cs\gamma + c^2\beta \end{pmatrix} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix}.$$

Dobieramy θ tak, aby $cs(\alpha - \beta) + \gamma(c^2 - s^2) = 0$, czyli

$$\gamma(\cos^2\theta - \sin^2\theta) = \cos\theta \sin\theta \cdot (\beta - \alpha),$$
$$\gamma \cos 2\theta = \tfrac{1}{2} \sin 2\theta(\beta - \alpha),$$
$$\operatorname{tg} 2\theta = \frac{2\gamma}{\beta - \alpha},$$
$$\theta = \tfrac{1}{2}\operatorname{arctg}\tfrac{2\gamma}{\beta - \alpha}, \text{ o ile } \alpha \neq \beta.$$

Jeśli $\alpha = \beta$ to bierzemy

$$\theta = \begin{cases} \frac{\pi}{4}, & \gamma > 0 \\ -\frac{\pi}{4}, & \gamma < 0 \end{cases}$$

Przy takim wyborze θ mamy

$$T^{-1}AT = \begin{pmatrix} b_{11} & 0 \\ 0 & b_{22} \end{pmatrix}.$$

Uwaga 3.20. *Suma kwadratów na przekątnej wzrosła o* $2(a^{(k)}_{p_k q_k})^2$.

Dowód.

$$G = T^T A T = \begin{pmatrix} U^T & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} U & 0 \\ 0 & I \end{pmatrix} = \begin{pmatrix} U^T A_{11} U & U^T A_{12} \\ A_{12} U & A_{22} \end{pmatrix},$$

$$\begin{aligned} N^2(G) &= N^2(U^T A_{11} U) + N^2(U^T A_{12}) + N^2(A_{21} U) + N^2(A_{22}) \\ &= N^2(U^T A_{11} U) + N^2(A_{12}) + N^2(A_{21}) + N^2(A_{22}). \end{aligned}$$

□

Twierdzenie 3.21. *Jeśli dla każdego* k *wybierzemy* (p_k, q_k) *takie, że*

$$|a^{(k)}_{p_k q_k}| = \max\{|a^{(k)}_{ij}| : i \neq j\},$$

to metoda obrotów jest zbieżna.

Dowód. Niech

$$t_k^2 = \sum_{i \neq j} |a^{(k)}_{ij}|^2.$$

Na mocy poprzedniej uwagi

$$t^2_{k+1} = t^2_k - 2|a^{(k)}_{p_k q_k}|^2.$$

Skoro

$$|a^{(k)}_{p_k q_k}| \geq |a^{(k)}_{ij}|, \ i \neq j,$$

to

$$|a^{(k)}_{p_k q_k}|^2 \geq |a^{(k)}_{ij}|^2, \ i \neq j,$$

więc

$$t_k^2 = \sum_{i \neq j} |a^{(k)}_{ij}|^2 \leq n(n-1)|a^{(k)}_{p_k q_k}|^2.$$

$\frac{2}{n(n-1)} < 1$, $n > 2$

Mamy więc

$$\begin{aligned} t^2_{k+1} &\leq t^2_k - \tfrac{2t_k^2}{n(n-1)} = t^2_k\big(1 - \tfrac{2}{n(n-1)}\big) \leq \ldots \leq t^2_0\big(1 - \tfrac{2}{n(n-1)}\big)^{k+1} \\ &\leq N^2(A)\big(1 - \tfrac{2}{n(n-1)}\big)^{k+1}, \end{aligned}$$

zatem

$$\lim_{k \to \infty} t_k^2 = 0,$$

więc

$$\lim_{k \to \infty} A^{(k)} = \operatorname{diag}(d_1, \ldots, d_n).$$

□

3.3.2 Metoda QR wyznaczania wartości własnych macierzy.

Jak pokazaliśmy w twierdzeniu o faktoryzacji QR (tw. 2.9), jeśli macierz A jest nieosobliwa, to istnieją macierze Q – ortogonalna i R – trójkątna górna takie, że

$$A = QR.$$

Z dowodu tego twierdzenia wynika, że rozkład ten jest jednoznaczny.

Załóżmy o macierzy Q więcej, że jest ortonormalna, czyli

$$Q^T Q = I.$$

Zauważmy, że skoro $Q^T Q = D^2$ to $D^{-1} Q^T Q D^{-1} = I$. Zatem jako naszą macierz ortonormalną wystarczy wziąć QD^{-1}.

Jeśli macierz A jest osobliwa, to można otrzymać podobny rozkład z tym, że macierz R też będzie osobliwa (metoda Householdera).

Przejdźmy teraz do samej metody QR wyznaczania wartości własnych macierzy A. Załóżmy, że A jest nieosobliwa i niech $A_0 = A$. Istnieją zatem macierze Q_0 – ortonormalna i R_0 – trójkątne górna takie, że $A_0 = Q_0 R_0$. Zdefiniujmy ciąg

$$A_k = R_{k-1} Q_{k-1}, \ k = 1, 2, \ldots \tag{55}$$

Lemat 3.22. *Dla ciągu A_k zdefiniowanego wzorem (55)*

$$A_k \sim A_{k+1}, \ k = 0, 1, \ldots$$

Dowód. Skoro $A_k = Q_k R_k$ to $R_k = Q_k{}^T A_k$ albo $Q_k = A_k R_k^{-1}$. Więc

$$A_{k+1} = R_k Q_k = Q_k{}^T A_k Q_k,$$

albo

$$A_{k+1} = R_k Q_k = R_k A_k R_k^{-1}.$$

□

Pytanie jakie należy sobie teraz postawić, to czy ciąg $\{A_k\}_{k=0}^{\infty}$ jest zbieżny, a jeśli tak to do czego. Zauważmy, że

$$A_{k+1} = Q_k{}^T A_k Q_k = Q_k{}^T Q_{k-1}{}^T A_{k-1} Q_{k-1} Q_k = \ldots = Q_k{}^T \ldots Q_0{}^T A_0 Q_0 \ldots Q_k,$$

a więc

$$A_{k+1} = U_k{}^T A U_k, \ U_k = Q_0 \ldots Q_k.$$

Podobnie otrzymamy, że

$$A_{k+1} = G_k A G_k^{-1}, \ G_k = R_k \ldots R_0.$$

Oczywiście U_k jest macierzą ortogonalną, a R_k macierzą trójkątną górną. Co więcej, jeśli $A_0{}^* = A_0$ lub A_0 jest trójprzekątniowa, to $A_k{}^* = A_k$ lub odpowiednio A_k jest trójprzekątniowa (dowód ze względu na indukcje – nie wymagany).

Pokażemy jeszcze, że $U_k G_k = A^{k+1}$. Istotnie

$$\begin{aligned}
&A_k = U_{k-1}{}^T A U_{k-1} \Rightarrow U_{k-1} A_k = A U_{k-1}, \\
&U_k G_k = Q_0 \ldots Q_{k-1} Q_k R_k R_{k-1} \ldots R_0 = Q_0 \ldots Q_{k-1} A_k R_{k-1} \ldots R_0 = U_{k-1} A_k G_{k-1} \\
&\qquad = A U_{k-1} G_{k-1} = \ldots = A^{k+1}.
\end{aligned}$$

Twierdzenie 3.23. *Jeśli ciąg* $\{U_k\}_k$ *jest zbieżny, to ciąg* $\{A_k\}_k$ *jest zbieżny do macierzy trójkątnej górnej.*

Dowód. Niech $U_k = Q_0 \dots Q_{k-1}Q_k$, więc (ze zbieżności $\{U_k\}_k$) mamy

$$\begin{aligned} &U_k = U_{k-1}Q_k, \\ &Q_k = {U_{k-1}}^T U_k, \\ &\lim_{k\to\infty} Q_k = (\lim_{k\to\infty} {U_{k-1}}^T)(\lim_{k\to\infty} U_k) = (\lim_{k\to\infty} U_{k-1})^T \lim_{k\to\infty} U_k = I, \end{aligned}$$

oraz

$$\begin{aligned} &A_k = R_{k-1}Q_{k-1}, \\ &R_{k-1} = A_k {Q_{k-1}}^T = {U_{k-1}}^T A U_{k-1} {Q_{k-1}}^T = {U_{k-1}}^T A Q_0 \dots Q_{k-2}Q_{k-1}{Q_{k-1}}^T \\ &\qquad = {U_{k-1}}^T A U_{k-2}. \end{aligned}$$

Więc ciągi $\{Q_k\}_k$ i $\{R_k\}_k$ są zbieżne, zatem ciąg $\{A_k\}_k$ też jest zbieżny.

Oznaczmy przez A_∞ granicę ciągu $\{A_k\}_k$ oraz przez R_∞ granicę ciągu $\{R_k\}_k$, wówczas

$$A_\infty = \lim_{k\to\infty} A_k = (\lim_{k\to\infty} Q_k)(\lim_{k\to\infty} R_k) = \lim_{k\to\infty} R_k = R_\infty.$$

Macierz R_∞ (a tym samym A_∞) jest macierzą trójkątną górną (bo R_k są trójkątne rórne), co kończy dowód. □

Wniosek 3.24. *Niech ciąg* $\{U_k\}_k$ *będzie zbieżny i niech* $\tilde{A} = \lim_{k\to\infty} A_k$. *Wówczas* $\tilde{a}_{ii}$ *są wartościami własnymi macierzy* A.

Dowód. Twierdzenie 3.23 + lemat 3.22 + wniosek 2.22. □

Uwaga 3.25. *Jeśli macierz* A *jest symetryczna i dodatnio określona, to ciąg* $\{U_k\}_k$ *jest zbieżny.*

4 Interpolacja.

Dana niech będzie funkcja $f : [a, b] \to \mathbb{R}$ oraz $(n+1)$-parametrowa rodzina funkcji $\Phi(x; c_0, \dots, c_n) : [a, b] \times \mathbb{R}^{n+1} \to \mathbb{R}$. Dla zadanych $n+1$ punktów $x_0, \dots, x_n \in [a, b]$ będziemy chcieli dobrać parametry $c_0, \dots, c_n$ tak, aby zachodziły równości

$$f(x_i) = \Phi(x_i; c_0, \dots, c_n),\ i = 0, \dots, n. \tag{56}$$

Punkty $x_0, \dots, x_n$ nazywamy **węzłami interpolacyjnymi**, natomiast równości (56) – **warunkami interpolacyjnymi. Interpolacją** nazywamy funkcję Φ.

Tak więc chcemy dobrać parametry $c_0, \dots, c_n$ tak, aby wykres Φ zawierał punkty $(x_j, f(x_j)),\ j = 0, \dots, n$.

Jeżeli Φ jest funkcją liniową zmiennych $c_0, \dots, c_n$, to interpolację nazywamy **liniową**:

$$\Phi(x; c_0, \dots, c_n) = \sum_{j=0}^{n} c_j \varphi_j(x).$$

Dla interpolacji liniowej warunki interpolacyjne przyjmują postać

$$(57)\quad f(x_i) = \sum_{j=0}^{n} c_j \varphi_j(x_i),\ i = 0, \dots n.$$

Przykład 4.1.

(i) Interpolacja wielomianowe: $\varphi_j(x) = x^j,\ j = 0, \dots, n,$

(ii) Interpolacja trygonometryczna: $\varphi_j(x) = e^{ijx},\ j = 0, \dots, n,\ i = \sqrt{-1}.$

Uwaga 4.2. *Układ (57) jest układem* $(n+1)$ *równań liniowych o niewiadomej*

$$c = (c_0, \dots, c_n)^T \in \mathbb{R}^{n+1},$$

wyrazie wolnym

$$f = (f(x_0), \dots, f(x_n))^T \in \mathbb{R}^{n+1}$$

oraz macierzy

$$A = (a_{ij}),\ a_{ij} = \varphi_j(x_i).$$

Równoważnie zapisać go można jako

$$Ac = f.$$

Zatem warunek ten równoważny jest

$$f \in \operatorname{span}\{\varphi_0, \varphi_1, \dots, \varphi_n\}.$$

Wówczas parametry c_j *są współrzędnymi* f *względem bazy* $\{\varphi_j\}$.

Twierdzenie 4.3. *Problem interpolacji liniowej ma dla każdego wektora* $y \in \mathbb{R}^{n+1}$ *dokładnie jedno rozwiązanie wtedy i tylko wtedy, gdy* $\det(\varphi_j(x_i)) \neq 0$ *(*A *jest macierzą nieosobliwą).*

Uwaga 4.4.

(i) Twierdzenie 4.3 jest prawdziwe także dla x_i *oraz* $f(x_i)$ *zespolonych.*

(ii) Do wyznaczenia c_j *wystarczy znajomość* x_i *oraz* $f(x_i),\ i = 0, \dots, n.$

4.1 Interpolacja wielomianowa.

Niech

$$\Pi_n = \{w |\ w(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0,\ a_j \in \mathbb{C}\}$$

oznacza **przestrzeń wielomianów stopnia co najwyżej n**.

Uwaga 4.5.

(i) Π_n *jest przestrzenią wektorową.*

(ii) $\dim \Pi_n = n + 1$.

Dowód.

ad*(i)* $v, w \in \Pi_n \Rightarrow \alpha v + \beta w \in \Pi_n$.

ad*(ii)* Wystarczy pokazać, że $1, x, \ldots, x^n$ jest bazą Π_n. Oczywiści $\Pi_n = \operatorname{span}\{1, x, \ldots, x^n\}$. Niech $P(x) = \sum_{j=0}^{n} c_j x^j = 0$. Oczywiście $P(x) \in \Pi_n$ i ma on nieskończenie wiele zer, zatem na mocy zasadniczego twierdzenie algebry $c_j = 0,\ j = 0, \ldots, n$, czyli $1, x, \ldots, x^n$ są liniowo niezależne.

□

Zagadnienie interpolacji wielomianowej polega na tym, że dla danych par punktów $(x_i, f(x_i)),\ i = 0, \ldots n$ szukamy wielomianu $P \in \Pi_n$ takiego, że

(58) $P(x_i) = f(x_i),\ i = 0, \ldots, n.$

Twierdzenie 4.6. *(istnienie i jednoznaczność zagadnienia interpolacji wielomianowej) Jeżeli* $x_0, \ldots, x_n \in [a, b]$ *są parami różne, to dla dowolnego wektora wartości* $f = (f(x_0), \ldots, f(x_n))^T$ *problem (58) ma dokładnie jedno rozwiązanie.*

Dowód. Zacznijmy od istnienia takiego wielomianu. Niech

$$l_i(x) = \frac{(x - x_0) \ldots (x - x_{i-1})(x - x_{i+1}) \ldots (x - x_n)}{(x_i - x_0) \ldots (x_i - x_{i-1})(x_i - x_{i+1}) \ldots (x_i - x_n)} = \prod_{j \neq i} \frac{x - x_j}{x_i - x_j},$$

Delta Kroneckera.

zatem $l_i(x_j) = \delta_{ij}$. Niech

(59) $L(x) = \sum_{i=0}^{n} f(x_i) l_i(x)$ – **wielomian interpolacyjny Lagrange'a.**

Więc

$$L(x_j) = \sum_{i=0}^{n} f(x_i) l_i(x_j) = f(x_j),\ j = 0, \ldots, n.$$

Aby wykazać jedyność przypuśćmy, że $P, Q \in \Pi_n$ spełniają założenia twierdzenia. Wtedy $W(x) = P(x) - Q(x) \in \Pi_n$ zeruje się w $n + 1$ punktach $x_0, \ldots, x_n$. Na mocy zasadniczego twierdzenia algebry $W \equiv 0$, zatem $P = Q$. □

Uwaga 4.7. *W powyższym twierdzeniu zamiast przedziału* $[a, b] \subset \mathbb{R}$ *można rozpatrywać zbiór zwarty* $K \subset \mathbb{C}$.

Uwaga 4.8. *Układ* $\{l_j\}_j$ *tworzy bazę* Π_n.

Dowód. Oczywiście $\operatorname{span}\{l_0, \ldots, l_n\} = \Pi_n$. Wystarczy wykazać, że $l_0, \ldots, l_n$ są liniowo niezależne. Niech

$$P(x) = \sum_{j=0}^{n} d_j l_j(x) = 0.$$

Skoro $l_j(x_i) = \delta_{ij}$ to

$$P(x_i) = d_i, \ i = 0, \ldots, n.$$

Więc $d_0 = d_1 = \ldots = d_n = 0$, co kończy dowód. □

Z istnienia i jednoznaczności interpolacji wielomianowej dla $x_i \neq x_j$ dostajemy, że

$$\det(x_i^j) = V(x_0, \ldots, x_n) = \det \begin{pmatrix} 1 & x_0 & \ldots & x_0^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \ldots & x_n^n \end{pmatrix} \neq 0.$$

Dzieje się tak dlatego, że dla bazy Lagrange'a dostajemy macierz jednostkową I_{n+1}.

Wyznacznik $V(x_0, \ldots, x_n)$ nazywamy **wyznacznikiem Vandermonde'a**.

Jak widać po dodaniu nowego węzła, aby znaleźć wielomian interpolacyjny Lagrange'a należy ponownie wyznaczyć wielomiany $l_j(x)$, $j = 0, \ldots, n+1$. Spróbujmy pozbyć się tej niedogodności. Jako bazę Π_n przyjmijmy

$$\begin{aligned} &p_0(x) = 1, \\ &p_1(x) = x - x_0, \\ &\vdots \\ &p_n(x) = (x - x_0)(x - x_1) \ldots (x - x_{n-1}). \end{aligned}$$

Oczywiście $\{p_j\}_{j=0}^n \subset \Pi_n$ oraz $\Pi_n = \operatorname{span}\{p_0, \ldots, p_n\}$. Niech

$$P(x) = \sum_{s=0}^{n} b_s p_s(x)$$

oznacza **wielomian interpolacyjny Newtona**.

Korzystając z uwagi 4.2 (dla $c = (b_0, \ldots, b_n)$) macierz A przyjmuje postać

$$A = (p_j(x_i)) = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 1 & x_1 - x_0 & 0 & \ldots & 0 \\ 1 & x_2 - x_0 & (x_2 - x_0)(x_2 - x_1) & & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n - x_0 & (x_n - x_0)(x_n - x_1) & \ldots & (x_n - x_0) \cdot \ldots \cdot (x_n - x_{n-1}) \end{pmatrix}$$

Zatem, jako wniosek z twierdzenia 4.6 otrzymujemy

Twierdzenie 4.9. *Jeśli $x_i \neq x_j$ dla $i \neq j$ to wielomian interpolacyjny w postaci Newtona jest wyznaczony jednoznacznie. Liczby b_j wyznacza się z układu równań o macierzy trójkątnej.*

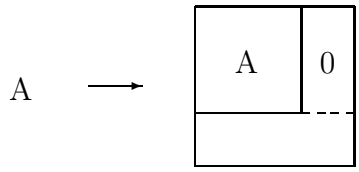
Zaletą postaci Newtona jest to, że dodanie nowego węzła x_{n+1} (takiego, że $x_{n+1} \neq x_j$, $j = 0, \ldots, n$) nie zmienia wartości $b_0, \ldots, b_n$. Istotnie, oznaczmy przez $P_{0..n}(x)$ (dowolny) wielomian interpolacyjny przechodzący przez węzły $x_0, \ldots, x_n$, wówczas

$$P_{0..n+1}(x) = P_{0..n}(x) + b_{n+1}p_{n+1}(x),$$

gdzie

$$p_{n+1}(x) = (x - x_0) \ldots (x - x_n).$$

Dzieje się tak dlatego, że macierz A zostaje powiększona o jeden wiersz i jedną kolumnę



Rysunek 8:

Istnieją wzory na współczynniki b_j wielomianu interpolacyjnego Newtona. Związane są one z ilorazami różnicowymi.

4.2 Ilorazy różnicowe.

Dana niech będzie funkcja $f : [a, b] \to \mathbb{R}$ oraz podział $\Delta_n = \{x_0, \ldots, x_n\}$ przedziału $[a, b]$ ($x_i \neq x_j$ dla $i \neq j$). Definiujemy **ilorazy różnicowe** następująco:

$$f[x_i] = f(x_i),$$

$$f[x_i, x_{i+1}] = \frac{f[x_{i+1}] - f[x_i]}{x_{i+1} - x_i},$$

$$\vdots$$

$$f[x_i, x_{i+1}, \ldots, x_{i+k}] = \frac{f[x_{i+1}, x_{i+2}, \ldots, x_{i+k}] - f[x_i, x_{i+1}, \ldots, x_{i+k-1}]}{x_{i+k} - x_i}.$$

W celu obliczania kolejnych ilorazów różnicowych najlepiej korzystać ze schematu

$$\begin{array}{c|cccc}
x_0 & f[x_0] & & & \\
 & & f[x_0, x_1] & & \\
x_1 & f[x_1] & & f[x_0, x_1, x_2] & \\
 & & f[x_1, x_2] & & f[x_0, x_1, x_2, x_3] \\
x_2 & f[x_2] & & f[x_1, x_2, x_3] & \\
 & & f[x_2, x_3] & & \\
x_3 & f[x_3] & & &
\end{array}$$

Twierdzenie 4.10. *(współczynniki wielomianu interpolacyjnego w postaci Newtona) Jeśli* $P_{i..i+k}(x)$ *jest wielomianem interpolacyjnym Newtona przechodzącym przez węzły* $x_i, \ldots, x_{i+k}$, *to*

$$P(x)_{i..i+k} = f[x_i] + f[x_i, x_{i+1}](x - x_i) + \ldots + f[x_i, \ldots, x_{i+k}](x - x_i) \ldots (x - x_{i+k-1}).$$

W szczególności

$$P_{0..n}(x) = \sum_{s=0}^{n} b_s p_s(x),\ b_s = f[x_0, \ldots, x_s].$$

Dowód. Indukcja względem k. Dla $k = 0$ wzór jest oczywiście prawdziwy. Załóżmy, że wzór zachodzi dla $k-1$ punktów. Niech

$$\beta(x) = P_{i..i+k-1}(x),$$
$$\gamma(x) = P_{i+1..i+k}(x)$$

i niech

$$\alpha(x) = \frac{\beta(x)(x_{i+k} - x) + \gamma(x)(x - x_i)}{x_{i+k} - x_i}$$

będzie wypukłą kombinacją tych wielomianów (o współczynnikach $\frac{x_{i+k}-x}{x_{i+k}-x_i}$, $\frac{x-x_i}{x_{i+k}-x_i}$). Dla $i < j < i + k$ mamy

$$\alpha(x_j) = \frac{f(x_j)(x_{i+k} - x_j) + f(x_j)(x_j - x_i)}{x_{i+k} - x_i} = f(x_j),$$

dla $j = i$

$$\alpha(x_i) = \frac{P_{i..i+k-1}(x_i)(x_{i+k} - x_i)}{x_{i+k} - x_i} = f(x_i),$$

dla $j = i + k$

$$\alpha(x_{i+k}) = \frac{P_{i+1..i+k}(x_{i+k})(x_{i+k} - x_i)}{x_{i+k} - x_i} = f(x_{i+k}).$$

Zatem wielomian $\alpha(x)$ jest wielomianem interpolacyjnym (bo spełania warunki interpolacyjne $\alpha(x_j) = f(x_j)$), więc

$$\alpha(x) = P_{i..i+k}(x).$$

Zatem (oznaczając przez r „pozostałe wyrazy“) mamy

$$\alpha(x) = b_k x^k + r.$$

Wystarczy wykazać, że $b_k = f[x_i, \ldots, x_{i+k}]$. Na mocy założenia indukcyjnego

$$\beta(x) = f[x_i, \ldots, x_{i+k-1}]x^{k-1} + r_\beta,$$
$$\gamma(x) = f[x_{i+1}, \ldots, x_{i+k}]x^{k-1} + r_\beta.$$

Na mocy definicji wielomianu $\alpha(x)$ oraz założenia indukcyjnego otrzymujemy

$$\begin{aligned}\alpha(x) &= \tfrac{1}{x_{i+k}-x_i}(\beta(x)(x_{i+k} - x) + \gamma(x)(x - x_i)) \\ &= \tfrac{1}{x_{i+k}-x_i}[(-f[x_i, \ldots, x_{i+k-1}] + f[x_{i+1}, \ldots, x_{i+k}])x^k + r] \\ &= f[x_i, \ldots, f_{i+k}]x^k + r.\end{aligned}$$

□

Twierdzenie 4.11. *(Własności ilorazów różnicowych)*

(i) Dla dowolnej permutacji σ *zbioru* $\{i, \ldots, i+k\}$*:*

$$f[x_i, \ldots, x_{i+k}] = f[x_{\sigma(i)}, \ldots, x_{\sigma(i+k)}].$$

(ii) Iloraz różnicowy jest liniowy, to znaczy

$$(\alpha f + \beta g)[x_i, \ldots, x_{i+k}] = \alpha f[x_i, \ldots, x_{i+k}] + \beta g[x_i, \ldots, x_{i+k}].$$

(iii) Jeżeli $f \in \Pi_N,\ f_j \equiv f(x_j),\ j = i, \ldots, i+k$ *to*

$$f[x_i, \ldots, x_{i+k}] = 0 \text{ dla } k > N.$$

$\widehat{x_j}$ – *element opuszczony*

(iv) Dla $s \neq t$

$$f[x_i, \ldots, x_{i+k}] = \frac{f[x_i, \ldots, \widehat{x_s}, \ldots, x_{i+k}] - f[x_i, \ldots, \widehat{x_t}, \ldots, x_{i+k}]}{x_t - x_s}.$$

Dowód.

ad*(i)* Zmiana numeracji węzłów nie wpływa na wartości wielomianów.

ad*(ii)* Indukcja względem k.

ad*(iii)*

$$\begin{aligned} f(x) &= a_N x^N + \ldots + a_0 = 0 \cdot x^k + \ldots + 0 \cdot x^{N+1} + a_N x^N + \ldots + a_0 \\ &= f[x_0, \ldots, x_k] x^k + \ldots \Rightarrow f[x_0, \ldots, x_k] = 0. \end{aligned}$$

ad*(iv)* Wynika z definicji i punktu *(i)*.

□

4.3 Wielomiany Hermite'a.

Dana niech będzie funkcja $f : [a, b] \to \mathbb{R}$ dostatecznie regularna. Niech $\Delta = \{x_0, \ldots, x_n\} \subset [a, b]$ będzie takim podziałem, że $x_i \neq x_j$ dla $i \neq j$. Szukamy wielomianu $H_N \in \Pi_N$ spełniającego warunki

$$H_N^{(j)}(x_i) = f^{(j)}(x_i),\ 0 \leq j \leq m_i - 1,\ i = 0, \ldots, n, \tag{60}$$

gdzie $\sum_{i=0}^{n} m_i = N + 1$. Wartość m_i nazywamy **krotnością węzła** x_i. Wielomian H_N nazywamy **wielomianem interpolacyjnym Hermite'a**.

Przypomnijmy, że $\bar{x}$ jest zerem m-krotnym wielomianu P, gdy

$$\begin{aligned} &P^{(i)}(\bar{x}) = 0,\ i = 0, \ldots, m-1, \\ &P^{(m)}(\bar{x}) \neq 0. \end{aligned}$$

Jeśli krotność każdego węzła wynosi 1, to H_N jest wielomianem Lagrange'a

Twierdzenie 4.12. *Jeśli* $x_s \neq x_t$ *dla* $s \neq t$*, to zagadnienie interpolacji Hermite'a (60) ma jednoznaczne rozwiązanie dla każdego układu* $\{f^{(j)}(x_i)\}$.

Dowód. Zacznijmy od jednoznaczności. Niech P, Q spełniają warunki twierdzenia. Zatem $W = P - Q \in \Pi_N$ oraz

$$W^{(j)}(x_i) = 0,\ i = 0, \ldots, n,\ 0 \leq j \leq m_i - 1.$$

Wobec tego W ma zera o łącznej krotności $N + 1$, więc na podstawie zasadniczego twierdzenia algebry $W \equiv 0$.

Przejdźmy teraz do istnienia. Niech

$$H_N(x) = \sum_{l=0}^{N} b_l p_l(x). \tag{61}$$

Jeśli $0 \leq l \leq N$, to istnieją $i \in \{0, \ldots, n\},\ j \in \{0, \ldots, m_i - 1\}$ takie, że

$$l = s(i) + j,$$

gdzie

$$s(i) = \begin{cases} 0, & i = 0 \\ m_0 + \ldots + m_{i-1}, & i > 0 \end{cases}.$$

Wtedy

$$\begin{aligned}
&p_0(x) = 1,\\
&p_1(x) = (x - x_0),\\
&\vdots\\
&p_{m_0}(x) = (x - x_0)^{m_0},\\
&p_{m_0+1}(x) = (x - x_0)^{m_0}(x - x_1),\\
&\vdots\\
&p_N(x) = (x - x_0)^{m_0} \ldots (x - x_n)^{m_n - 1},\\
&p_{s(i)+j}(x) = (x - x_0)^{m_0} \ldots (x - x_{i-1})^{m_{i-1}}(x - x_j)^j.
\end{aligned}$$

Z postaci (61) wielomianu H_N mamy, że

$$H_N^{(j)}(x_i) = \sum_{l=0}^{N} b_l p_l^{(j)}(x_i).$$

Dostajemy więc układ

$$\sum_{l=0}^{N} b_l p_l^{(j)}(x_i) = f^{(j)}(x_i), 0 \leq j \leq m_i - 1,\ i = 0, \ldots, n,$$

którego macierz jest trójkątna dolna. Wystarczy pokazać, że istnieje rozwiązanie tego układu. Oczywiście

$$b_0 = f(x_0).$$

Przypuśćmy, że znamy już $b_0, \ldots, b_{k-1}$ i przyjmijmy

$$A(x) = \sum_{l=0}^{k-1} b_l p_l(x) \text{ – znany wielomian,}$$
$$B(x) = \sum_{l=k+1}^{N} b_l p_l(x).$$

Wówczas

$$H_N(x) = A(x) + b_k p_k(x) + B(x). \tag{62}$$

Wystarczy więc, że znajdziemy $i \in \{0, \ldots, n\},\ j \in \{0, \ldots, m_i - 1\}$ takie, że

$$p_k^{(j)}(x_i) \neq 0,$$
$$B^{(j)}(x_i) = 0.$$

Jeżeli takie i, j znajdziemy, to z (62) otrzymamy, że

$$b_k = \frac{f^{(j)}(x_i) - A^{(j)}(x_i)}{p_k^{(j)}(x_i)}.$$

Wiemy, że istnieją $i \in \{0, \ldots, n\},\ j \in \{0, \ldots, m_i - 1\}$ takie, że

$$k = s(i) + j.$$

Wówczas

$$p_k(x) = (x - x_0)^{m_0} \ldots (x - x_{i-1})^{m_{i-1}} (x - x_i)^j.$$

Przyjmując

$$\alpha(x) = (x - x_0)^{m_0} \ldots (x - x_{i-1})^{m_{i-1}},$$
$$\beta(x) = (x - x_i)^j,$$

mamy

$$p_k(x) = \alpha(x)\beta(x).$$

Ze wzoru Leibniza dostajemy

$$p_k^{(j)}(x) = \sum_{t=0}^{j} \binom{j}{t} \alpha^{(t)}(x) \beta^{(j-t)}(x).$$

Oczywiście

$$\beta^{(t)}(x_i) = 0,\ t < j,$$
$$\beta^{(j)}(x_i) = j!,$$
$$\alpha(x_i) = (x_i - x_0)^{m_0} \ldots (x_i - x_{i-1})^{m_{i-1}} \neq 0,$$

zatem

$$p_k^{(j)}(x_i) = \alpha(x_i)\beta^{(j)}(x_i) \neq 0.$$

Z definicji wielomianu $B(x)$ otrzymujemy, że

$$B(x_i) = 0,$$

bo

$$p_l(x_i) = 0,\ l > k.$$

□

4.4 Reszta interpolacji wielomianu.

Lemat 4.13. *Jeśli* f *ma* $n+1$ *zer* $x_0, \ldots, x_n$ *w* (a, b) *o łącznej krotności* k*, to pochodna* f' *ma w* (a, b) *zera o łącznej krotności co najmniej równej* $k-1$*.*

Dowód. Jeśli f ma zero o krotności k, to f' ma zero o krotności $k-1$.

Jeśli f ma $n+1$ różnych miejsc zerowych $x_0, \ldots, x_n$ o krotnościach $m_0, \ldots, m_n$ ($m_0 + \ldots + m_n = k$), to są to miejsca zerowe f' o krotnościach $m_0 - 1, \ldots, m_n - 1$ odpowiednio. Wobec tego łączna krotność wynosi

$$\sum_{i=0}^{n}(m_i - 1) = \sum_{i=0}^{n} m_i - (n+1) = k - (n+1).$$

Ale z twierdzenia Rolle'a f ma zera w przedziałach (x_i, x_{i+1}), $i = 0, \ldots, n-1$, zatem łączna krotność zer f' jest co najmniej równa $k - (n+1) + n = k - 1$. □

Definicja 4.14. *Dana niech będzie funkcja* $f : [a, b] \to \mathbb{R}$*, dostatecznie regularna i dany niech będzie podział* $\Delta = \{x_0, \ldots, x_n\} \subseteq [a, b]$*. Jeżeli* $P(x)$ *oznacza wielomian interpolacyjny w postaci Lagrange'a oparty na podziale* Δ*, a* $H_N(x)$ *oznacza wielomian Hermite'a oparty na podziale* Δ *o krotnościach* $m_0, \ldots, m_n$ *odpowiednio, to dla* $\bar{x} \in [a, b]$ ***resztą interpolacyjną Lagrange'a*** *nazywamy*

(63) $r(\bar{x}) = f(\bar{x}) - P(\bar{x})$,

zaś ***resztą interpolacyjną Hermite'a*** *nazywamy*

(64) $r(\bar{x}) = f(\bar{x}) - H_N(\bar{x})$.

Twierdzenie 4.15. *(o reszcie)*
Jeżeli funkcja f *jest klasy* $n+1$ *w przypadku interpolacji Lagrange'a, bądź klasy* $N+1$ *w przypadku interpolacji Hermite'a, to dla* $\bar{x} \in [a, b]$ *istnieje* $\xi \in I(\bar{x}, x_0, \ldots, x_n)$ *(najmniejszy przedział zawierający punkty* $\bar{x}, x_0, \ldots, x_n$*) taki, że*

(i) $r(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(\bar{x})$*, gdzie* $\omega(x) = (x - x_0)\ldots(x - x_n)$ *w przypadku interpolacji Lagrange'a,*

(ii) $r(\bar{x}) = \frac{f^{(N+1)}(\xi)}{(N+1)!} p_{N+1}(\bar{x})$, *gdzie* $p_{N+1}(x) = (x - x_0)^{m_0} \dots (x - x_n)^{m_n}$ *w przypadku interpolacji Hermite'a.*

Dowód. Wykażemy *(i)* (w drugim przypadku należy przeprowadzić podobne rozumowanie).

Niech L będzie wielomianem interpolacyjnym w postaci Lagrange'a dla funkcji f opartym o węły $x_0, \dots, x_n$. Niech

$$F(x) = f(x) - L(x) - K\omega(x),$$

gdzie K dobieramy tak, aby $F(\bar{x}) = 0$. Skoro

$$F(x_i) = f(x_i) - L(x_i) - K\omega(x_i) = 0,$$

to F ma w $[a, b]$ zera o łącznej krotności $n + 2$ w przypadku interpolacji Lagrange'a (w przypadku interpolacji Hermite'a – $N + 2$). Na mocy lematu 4.13 $F^{(n+1)}$ zeruje się w co najmniej jednym punkcie, tzn.:

$$\exists\, \xi \in I(\bar{x}, x_0, \dots, x_n) :\ F^{(n+1)}(\xi) = 0.$$

Z drugiej strony

$$F^{(n+1)}(x) = f^{(n+1)}(x) - L^{(n+1)}(x) - K\omega^{(n+1)}(x) = f^{(n+1)}(x) + K(n+1)!,$$

(w przypadku interpolacji Hermite'a $p^{(N+1)}(x) = (N+1)!$). Stąd

$$f^{(n+1)}(\xi) - K(n+1)! = 0 \Rightarrow K = \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

czyli

$$r(\bar{x}) = \frac{f^{(n+1)}(\xi)}{(n+1)!}\omega(\bar{x}).$$

□

Twierdzenie 4.16. *Jeżeli funkcja* f *spełnia założenia poprzedniego twierdzenia, to*

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!},\ \textit{dla pewnego}\ \xi \in I(x_0, \dots, x_n) \tag{65}$$

Dowód. Oznaczmy przez $P_{0..k}(x)$ wielomian interpolacyjny w postaci Newtona dla funkcji f, oparty na punktach $x_0, \dots, x_k$. Na mocy twierdzenia 4.10 wiemy, że

$$P_{0..n}(x) = P_{0..n-1}(x) + f[x_0, \dots, x_n](x - x_0) \dots (x - x_{n-1}).$$

Ale $P_{0..n}$ jest również wielomianem interpolacyjnym dla funkcji $P_{0..n-1}$, wobec tego odpowiednia reszta w punkcie x_n jest równa

$$r(x_n) = P_{0..n}(x_n) - P_{0..n-1}(x_n) = f[x_0, \dots, x_n](x_n - x_0) \dots (x_n - x_{n-1}).$$

Z drugiej strony

$$r(x_n) \overset{tw.\ 4.15}{=} \frac{f^{(n)}(\xi)}{n!}(x_n - x_0)\dots(x_n - x_{n-1}),$$

zatem

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!}.$$

□

Zauważmy, że skoro

$$r(\bar{x}) = f(\bar{x}) - P(\bar{x}) = \frac{f^{(n)}(\xi)}{n!}(\bar{x} - x_0)\dots(\bar{x} - x_n),$$

to jeśli $\bar{x} \notin [x_0, x_n]$, to błąd $r(\bar{x})$ szybko rośnie.

Twierdzenie 4.17. *(Faber)*
Załóżmy, że $\Delta_n \subset [a, b]$ *jest takim podziałem, że* $\Delta_n = \{x_0^m, \dots, x_n^m\}$, P_n *jest wielomianem interpolacyjnym dla* $f \in \mathcal{C}([a, b], \mathbb{R})$ *skojarzonym z* Δ_n. *Wówczas dla dowolnego ciągu podziałów* $\{\Delta_n\}_n$ *istnieje* $f \in \mathcal{C}([a, b], \mathbb{R})$ *taka, że* P_n *nie zmierza jednostajnie w* $[a, b]$ *do* f.

4.5 Węzły równoodległe

Niech

$$x_i = x_0 + ih,\ i = 0, \pm 1,\ \pm 2, \dots$$

będzie siatką na $\mathbb{R}$, oraz niech $f : \mathbb{R} \to \mathbb{R}$.

Definicja 4.18. ***Różnicą progresywną*** *nazywać będziemy operator*

$$\Delta f(x) = f(x + h) - f(x). \tag{66}$$

Różnicą wsteczną *nazywać będziemy operator*

$$\nabla f(x) = f(x) - f(x - h). \tag{67}$$

Operatorem przesunięcia *będziemy nazywać*

$$Ef(x) = f(x + h),\ E^{-1}f(x) = f(x - h). \tag{68}$$

Wprowadźmy jeszcze następujące oznaczenia:

- I – identyczność, tzn.: $If(x) = f(x)$,
- $\Delta^k = \Delta(\Delta^{k-1}),\ \Delta^0 f(x) = f(x)$,
- $\nabla^k = \nabla(\nabla^{k-1}),\ \nabla^0 f(x) = f(x)$.

Uwaga 4.19.

(i) Operatory $\Delta,\ \nabla,\ \mathsf{E}$ *oraz* E^{-1} *są liniowe, to znaczy*

$$\bullet(\alpha f + \beta g)(x) = \alpha \bullet f(x) + \beta \bullet g(x),\ \bullet \in \{\Delta,\ \nabla,\ \mathsf{E},\ \mathsf{E}^{-1}\}.$$

(ii)

$$f[x_0, \ldots, x_k] = \frac{\Delta^k f(x_0)}{k! h^k}.$$

(iii) Jeżeli $f_i = f(x_i)$*, to*

$$\Delta^r f_k = \sum_{s=0}^{r} (-1)^s \binom{r}{s} f_{k+r-s},$$

$$\nabla^r f_k = \sum_{s=0}^{r} (-1)^s \binom{r}{s} f_{k-s}.$$

(iv)

$$\Delta(f(x)g(x)) = g(x+h)\Delta f(x) + f(x)\Delta g(x) = g(x)\Delta f(x) + f(x+h)\Delta g(x),$$

$$\sum_{s=a}^{b} \Delta f_s = f_{b+1} - f_a,$$

$$\sum_{s=a}^{b} f_s \Delta g_s = f_s g_s\Big|_a^{b+1} - \sum_{s=a}^{b} g_{s+1} \Delta f_s.$$

Dowód.

ad*(ii)* Indukcja względem k. Oczywiście

$$f[x_0] = f(x_0) = \frac{\Delta^0 f(x_0)}{0! h^0}.$$

Załóżmy, że twierdzenie jest prawdziwe dla k, wówczas

$$\begin{aligned} f[x_0, \ldots, x_{k+1}] &= \frac{f[x_1, \ldots, x_{k+1}] - f[x_0, \ldots, x_k]}{x_{k+1} - x_0} \\ &= \frac{1}{x_{k+1} - x_0}\Big(\frac{\Delta^k f(x_1)}{k! h^k} - \frac{\Delta^k f(x_0)}{k! h^k}\Big) = \frac{\Delta^k f(x_1) - \Delta^k f(x_0)}{(k+1)! h^{k+1}} \\ &= \frac{(\Delta^{k+1} f)(x_0)}{(k+1)! h^{k+1}} \end{aligned}$$

$x_{k+1} = x_0 + h(k+1)$

ad*(iii)*

$$\Delta f(x) = f(x+h) - f(x) = (\mathsf{E} - \mathrm{I}) f(x),$$

$$\Delta^r f_k = (\mathsf{E} + (-1)\mathrm{I})^r f_k = \sum_{s=0}^{r} \binom{r}{s} \mathsf{E}^{r-s} (-\mathrm{I})^s f_k = \sum_{s=0}^{r} \binom{r}{s} (-1)^s f_{k+r-s}.$$

$\mathrm{I}\mathsf{E} = \mathsf{E}\mathrm{I}$

Podobnie dowodzi się drugi wzór, z tym, że

$$\nabla f(x) = f(x) - f(x-h) = (\mathrm{I} - \mathsf{E}^{-1}) f(x).$$

ad*(iv)* Są to wzory analogiczne do wzorów na pochodną iloczynu (różnica iloczynu) i całkowania przez części (sumowanie przez części). Wprost z definicji

$$\begin{aligned}\Delta(f(x)g(x)) &= f(x+h)g(x+h) - f(x)g(x)\\ &= f(x+h)g(x+h) - f(x)g(x+h) + f(x)g(x+h) - f(x)g(x)\\ &= (f(x+h) - f(x))g(x+h) + f(x)(g(x+h) - g(x))\\ &= g(x+h)\Delta f(x) + f(x)\Delta g(x),\end{aligned}$$

$$\begin{aligned}\sum_{s=a}^{b} \Delta f_s &= \Delta f_a + \ldots + \Delta f_s = (f_{a+1} - f_a) + (f_{a+2} - f_{a+1}) + \ldots + (f_{b+1} - f_b)\\ &= f_{b+1} - f_a.\end{aligned}$$

Aby wykazać ostatni wzór skorzystamy ze wzoru na różnicę iloczynu

$$\Delta(f_s g_s) = f_s \Delta g_s + g_{s+1}\Delta f_s.$$

Obkładając tą równość obustronnie przez $\sum_{s=a}^{b}$ mamy

$$f_s g_s\Big|_a^{b+1} = \sum_{s=a}^{b} \Delta(f_s g_s) = \sum_{s=a}^{b} f_s \Delta g_s + \sum_{s=a}^{b} g_{s+1}\Delta f_s.$$

□

4.6 Potęga symboliczna (wielomian czynnikowy)

Dla $n \in \mathbb{N}$ definiujemy **potęgę symboliczną**[13] wzorem

$$\begin{aligned}x^{[0]} &= 1,\\ x^{[n]} &= x(x-1)\cdot\ldots\cdot(x-n+1).\end{aligned}$$

Twierdzenie 4.20. *Potęga symboliczna ma następujące własności:*

$0 \notin \mathbb{N}$

(i) $x^{[k]} = 0$, *gdy* $k - x \in \mathbb{N}$ *i* $x \geq 0$,

(ii) $x^{[n+k]} = x^{[n]} \cdot (x-n)^{[k]}$,

(iii) $\Delta x^{[n]} = n x^{[n-1]}$.

Dowód.

ad*(i)* Niech x będzie takie, że $k - x \in \mathbb{N}$, tzn. $0 \leq x \leq k-1$ i $x \in \mathbb{N}$. Wówczas

$$x^{[k]} = x(x-1)\ldots(x-k+1) = 0.$$

ad*(ii)* Z definicji.

[13] W literaturze nazywana jest też potęgą ubywającą i oznaczana jako $x^{\underline{n}}$. Analogicznie można zdefiniować $x^{\overline{n}}$

ad*(iii)* Przyjmijmy $h = 1$, wówczas $\Delta f(x) = f(x+1) - f(x)$ i

$$\begin{aligned}\Delta x^{[n]} &= (x+1)^{[n]} - x^{[n]} \\ &= (x+1)x(x-1)\dots(x+1-n+1) - x(x-1)\dots(x-n+1) \\ &= x(x-1)\dots(x-n+2)[(x+1)-(x-n+1)] = nx^{[n-1]}.\end{aligned}$$

□

Wniosek 4.21. *Jeśli* $(x+n)^{[n]} \neq 0$*, to*

$$x^{[-n]} = \frac{1}{(x+n)^{[n]}}.$$

Dowód. Wystarczy skorzystać z poprzedniego twierdzenia podpunkt *(ii)* dla $n = -n$ i $k = n$. □

Przykład 4.22.

1. Dla $k+1 \neq 0$

$$\sum_{s=a}^{b} s^{[k]} = \sum_{s=a}^{b} \Delta \frac{s^{[k+1]}}{k+1} = \left.\frac{s^{[k+1]}}{k+1}\right|_a^{b+1}$$

2.

$$\begin{aligned}S_n &= 1\cdot 3 + 2\cdot 4 + \ldots + n(n+2) = \sum_{s=1}^{n} s(s+2) = \sum_{s=1}^{n} (3s + s(s-1)) \\ &= \sum_{s=1}^{n} (3s^{[1]} + s^{[2]})\end{aligned}$$

i dalej jak w poprzednim przykładzie.

3.

$$\sum_{s=1}^{n} \frac{1}{s(s+1)(s+2)} = \sum_{s=1}^{n} \frac{1}{(s-1+3)^{[3]}} = \sum_{s=1}^{n} (s-1)^{[-3]}$$

i dalej jak w przykładzie pierwszym.

4.7 Interpolacja trygonometryczna

Przypomnijmy, że w przypadku interpolacji trygonometrycznej

$$\Phi(x; c_0, \dots, c_n) = \sum_{j=0}^{n} c_j \varphi_j(x), \;\; \varphi_j(x) = e^{ijx} = \cos jx - i \sin jx.$$

Tak jak poprzednio daną mamy funkcję $f : [a, b] \to \mathbb{R}$ oraz $\Delta = \{x_0, \ldots, x_n\}$ – podział przedziału $[a, b]$ (węzły interpolacji). Szukać będziemy takich parametrów $c_0, \ldots, c_n$, aby

$$(69)\quad f(x_k) = \sum_{j=0}^{n} c_j e^{ijx_k},\ k = 0, \ldots, n.$$

Niech

$$f_k = f(x_k),\ k = 0, \ldots, n.$$

Twierdzenie 4.23. *Jeżeli $x_s \neq x_t$ dla $s \neq t$ oraz $x_k \in [0, 2\pi)$, to problem (69) ma jednoznaczne rozwiązanie dla każdego zespołu wartości f_k, $k = 0, \ldots, n$.*

Dowód. Problem ten jest równoważny problemowi

$$(70)\quad Ac = f,$$

gdzie

$$c = \begin{pmatrix} c_0 \\ \vdots \\ c_n \end{pmatrix},\ f = \begin{pmatrix} f_0 \\ \vdots \\ f_n \end{pmatrix},\ A = (e^{ijx_k})_{j,k}.$$

Aby udowodnić twierdzenie wystarczy wykazać, że $\det A \neq 0$. Niech więc $z_k = e^{ix_k}$, wówczas

$$A = \begin{pmatrix} 1 & z_0 & z_0^2 & \ldots & z_0^n \\ 1 & z_1 & z_1^2 & \ldots & z_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \ldots & z_n^n \end{pmatrix}$$

czyli A jest macierzą Vandermonde'a skojarzoną z wektorem $(z_0, \ldots, z_n)$, zatem jej wyznacznik $V(z_0, \ldots, z_n)$ jest niezerowy o ile $z_s \neq z_t$ dla $s \neq t$ (wynika to z jednoznaczności interpolacji wielomianowej). Niech więc $s \neq t$ i przypuśćmy, że

$$z_s = z_t \Leftrightarrow e^{ix_s} = e^{ix_t} \Leftrightarrow e^{i(x_s - x_t)} = 1.$$

Zatem $x_s - x_t = 2l\pi,\ l \in \mathbb{Z}$. Skoro $x_k \in [0, 2\pi)$, to $x_s - x_t \in (-2\pi, 2\pi)$, czyli $x_s = x_t$ – sprzeczność. □

Przyjrzyjmy się wyznacznikowi Vandermonde'a

$$V(z, z_1, \ldots, z_n) = \begin{vmatrix} 1 & z & z^2 & \ldots & z^n \\ 1 & z_1 & z_1^2 & \ldots & z_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \ldots & z_n^n \end{vmatrix}.$$

Jest to wielomian zmiennej z stopnia n. Niech $P(z) = V(z, z_1, \ldots, z_n)$. Wówczas

$$P(z_1) = \ldots = P(z_n) = 0,$$

(bo mamy dwa takie same wiersze). W takim razie

$$P(z) = a(z - z_1) \cdot \ldots \cdot (z - z_n) = az^n + \text{pozostałe},$$

gdzie

$$a = (-1)^{n+1} \begin{vmatrix} 1 & z_1 & z_1^2 & \ldots & z_1^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_n & z_n^2 & \ldots & z_n^{n-1} \end{vmatrix} =: (-1)^{n+1} Q(z_1).$$

Wtedy

$$Q(z) = b(z - z_2) \cdot \ldots \cdot (z - z_n) = bz^{n-1} + \text{pozostałe},$$

itd. Stąd

$$\begin{aligned} V(z_0, \ldots, z_n) = P(z_0) &= (z_0 - z_1) \ldots (z_0 - z_n)(-1)^{n+1}(z_1 - z_2) \ldots (z_1 - z_n) \\ &= k(z_0 - z_1) \ldots (z_0 - z_n)(z_1 - z_2) \ldots (z_1 - z_n) \ldots (z_{n-1} - z_n), \end{aligned}$$

gdzie k jest pewną stałą. Wobec tego

$$V(z_0, \ldots, z_n) \neq 0 \Leftrightarrow z_t \neq z_s, \ t \neq s.$$

Zajmijmy się przypadkiem gdy węzły są równoodległe, to znaczy

$$x_k = \frac{2\pi k}{n+1}, \ k = 0, \ldots, n. \tag{71}$$

Oczywiście

$$x_k \in [0, 2\pi), \ k = 0, \ldots, n.$$

Twierdzenie 4.24. *Jeżeli spełniony jest warunek (71), to dla* $A = (e^{ijx_k})_{j,k=0,\ldots,n}$ *spełniony jest wzór*

$$A^\star A = (n+1)I. \tag{72}$$

Dowód. Oznaczmy

$$\begin{aligned} &z_k = e^{ix_k} = \exp\left(\tfrac{2\pi ik}{n+1}\right), \\ &A = (a_{jk})_{jk}, a_{jk} = z_k^j = \exp\left(\tfrac{2\pi ijk}{n+1}\right), \\ &A^\star = (a^\star_{jk})_{jk}, \ a^\star_{jk} = \bar{a}_{kj} = \exp\left(-\tfrac{2\pi ijk}{n+1}\right). \end{aligned}$$

Niech $A^\star A = C = (c_{st})_{st}$. Zatem

$$c_{st} = \sum_{j=0}^{n} a^\star_{sj} a_{jt} = \sum_{j=0}^{n} \exp\left(-\tfrac{2\pi isj}{n+1}\right) \exp\left(\tfrac{2\pi itj}{n+1}\right) = \sum_{j=0}^{n} \exp\left(\tfrac{2\pi ij}{n+1}(t-s)\right) = \sum_{j=0}^{n} q^j,$$

gdzie $q = \exp\left(\frac{2\pi i}{n+1}(t-s)\right)$. Oczywiście

$$\sum_{j=0}^{n} q^j = \begin{cases} \frac{q^{n+1}-1}{q-1}, & q \neq 1 \\ n+1, & q = 1 \end{cases}$$

Zauważmy, że

$$q^{n+1} = \exp\left(\tfrac{2\pi}{n+1} i(t-s)(n+1)\right) = \exp\left(2\pi i(t-s)\right) = 1,$$

zatem

$$c_{st} = \tfrac{q^{n+1}-1}{q-1} = 0,\ s \neq t,$$
$$c_{ss} = n+1.$$

Stąd

$$C = (n+1)I.$$

□

Wniosek 4.25. *Rozwiązaniem problemu (70) jest*

$$c = \frac{1}{n+1} A^{\star} f.$$

Dowód.

$$Ac = f,$$
$$A^{\star}Ac = A^{\star}f,$$
$$(n+1)c = A^{\star}f.$$

□

Możemy więc zapisać wzory na szukane przez nas c_k:

$$c_k = \frac{1}{n+1}\sum_{j=0}^{n} a_{kj}^{\star} f_j = \frac{1}{n+1}\sum_{j=0}^{n} f_j \exp\left(-\tfrac{2\pi}{n+1} ijk\right) = \frac{1}{n+1}\sum_{j=0}^{n} f_j \exp\left(-ijx_k\right)$$
$$= \frac{1}{n+1}(f|a_k),$$

gdzie a_k jest k-tą kolumną macierzy A,

$$(u|v) = \sum_{j=0}^{n} u_j \bar{v}_j,\ u, v \in \mathbb{C}^{n+1}$$

oznacza zespolony iloczyn skalarny. Dostajemy zatem dwa równoważne wzory

$$c_k = \frac{1}{n+1}\sum_{j=0}^{n} f_j \bar{a}_{jk},\ k = 0, \ldots, n, \tag{73}$$

oraz

$$c_k = \frac{1}{n+1}\sum_{j=0}^{n} f_j e^{-\frac{2\pi i}{n+1} jk},\ k = 0, \ldots, n. \tag{74}$$

Skoro $Ac = f$, to możemy zapisać też wzory na f_k:

$$f_k = \sum_{j=0}^{n} c_j e^{\frac{2\pi i}{n+1} jk},\ k = 0, \ldots, n. \tag{75}$$

mamy więc dwa zagadnienia

(i) Synteza Fouriera: Dany jest układ $\{c_k\}_k$ i szukamy $\{f_k\}_k$.

(ii) Analiza Fouriera: Dany jest układ $\{f_k\}_k$ i szukamy $\{c_k\}_k$.

Zauważmy, że dla dowolnego $k \in \{0, \ldots, n\}$

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} f_j \bar{a}_{jk} = \frac{1}{n+1} \overline{\sum_{j=0}^{n} \bar{f}_j a_{jk}},$$

zatem oba zagadnienia rozwiązuje się wyliczając sumę postaci

$$X_j = \sum_{k=0}^{N-1} A_k w^{jk}, \ j = 0, \ldots, N-1, \tag{76}$$

gdzie $w = e^{\frac{2\pi i}{N}}$ (w jest pierwiastkiem N-tego stopnia z jedności). Dla wyliczenia X_j potrzeba N mnożeń zespolonych, wobec czego dla wyznaczenia $X = (X_j)_{j=0,\ldots,N-1}$ potrzeba N^2 mnożeń zespolonych. Dla dużych N algorytm ten jest mało efektywny. Będziemy chcieli zmniejszyć ilość mnożeń.

4.7.1 Algorytm szybkiej transformaty Fouriera

Zakładamy, że $N = N_1 \cdot \ldots \cdot N_p$ (np. $N = 2^p$). Sposób postępowania opiera się na grupowaniu wyrażeń sumy (76). Niech $N = N_1 \cdot N_1^\star$, $N_1^\star = N_2 \cdot \ldots \cdot N_p$. Przedstawiamy $0 \le j, k \le N-1$ w postaci

$$j = j_1 \cdot N_1^\star + j_0, \ 0 \le j_0 < N_1^\star, \ 0 \le j_1 < N_1,$$
$$k = k_1 \cdot N_1 + k_0, \ 0 \le k_0 < N_1, \ 0 \le k_1 < N_1^\star.$$

Zatem para (j_0, j_1) jednoznacznie odpowiada liczbie j, a para (k_0, k_1) jednoznacznie liczbie k. Teraz, na mocy (76) mamy

$$X_j = X(j_0, j_1) = \sum_{k_0=0}^{N_1-1} \sum_{k_1=0}^{N_1^\star-1} A(k_0, k_1) w^{j(k_1 N_1 + k_0)}.$$

Policzmy

$$w^{jk_1N_1} = w^{(j_1N_1^\star + j_0)N_1k_1} = w^{jN_1^\star N_1 k_1} w^{j_0 N_1 k_1} = w^{j_0 N_1 k_1},$$

bo $w^N = 1$ oraz $N_1 N_1^\star = N$. Stąd

$$X(j_0, j_1) = \sum_{k_0=0}^{N_1-1} \Big(\sum_{k_1=0}^{N_1^\star-1} A(k_0, k_1) w^{j_0 N_1 k_1} \Big) w^{jk_0} = \sum_{k_0=0}^{N_1-1} A_1(j_0, k_0) w^{jk_0},$$

gdzie

$$A_1(j_0, k_0) = \sum_{k_1=0}^{N_1^\star-1} A(k_0, k_1) w^{j_0 N_1 k_1}.$$

Skoro do wyliczenia A_1 należy wykonać $N_1^\star$ mnożeń, to aby wyznaczyć X_j wykonać należy $N_1 + N_1^\star$ mnożeń. Łącznie, aby obliczyć X wykonujemy

$$N(N_1 + N_1^\star)$$

mnożeń zespolonych.

Stosując podobne rozumowanie, tym razem dla obliczenia A_1, to znaczy biorąc $N_1^\star = N_2 \cdot N_2^\star$ dostaniemy redukcję obliczeń do

$$N_1 \cdot N_1^\star(N_2 + N_2^\star).$$

Wyznaczenie X wymaga wtedy

$$N \cdot N_1 + N(N_2 + N_2^\star) = N(N_1 + N_2 + N_2^\star),$$

itd. Jeżeli $N = N_1 \cdot \ldots \cdot N_p$, to stosując redukcję p-krotnie będziemy musieli wykonać

$$M = N(N_1 + \ldots + N_p)$$

operacji. Na przykład dla $N_i = 2$ mamy

$$M = 2^p(2p) = p2^{p+1}.$$

Przykład 4.26. Niech $N = r^p$. Porównajmy tą metodę z metodą poprzednią, policzmy stosunek ilości mnożeń

$$\frac{M}{N^2} = \frac{pr^{p+1}}{r^{2p}} = \frac{p}{r^{p-1}},$$

zatem dla $r = 2$ i $p = 8$ stosunek ten wynosi $\frac{1}{16}$, widać więc, że korzyść płynąca ze stosowania FFT jest wymierna.

4.8 Funkcje sklejane

Rozpatrzmy teraz interpolację wielomianową, gdy daną mamy dużą ilość węzłów. Wielomian tak otrzymany będzie miał bardzo dużo miejsc zerowych, „będzie się miotał". Jak można temu zaradzić? Jednym ze sposobów jest połączenie węzłów odcinkami. Otrzymamy wówczas funkcję łamaną. Następnie wystarczy wygładzić załamania tak, aby uzyskać rządną regularność.

Na potrzeby tego podrozdziału ustalmy $[a, b]$ – przedział zwarty oraz $\Delta = \{x_0, \ldots, x_n\}$ jego podział taki, że $a = x_0 < x_1 < \ldots < x_n = b$.

Definicja 4.27. *Funkcję $S : \mathbb{R} \to \mathbb{R}$ nazywamy funkcją sklejaną stopnia m $\equiv$*

$x_{-1} = -\infty$, $x_{n+1} = \infty$

(i) $S_{|(x_{i-1}, x_i)} \in \Pi_m$, $i = 0, \ldots, n+1$,

(ii) $S \in \mathcal{C}^{m-1}(\mathbb{R}, \mathbb{R})$.

Definicja 4.28. *Funkcję sklejaną stopnia $2m-1$ nazywamy naturalną funkcją sklejaną $\equiv$*

(iii) $S_{|(-\infty, x_0)}$ *oraz* $S_{|(x_n, \infty)}$ *są wielomianami stopnia co najwyżej* $m-1$.

Oznaczenia 4.29.

$\mathcal{S}_m(\Delta) =$ *zbiór funkcji sklejanych stopnia* m,
$\mathcal{N}_{2m-1}(\Delta) =$ *zbiór naturalnych funkcji sklejanych*,
$x_+ = \begin{cases} x, & x \geq 0 \\ 0, & x < 0 \end{cases}$

Przykład 4.30.

$S \in \mathcal{S}_1(\Delta) \Rightarrow S$ jest funkcją łamaną.

Twierdzenie 4.31. *(o postaci funkcji łamanych)*

(i) Jeżeli $S \in \mathcal{S}_m(\Delta)$, *to*

$$S(x) = p(x) + \sum_{j=0}^{n} a_j (x - x_j)_+^m,$$

gdzie $p \in \Pi_m$.

(ii) Jeżeli $S \in \mathcal{N}_{2m-1}(\Delta)$, *to*

$$S(x) = p(x) + \sum_{j=0}^{n} a_j (x - x_j)_+^{2m-1}, \ p \in \Pi_{m-1},$$

oraz

$$(77) \quad \sum_{j=0}^{n} a_j x_j^r = 0, \ r = 0, \ldots, m-1.$$

Dowód. Oznaczmy przez $Y(x)$ funkcję Heaviside'a (funkcję skoku jednostkowego):

$$Y(x) = \begin{cases} 1, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Wówczas

$$x_+ = \int_{-\infty}^{x} Y(s)ds,$$

oraz

$$(78) \quad \frac{(x_+)^m}{m!} = (m) \int_{-\infty}^{x} Y(s)ds - m - \text{krotna całka}.$$

ad *(i)* (Trochę inny niż na wykładzie.) Skoro $S \in \mathcal{S}_m(\Delta)$, to S jest na każdym przedziale (x_{i-1}, x_i) (dla $i = 0, \ldots, n+1$) wielomianem stopnia co najwyżej m, zatem $S^{(m)}(x) = b_i$ dla $x \in (x_{i-1}, x_i)$, $i = 0, \ldots, n+1$. Weźmy wielomian $p \in \Pi_m$ taki, że

$$p_{|(-\infty, x_0)} = S_{|(-\infty, x_0)}.$$

Wielomian taki istnieje, wystarczy przedłużyć wielomian który składa się na naszą funkcję S na przedziale $(-\infty, x_0)$. Wówczas

$$S^{(j)}(x) - p^{(j)}(x) = 0, \ x \in (-\infty, x_0), j \geq 0.$$

Niech

$$T(x) = S(x) - p(x), \tag{79}$$

wówczas

$$T^{(j)}(x) = 0, \ x \in (-\infty, x_0), j = 1, \ldots, m-1.$$

Zatem

$$\int_{-\infty}^{x} T^{(m)}(s)ds = T^{(m-1)}(x),$$

$$\int_{-\infty}^{x} T^{(m-1)}(s)ds = T^{(m-2)}(x),$$

$$\ldots$$

$$\int_{-\infty}^{x} T'(s)ds = T(x),$$

czyli

$$T(x) = (m) \int_{-\infty}^{x} T^{(m)}(s)ds. \tag{80}$$

Ale

$$\begin{aligned} S^{(m)}(x) = b_0 &+ (b_1 - b_0)Y(x - x_0) + (b_2 - b_1)Y(x - x_1) + \ldots \\ &+ (b_n - b_{n-1})Y(x - x_{n-1}) + (b_{n+1} - b_n)Y(x - x_n), \end{aligned}$$

więc

$$T^{(m)}(x) = \sum_{j=0}^{n} (b_{j+1} - b_j)Y(x - x_j), \tag{81}$$

zatem

$$S(x) \overset{(79)}{=} p(x) + T(x) \overset{(80)}{=} p(x) + (m) \int_{-\infty}^{x} T^{(m)}(s)ds$$

$$\overset{(81)}{=} p(x) + \sum_{j=0}^{n} (b_{j+1} - b_j) \int_{-\infty}^{x} Y(s - x_j)ds$$

$$\overset{(78)}{=} p(x) + \sum_{j=0}^{n} \frac{b_{j+1}-b_j}{m!}(x - x_j)_+^m.$$

ad*(ii)* Skoro $S \in \mathcal{N}_{2m-1}(\Delta)$, to z poprzedniego punktu

$$S(x) = p(x) + \sum_{j=0}^{n} a_j(x - x_j)_+^{2m-1}, \ p \in \Pi_{2m-1}.$$

Wystarczy więc wykazać, że $p \in \Pi_{m-1}$ oraz zachodzi (77).

Skoro $S \in \mathcal{N}_{2m-1}(\Delta)$, to dla $x < x_0$ mamy

$$S(x) = p(x) \in \Pi_{m-1}.$$

Dla $x \geq x_n$

$$S(x) = p(x) + (2m-1)! \sum_{j=0}^{n} \frac{a_j}{(2m-1)!}(x - x_j)^{2m-1} \in \Pi_{m-1},$$

czyli $S^{(m)}(x) = 0$ dla $x > x_n$, więc

$$0 = S^{(m)}(x) = (2m-1)! \sum_{j=0}^{n} \frac{a_j}{(m-1)!}(x - x_j)^{m-1} = W(x), \ x > x_n.$$

Więc W jako wielomian (stopnia co najwyżej $m-1$) zerujący się na odcinku jest wielomianem zerowym, więc $W^{(r)}(0) = 0$ dla $r = 0, \ldots, m-1$. Ale

$$W^{(r)}(x) = (2m-1)! \sum_{j=0}^{n} a_j \frac{(x-x_j)^{m-1-r}}{(m-1-r)!}, r = 0, \ldots, m-1.$$

Dla $r = 0, \ldots, m-1$, przyjmując $k = m-1-r$ dostajemy

$$0 = W^{(r)}(0) = (2m-1)! \sum_{j=0}^{n} a_j \frac{(-1)^k x_j^k}{k!} = \frac{(2m-1)!(-1)^k}{k!} \sum_{j=0}^{n} a_j x_j^k.$$

Wykazaliśmy więc (77), bo jeśli $r \in \{0, \ldots, m-1\}$, to $k \in [0..m-1]$.

□

Lemat 4.32. *Dla dowolnego przedziału* $[\alpha,\beta]\subset\mathbb{R}$ *oraz podziału* $\Delta=\{x_0,\ldots,x_n\}\subset[\alpha,\beta]$ $(x_i\neq x_j)$*, jeżeli* $g\in\mathcal{C}^{m-1}([\alpha,\beta],\mathbb{R})$ *ma ciągłą pochodną rzędu* m *dla* $x\in(x_{i-1},x_i)$*,* $i=1,\ldots,n$*, to*

$$S\in\mathcal{N}_{2m-1}(\Delta)\Rightarrow\int_\alpha^\beta g^{(m)}(x)S^{(m)}(x)dx=(-1)^m(2m-1)!\sum_{j=0}^{n}a_jg(x_j),$$

gdzie a_j *jak w poprzednim twierdzeniu.*

Dowód. Niech $S\in\mathcal{N}_{2m-1}(\Delta)$, $x_0<\ldots<x_n$. Wówczas

$$S^{(m+j)}_{|(-\infty,x_0)}=S^{(m+j)}_{|(x_n,\infty)}=0,\ j\geq 0.$$

Stąd, całkując przez części

$$\begin{aligned}I&=\int_\alpha^\beta g^{(m)}(x)S^{(m)}(x)dx=S^{(m)}(x)g^{(m-1)}(x)\Big|_\alpha^\beta-\int_\alpha^\beta g^{(m-1)}(x)S^{(m+1)}(x)dx\\&=(-1)^{m-1}\int_\alpha^\beta g'(x)S^{(2m-1)}(x)dx=(-1)^{m-1}\sum_{i=0}^{n-1}\int_{x_i}^{x_{i+1}}S^{(2m-1)}(x)g'(x)dx.\end{aligned}$$

Na mocy twierdzenia o postaci naturalnej funkcji sklejanej wiemy, że

$$S(x)=p(x)+\sum_{j=0}^{n}a_j(x-x_j)_+^{2m-1},$$

zatem dla $x\in(x_i,x_{i+1})$ $(i=0,\ldots,n-1)$

$$S(x)=p(x)+\sum_{j=0}^{i}a_j(x-x_j)_+^{2m-1},$$

$$S^{(2m-1)}(x)=(2m-1)!\sum_{j=0}^{i}a_j=:\nu_i.$$

Oczywiście

$$\begin{aligned}&\Delta\nu_i=\nu_{i+1}-\nu_i=(2m-1)!a_{i+1},\ i=0,\ldots,n-1,\\&\nu_0=(2m-1)!a_0,\\&\nu_n=0\text{ bo }S\in\mathcal{N}_{2m-1}(\Delta).\end{aligned}$$

Wróćmy do naszej całki.

$$I = (-1)^{m-1} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} S^{(2m-1)}(x) g'(x) dx = (-1)^{m-1} \sum_{i=0}^{n-1} \int_{x_i}^{x_{i+1}} \nu_i g'(x) dx$$

$$= (-1)^{m-1} \sum_{i=0}^{n-1} \nu_i [g(x_{i+1}) - g(x_i)] = (-1)^{m-1} \sum_{i=0}^{n-1} \nu_i \Delta g(x_i)$$

$$\overset{4.19\ (iv)}{=} (-1)^{m-1} [\nu_n g(x_n) - \nu_0 g(x_0) - \sum_{i=0}^{n-1} g(x_{i+1}) \Delta \nu_i]$$

$$= (-1)^m [(2m-1)! a_0 g(x_0) + (2m-1)! \sum_{i=0}^{n-1} g(x_{i+1}) a_{i+1}]$$

$$= (-1)^m (2m-1)! [a_0 g(x_0) + \sum_{i=1}^{n} g(x_i) a_i] = (-1)^m (2m-1)! \sum_{i=0}^{n} g(x_i) a_i$$

□

Twierdzenie 4.33. *(istnienie i jednoznaczność interpolacji naturalnymi funkcjami sklejonymi)*
Jeżeli $\Delta = \{x_0, \ldots, x_n\}$ *jest podziałem przedziału* $[a, b]$ *takim, że* $x_i \neq x_j$ *dla* $i \neq j$, $f : [a, b] \to \mathbb{R}$ *oraz* $1 \leq m \leq n + 1$, *to dla dowolnego zespołu wartości* $\{y_i\}_{i=0,\ldots,n}$ *istnieje dokładnie jedna naturalna funkcja sklejana* $S \in \mathcal{N}_{2m-1}(\Delta)$ *taka, że*

(82) $S(x_i) = y_i,\ i = 0, \ldots, n.$

Dowód. Na mocy twierdzenia o postaci naturalnej funkcji sklejanej wiemy, że jeśli S istnieje, to jest postaci

$$S(x) = p(x) + \sum_{j=0}^{n} a_j (x - x_j)_+^{2m-1},\ p \in \Pi_{m-1},$$

oraz zachodzi (77). Do wyznaczenia mamy m współczynników wielomianu p oraz $n + 1$ stałych a_j. Warunki (82) ($n + 1$ warunków) oraz (77) (m warunków) są liniowe ze względu na szukane przez nas niewiadome (a_j oraz współczynniki wielomianu p), zatem wystarczy wykazać, że jeżeli

$$S(x_i) = 0,\ i = 0, \ldots, n,$$

to $S \equiv 0$. Wykorzystamy w tym celu poprzedni lemat. Przyjmując

$$y_i = 0,\ i = 0, \ldots, n,$$
$$g(x) = S(x),$$

otrzymujemy

$$\int_a^b (S^{(m)}(x))^2 dx = 0 \Rightarrow S^{(m)}(x) \equiv 0 \text{ w } [a, b] \Rightarrow S \in \Pi_{m-1},$$

więc

$$\deg S \leq m-1 \leq n < n+1.$$

Ale S ma $n+1$ miejsc zerowych $(x_0, \ldots, x_n)$, zatem na mocy zasadniczego twierdzenia algebry $S \equiv 0$. □

Przykład 4.34. Wyznaczymy $S \in \mathcal{N}_3(\Delta),\ \Delta = \{x_0, \ldots, x_N\},\ x_0 < \ldots < x_N$. Niech

$$S_{|[x_i, x_{i+1}]} = S_i,\ i = 0, \ldots, N-1.$$

Wiemy, że

$$S_i(x) = a_i + b_i t + c_i t^2 + d_i t^3,\ t = x - x_i,\ i = 0, \ldots, N-1,$$

Należy więc wyznaczyć $4N$ stałych $a_i, b_i, c_i, d_i,\ i = 0, \ldots, N-1$. Sprawdźmy jakie warunki mamy.

1° $N+1$ warunków interpolacyjnych:

$$S_i(x_i) = y_i,\ i = 0, \ldots, N-1,$$
$$S_{N-1}(x_N) = y_N.$$

2° $N-1$ warunków ciągłości S:

$$S_i(x_{i+1}) = S_{i+1}(x_{i+1}),\ i = 0, \ldots .N-2.$$

3° $N-1$ warunków ciągłość S':

$$S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}),\ i = 0, \ldots, N-2.$$

4° $N-1$ warunków ciągłość S'':

$$S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}),\ i = 0, \ldots, N-2.$$

5° 2 warunki naturalności:

$$S_0''(x_0) = 0,$$
$$S_{N-1}''(x_N) = 0.$$

Do rozwiązania mamy więc układ $4N$ równań z $4N$ niewiadomymi. Oczywiście, dla $i = 0, \ldots, N-1$

$$S_i(x_i) = a_i,$$
$$S_i'(x) = b_i + 2c_i t + 3d_i t^2,\ S_i'(x_i) = b_i,$$
$$S_i''(x) = 2c_i + 6d_i t,\ S''(x_i) = 2c_i.$$

Niech $h_i = x_{i+1} - x_i$, wówczas układ jaki należy rozwiązać to

$$\begin{cases} a_i = y_i, & i = 0, \dots, N-1 \\ a_{N-1} + b_{N-1}h_{N-1} + c_{N-1}h_{N-1}^2 + d_{N-1}h_{N-1}^2 = y_N, & \\ a_i + b_i h_i + c_i h_i^2 + d_i h_i^2 = a_{i+1}, & i = 0, \dots, N-2 \\ b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1}, & i = 0, \dots, N-2 \\ 2c_i + 6d_i h_i = 2c_{i+1}, & i = 0, \dots, N-2 \\ c_0 = 0, & \\ 2c_{N-1} + 6d_{N-1}h_{N-1} = 0. & \end{cases}$$

Od razu otrzymujemy

$$a_i = y_i,\ i = 0, \dots, N-1,$$
$$c_0 = 0$$

do rozwiązania pozostały więc równania

(i) $y_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_{i+1},\ i = 0, \dots, N-1,$

(ii) $b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1},\ i = 0, \dots, N-2,$

(iii) $c_i + 3d_i h_i = c_{i+1},\ i = 0, \dots, N-1,$ (przyjmujemy $c_N = 0$).

Mamy więc

$$\text{(iii)} \Rightarrow d_i = \tfrac{c_{i+1}-c_i}{3h_i},\ i = 0, \dots, N-1,$$
$$\text{(i)} \Rightarrow b_i = \tfrac{y_{i+1}-y_i}{h_i} - c_i h_i - d_i h_i^2 = \tfrac{y_{i+1}-y_i}{h_i} - \tfrac{h_i(2c_i+c_{i+1})}{3},\ i = 0, \dots, N-1.$$

Tak wyliczone d_i i b_i wstawiamy do (ii) i otrzymujemy

$$\text{(83)}\quad \tfrac{c_i}{3}\tfrac{h_i}{h_i+h_{i+1}} + \tfrac{2c_{i+1}}{3} + \tfrac{c_{i+2}}{3}\tfrac{h_{i+1}}{h_i+h_{i+1}} = \tfrac{1}{h_i+h_{i+1}}\big(\tfrac{y_{i+2}-y_{i+1}}{h_{i+1}} - \tfrac{y_{i+1}-y_i}{h_i}\big).$$

Niech

$$c_i^\star = \tfrac{c_i}{3},$$
$$u_i = \tfrac{h_{i-1}}{h_{i-1}+h_i},$$
$$w_i = \tfrac{h_i}{h_{i-1}+h_i} = 1 - u_i,$$
$$v_i = \tfrac{1}{h_{i-1}+h_i}\big(\tfrac{y_{i+1}-y_i}{h_i} - \tfrac{y_i-y_{i-1}}{h_{i-1}}\big),$$

wówczas (83) przyjmuje postać

$$\text{(84)}\quad u_{i+1}c_i^\star + 2c_{i+1}^\star + w_{i+1}c_{i+2}^\star = v_{i+1},\ i = 0, \dots, N-2.$$

Okazuje się, że równania te można zapisać w postaci macierzowej

$$Ac = v,$$

gdzie A jest macierzą trójprzekątniową, z 2 na przekątnej, $w_1, \dots, w_{N-2}$ nad przekątną oraz $u_2, \dots, u_{N-1}$ pod przekątną, $c = (c_1^\star, \dots, c_{N-1}^\star)^T$, $v = (v_1, \dots, v_{N-1})^T$. Macierz ta jest nieredukowalna, ale spełnia słabe kryterium sumy wierszy, zatem istnieje rozwiązanie zagadnienia interpolacji naturalnymi funkcjami sklejanymi stopnia 3.

Twierdzenie 4.35. *(oszacowanie błędu aproksymacji naturalnej funkcji sklejanej stopnia 3)*
Jeżeli $f \in \mathcal{C}^2([x_0, x_N], \mathbb{R})$ *oraz* $\max\{|f''(x)| : x \in [x_0, x_N]\} = M$, *to*

$$\max\{|f(x) - S(x)| : x \in [x_0, x_N]\} \leq 5M \max\{h_i^2 : i = 0, \ldots, N-1\}.$$

Dowód. Dla $x \in [x_i, x_{i+1}]$ niech

$$S(x) = S_i(x) = a_i + b_i t + c_i t^2 + d_i t^3, t = x - x_i.$$

Wówczas

$$\begin{aligned}|S_i(x) - f(x)| &= \left|f(x_i) - f(x) + \left[\tfrac{f(x_{i+1})-f(x_i)}{h_i} - h_i(c^\star_{i+1} + 2c^\star_i)\right]t + \ 3c^\star_i t^2 + \tfrac{c^\star_{i+1}-c^\star_i}{h_i}t^3\right| \\ &\leq \left|f(x_i) - f(x) + \tfrac{f(x_{i+1})-f(x_i)}{h_i}t\right| + \ \left| - h_i(c^\star_{i+1} + 2c^\star_i)t + 3c^\star_i t^2 + \tfrac{c^\star_{i+1}-c^\star_i}{h_i}t^3\right|.\end{aligned}$$

Oznaczmy przez A i B kolejne składniki ostatniej sumy. Ze wzoru Taylora:

$$\begin{aligned}&f(x) = f(x_i) + f'(x_i)t + \tfrac{1}{2}f''(\eta_i)t^2, \eta_i \in I(x_i, x), \\ &f(x_{i+1}) = f(x_i) + f'(x_i)h_i + \tfrac{1}{2}f''(\eta_i)h_i^2, \\ &A = \left|f(x_i) - f(x_i) - f'(x_i)t - \tfrac{1}{2}f''(\eta_i)t^2 + \tfrac{f'(x_i)h_i + \frac{1}{2}f''(\eta_i)h_i^2}{h_i}t\right| \leq Mh_i^2, \\ &B \leq h_i^2(|c^\star_{i+1}| + 2|c^\star_i| + 3|c^\star_i| + |c^\star_{i+1}| + |c^\star_i|) = h_i^2(2|c^\star_{i+1}| + 6|c^\star_i|) \\ &\quad \leq 8c^\star h_i^2,\end{aligned}$$

gdzie

$$c^\star = \max\{|c^\star_i| : i = 0, \ldots, N-1\}.$$

Zgodnie z poprzednim przykładem

$$\begin{aligned}&u_{i+1}c^\star_i + 2c^\star_{i+1} + w_{i+1}c^\star_{i+1} = v_{i+1}, \\ &u_{i+1} + w_{i+1} = 1.\end{aligned}$$

Zatem

$$|v_{i+1}| = |u_{i+1}c^\star_i + 2c^\star_{i+1} + w_{i+1}c^\star_{i+1}| \geq 2|c^\star_{i+1}| - (u_{i+1}|c^\star_i| + w_{i+1}|c^\star_{i+2}|).$$

Niech $i = \operatorname{argmax}\{|c^\star_i| : i = 0, \ldots, N-1\}$, wówczas

$$|v_{i+1}| \geq 2|c^\star| - |c^\star|(u_{r+1} + w_{r+1})) = c^\star.$$

Z drugiej strony:

$$\begin{aligned}v_i &= \tfrac{1}{h_i + h_{i+1}}\left(\tfrac{y_{i+1}-y_i}{h_i} - \tfrac{y_i - y_{i-1}}{h_{i-1}}\right) = \tfrac{1}{x_{i+1}-x_{i-1}}\left(\tfrac{f(x_{i+1})-f(x_i)}{x_{i+1}-x_i} - \tfrac{f(x_i)-f(x_{i-1})}{x_i - x_{i-1}}\right) \\ &= \tfrac{f[x_i, x_{i+1}] - f[x_{i-1}, x_i]}{x_{i+1}-x_{i-1}} = f[x_{i-1}, x_i, x_{i+1}] = \tfrac{1}{2}f''(\zeta_i).\end{aligned}$$

Stąd

$$|c^\star| \leq \tfrac{1}{2}M,$$

zatem

$$\begin{aligned}&|S_i(x) - f(x)| \leq 5Mh_i^2, \\ &|S(x) - f(x)| \leq 5M \cdot \max\{h_i^2 : i = 0, \ldots, N-1\}.\end{aligned}$$

□

5 Aproksymacja.

W rozdziale tym zajmiemy się następującym problemem. Dana niech będzie przestrzeń unormowana X oraz $F \subseteq X$ – jej podprzestrzeń. Dla $u \in X \setminus F$ szukać będziemy takiego $v \in F$, że $\|u - v\| = \min\{\|u - y\| : y \in F\}$. Jeżeli takie v istnieje, to nazywamy je **elementem optymalnym**. Wówczas

(85) $\mathcal{E}_F = \inf\{\|u - y\| : y \in F\}$

nazywać będziemy **błędem aproksymacji**.

Definicja 5.1. *Przestrzeń unormowaną X nazywamy unitarną, jeśli norma zadana jest przez iloczyn skalarny $\|x\| = \sqrt{(x|x)}$. Przestrzenią Banacha nazywamy zupełną przestrzeń unormowaną, przestrzenią Hilberta – zupełną przestrzeń unitarną.*

Przykład 5.2.

1. Przestrzenie unormowane:
   (a) $\mathcal{C}([a, b], \mathbb{R}^n)$, $\|x\| = \max\{|x(t)| : t \in [a, b]\}$, gdzie $|\cdot|$ jest normą w $\mathbb{R}^n$,
   (b) $\tilde{\mathcal{C}}(\mathbb{R}, \mathbb{R}) \subseteq \mathcal{C}([a, b], \mathbb{R})$ – przestrzeń funkcji T-okresowych, orzy czym $T = b - a$,
   (c) $F = \{t_n : t_n(x) = \sum_{j=0}^{n} a_j \cos jx + b_j \sin jx\} \subset \tilde{\mathcal{C}}(\mathbb{R}, \mathbb{R})$

   $$F = \operatorname{span}\{1, \cos x, \sin x, \ldots, \cos nx, \sin nx\},$$
   $$\dim F = 2n + 1.$$

   (d) $\Pi_n \subseteq \mathcal{C}([a, b], \mathbb{R})$,
2. Przestrzenie unitarne:
   (a) $L^2_p[a, b] = \{f : f$ mierzalna na $[a, b]$ i $\int_a^b p(x) f^2(x)\,dx < \infty\}$, przy czym $p > 0$ prawie wszędzie w $[a, b]$. Jest to przestrzeń funkcji całkowalnych z kwadratem z wagą p. Skoro, dla $u, v \in L^2_p$

   $$2|u(x)v(x)| \le |u|^2 + |v|^2,$$

   to

   $$(u|v) = \int_a^b p(x)u(x)v(x)\,dx$$

   jest określona, bo

   $$(u|v) \le \int_a^b \frac{p(x)}{2}(|u|^2 + |v|^2)\,dx < \infty.$$

   Więc $(u|v)$ jest iloczynem skalarnym.
   Aproksymacja w L^2_p nazywana jest **aproksymacją średniokwadratową**.

(b) $\mathcal{C}([a,b],\mathbb{R}) \subseteq L^2_p(a,b) \Rightarrow F, \Pi_n$ są przestrzeniami unitarnymi.

Twierdzenie 5.3. *(istnienie elementu optymalnego)*
Jeżeli $F \subseteq X$ *jest podprzestrzenią unormowaną przestrzeni wektorowej oraz* $\dim F < +\infty$*, to dla wszystkich* $u \in X$ *istnieje* $v \in F$ *– element optymalny dla* u*.*

Dowód.

Lemat 5.4. *(oszacowanie a priori elementu optymalnego)*
Jeżeli v *jest elementem optymalnym, to* $\|v\| \leq 2\|u\|$*.*

Dowód lematu. Niech $h \in F$ będzie elementem optymalnym u takim, że $\|h\| > 2\|u\|$. Wtedy

$$\|u-h\| = \|h-u\| \geq \|h\| - \|u\| > 2\|u\| - \|u\| = \|u\| = \|u-0\|,$$

zatem 0 lepiej przybliża u niż element h. □

Szukamy $\min\{\|u-y\| : y \in F\}$. Dzięki lematowi który przed chwilą wykazaliśmy mamy, że

$$\inf\{\|u-y\| : y \in F\} = \min\{\|u-y\| : y \in F \cap \overline{K}(0, 2\|u\|)\}.$$

Skoro $\dim F < +\infty$, to $F \cap \overline{K}(0, 2\|u\|)$ jest zbiorem zwartym (zatem kres dolny mogliśmy zastąpić przez minimum). Ponadto $\|u-y\|$ jest funkcją ciągłą zmiennej y, stąd na mocy twierdzenia Weierstrassa

$$\exists v \in F : \|u-v\| = \min\{\|u-y\| : y \in F \cap \overline{K}(0, 2\|u\|)\}.$$

□

Przykład 5.5. Niech $X = \mathbb{R}^2$, $x = \binom{x_1}{x_2} \in X$, $\|x\|_\infty = \max\{|x_1|, |x_2|\}$, $u = (0,1)$, $F = \mathbb{R} \times \{0\}$. Wówczas zbiorem rozwiązań optymalnych jest $[-1,1] \times \{0\}$, zatem mamy brak jednoznaczności elementu optymalnego. Jeżeli rozpatrywalibyśmy normę euklidesową, to istniałby dokładnie jeden element optymalny. Dzieje się tak dlatego, że kula w normie $\|\cdot\|_\infty$ nie jest silnie wypukła (w normie euklidesowej tak jest).

Twierdzenie 5.6. *(warunek konieczny i wystarczający optymalności)*
Jeżeli $F \subseteq X$ *jest skończenie wymiarową podprzestrzenią przestrzeni unitarnej* X*, to warunkiem koniecznym i wystarczającym na to, by* $v \in F$ *był elementem optymalnym dla* $u \in X$ *jest, aby* $\forall y \in F : (u-v|y) = 0$*.*

Dowód. Zauważmy, że dla $\|u\| = \sqrt{(u|u)}$, $\overline{K}(0,r) = \{u : \|u\| \leq r\}$ jest zbiorem silnie wypukłym. Dzięki poprzedniemu twierdzeniu mamy istnienie elementu optymalnego $v \in F$ dla $u \in X$.

Przypuśćmy, że istnieje $h \in F$ takie, że

$$(u-v|h) = \alpha > 0.$$

Wówczas

$$\begin{aligned}\|u-(v+\beta h)\|^2 &= (u-(v+\beta h)|u-(v+\beta h)) = ((u-v)-\beta h|(u-v)-\beta h)\\ &= \|u-v\|^2 - 2\beta(u-v|h) + \beta^2\|h\|^2\\ &= \|u-v\|^2 - 2\beta\alpha + \beta^2\|h\|^2\end{aligned}$$

Niech $\beta = \frac{\alpha}{\|h\|^2}$. Wówczas

$$\|u - (v + \beta h)\|^2 = \|u - v\|^2 - \beta\alpha < \|u - v\|^2$$

i v nie jest optymalne – sprzeczność.

Dla dowodu drugiej implikacji ($\Leftarrow$) weźmy dowolny $y \in F$. Wówczas

$$\begin{aligned}\|u - y\|^2 &= \|u - v - (y - v)\|^2 = (u - v - (y - v)|u - v - (y - v)) \\ &= \|u - v\|^2 - 2(u - v|y - v) + \|y - v\|^2 = \|u - v\|^2 + \|y - v\|^2,\end{aligned}$$

bo $y - v \in F$. Wobec tego, dla dowolnego $y \in F$

$$\|u - y\|^2 \geq \|u - v\|^2.$$

□

Wniosek 5.7. *Przy założeniach poprzedniego twierdzenia element optymalny jest wyznaczony jednoznacznie.*

Dowód. Dla dowodu nie wprost załóżmy, że $v_1 \neq v_2$ są elementami optymalnymi dla u. Postępując analogicznie jak w dowodzie powyższego twierdzenia dostaniemy np.:

$$\forall y \in F : \|u - y\|^2 \geq \|u - v_2\|^2.$$

Skoro $v_1 \in F$ i $v_1 \neq v_2$, to

$$\|u - v_1\|^2 > \|u - v_2\|^2,$$

czyli v_1 nie byłby optymalny. □

Twierdzenie 5.8. *(wyznaczanie elementu optymalnego)*
Przy założeniach twierdzenia 5.6, jeżeli $F = \operatorname{span}\{f_1, \ldots, f_n\}$ *i* $v \in F$ *jest elementem optymalnym dla* $u \in X$*, to* $\{\alpha_j\}_{j=1,\ldots,n}$ *takie, że*

$$v = \sum_{j=1}^{n} \alpha_j f_j$$

wyznacza się z układu równań

$$\sum_{j=1}^{n} \alpha_j (f_i|f_j) = (f_i|u), \ i = 1, \ldots, n. \tag{86}$$

Dowód. Na mocy twierdzenia 5.6 mamy

$$\forall y \in F : (u - v|y) = 0.$$

Równoważne to jest

$$(u - v|f_i) = 0, \ i = 1, \ldots, n,$$

czyli

$$\Big(u - \sum_{j=1}^{n} \alpha_j f_j | f_i\Big) = 0, \ i = 1, \ldots, n,$$

co kończy dowód. □

Uwaga 5.9. *Układ (86) ma jednoznaczne rozwiązanie, jeżeli macierz Gramma*

$$G = (g_{ij})_{i,j},\ g_{ij} = (f_i|f_j)$$

jest nieosobliwa.

Uwaga 5.10. *Macierz Gramma G jest nieosobliwa wtedy i tylko wtedy, gdy* $f_1, \ldots, f_n$ *są liniowo niezależne.*

$G^\star = G$, *gdy* X *nad* $\mathbb{C}$.

Dowód. Oczywiście $G^T = G$. Dla $z = \sum_{j=1}^{n} \beta_j f_j$ mamy:

$$(z|z) = \Big(\sum_{i=1}^{n} \beta_i f_i | \sum_{j=1}^{n} \beta_j f_j\Big) = \sum_{i,j=1}^{n} \beta_i \beta_j (f_i|f_j) = \beta^T G \beta,$$

gdzie $\beta = (\beta_1, \ldots, \beta_n)^T$. Zatem

$$(z|z) \geq 0 \Leftrightarrow \forall \beta : \beta^T G \beta \geq 0 \Leftrightarrow G \geq 0.$$

Po tej uwadze możemy przejść do właściwego dowodu:

$$f_1, \ldots, f_n \text{ są liniowo niezależne} \Leftrightarrow \forall \alpha = (\alpha_1, \ldots, \alpha_n)^T,\ \alpha \neq 0 : \sum_{j=1}^{n} \alpha_j f_j \neq 0$$

$$\Leftrightarrow \forall \alpha \neq 0 : \Big(\sum_{j=1}^{n} \alpha_j f_j | \sum_{j=1}^{n} \alpha_j f_j\Big) > 0 \Leftrightarrow \alpha^T G \alpha > 0 \Leftrightarrow G \text{ jest nieosobliwa.}$$

□

Uwaga 5.11. *Problem stwarza wybór bazy przestrzeni* F. *Sytuacją idealną jest, gdy* $\{f_i\}_i$ *tworzy ciąg ortogonalny, tzn.:*

$$(f_i|f_j) = 0,\ i \neq j.$$

Wówczas

$$(u|f_i) = \alpha_i (f_i|f_i),$$
$$G = \operatorname{diag}((f_1|f_1), \ldots, (f_n|f_n)) = \operatorname{diag}(\|f_1\|, \ldots, \|f_n\}).$$

Sytuacją „pomyślną" jest taka, gdy G *ma liczne zera. Metoda wyznaczenia takich elementów bazowych nazywa się* ***metodą elementu skończonego****.*

Przyjrzyjmy się jeszcze przestrzeni

$$X = \tilde{\mathcal{C}} = \{c \in \mathcal{C}(\mathbb{R}, \mathbb{R}) : u(x) = u(x + 2\pi)\},$$

oraz jej podprzestrzeni

$$F_n = \{t_n : t_n(x) = a_0 + \sum_{j=1}^{n} (a_j \cos jx + b_j \sin jx)\}$$

– przestrzeni wielomianów trygonometrycznych. Jej bazą ortogonalną jest $\{1, \cos x, \sin x, \ldots, \cos nx, \sin nx\}$. Dla $u \in \tilde{\mathcal{C}}$, S_n - element optymalny dla u ma współczynniki zadane za pomocą **wzorów Fouriera**

$$\|u - S_n\| \leq \|u - y\| \ \forall y \in F_n.$$

Znalezienie siatki na prostej czy na płaszczyźnie raczej nie sprawia problemu, ale w wyższych wymiarach już tak.

Uwaga 5.12. *Interpolację wielomianową można traktować jako problem aproksymacji przestrzeni* $\mathcal{C}([a, b], \mathbb{R})$ *podprzestrzenią* Π_n.

Dowód. Dla $\Delta = \{x_0, \ldots, x_n\} \subset [a, b]$ oraz iloczynu skalarnego

$$(u|v) = \sum_{k=0}^{n} u(x_k)v(x_k),$$

lub (w przypadku zespolonym)

$$(u|v) = \sum_{k=0}^{n} u(x_k)\overline{v(x_k)},$$

mamy

$$(u|u) = 0 \Leftrightarrow \sum_{k=0}^{n} u^2(x_k) = 0 \Leftrightarrow u(x_k) = 0, \ k = 0, \ldots, n,$$

zatem (na mocy zasadniczego twierdzenia algebry) wtedy i tylko wtedy, gdy $u \equiv 0$.

Jeśli $\Pi_n = \operatorname{span}\{\varphi_0, \ldots, \varphi_n\}$, to warunek optymalności $((u - v|y) = 0 \ \forall y \in \Pi_n)$ daje nam

$$(u|\varphi_i) = (v|\varphi_i), \ i = 0, \ldots, n,$$
$$\sum_{k=0}^{n} u(x_k)\varphi_i(x_k) = \sum_{k=0}^{n} v(x_k)\varphi_i(x_k), \ i = 0, \ldots, n.$$

Zatem, dla $v = \sum_{j=0}^{n} \alpha_j \varphi_j$ mamy

$$\sum_{k=0}^{n} u(x_k)\varphi_i(x_k) = \sum_{k=0}^{n} \Big(\sum_{j=0}^{n} \alpha_j \varphi_j(x_k) \Big) \varphi_i(x_k), \ i = 0, \ldots, n.$$

Dla

$$A = (a_{ij})_{ij=0,\ldots,n}, \ a_{ij} = \varphi_j(x_i), \ \alpha = \begin{pmatrix} \alpha_0 \\ \vdots \\ \alpha_n \end{pmatrix}, b = \begin{pmatrix} u(x_0) \\ \vdots \\ u(x_n) \end{pmatrix} = \begin{pmatrix} b_0 \\ \vdots \\ b_n \end{pmatrix}$$

mamy

$$\sum_{k=0}^{n} \sum_{j=0}^{n} \alpha_j a_{kj} a_{ki} = \sum_{k=0}^{n} b_k a_{ki}, \det A \neq 0,$$

bo $\varphi_0, \ldots, \varphi_n$ jest bazą, zatem

$$A^* A\alpha = A^* b \Rightarrow b = A\alpha \text{ (warunki interpolacji!).}$$

□

5.1 Ortogonalizacja

Przypomnijmy sobie (i ponownie zapiszmy) twierdzenie 2.8 o ortogonalizacji.

Twierdzenie 5.13. *(O ortogonalizacji.)*
Dana niech będzie przestrzeń unitarna $(X, (\cdot|\cdot))$ *oraz ciąg* $\{f_n\}_n \subset X$ *wektorów liniowo niezależnych* [14]. *Wówczas istnieje ciąg* $\{\varphi_n\}_n \subset X$ *taki, że*

(i) $(\varphi_i|\varphi_j) = 0$ *dla* $i \neq j$;

(ii) $\forall N \in \mathbb{N} : \operatorname{span}\{f_1, \ldots, f_N\} = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$.

Uwaga 5.14. *Warunki (i) i (ii) implikują, że* $\{\varphi_n\}_n$ *też jest liniowo niezależny.*

Uwaga 5.15. *Ortogonalizacja zależy od przestrzeni* X *oraz wyboru iloczynu skalarnego.*

Przykład 5.16.

1. $\{e^{inx}\}_{n=0,1\ldots}$ jest ortogonalny w $[-\pi, \pi]$ z wagą $p = 1$, gdy

$$X = L^2_p([a, b]),\ (u|v) = \int_a^b p(x)u(x)v(x)dx,\ u, v \in X,$$

przy czym $p(x) > 0$ prawie wszędzie w $[a, b]$ oraz p – mierzalna, tzn.

$$\int_a^b p(x)dx < +\infty.$$

2. $\left\{\frac{1}{\sqrt{\pi}}, \frac{\cos x}{\sqrt{\pi}}, \frac{\sin x}{\sqrt{\pi}}, \ldots\right\}$ – ortogonalny w $[-\pi, \pi]$ z wagą $p = 1$.

5.2 Wielomiany ortogonalne

Definicja 5.17. *Dany niech będzie przedział* $[a, b]$ *oraz funkcja sumowalna* $p : [a, b] \to \mathbb{R}^+$, *ciąg* $\{P_j\}_{j=0,\ldots,k} \subset \Pi_n \subset L^2_p([a, b])$ *wielomianów takich, że* $\deg P_j = j$ *nazywamy ciągiem wielomianów ortogonalnych na* $[a, b]$ *z wagą* p, *jeśli* $(P_i|P_j) = 0$ *dla* $i \neq j$, *gdzie*

$$(u|v) = \int_a^b p(x)u(x)v(x)dx.$$

Uwaga 5.18. *Ciąg* $\{P_k\}_k$ *istnieje, co wynika z twierdzenia o ortogonalizacji zastosowanego do ciągu* $\{x^n\}_{n\in\mathbb{N}}$, *bo* $\{x^n\}_{n=0,\ldots,N}$ *jest liniowo niezależny dla dowolnego* N *(z zasadniczego twierdzenia algebry).*

Twierdzenie 5.19. *(własności ciągu wielomianów ortogonalnych)*

(i) $\Pi_n = \operatorname{span}\{P_0, \ldots, P_n\}$.

(ii) Jeżeli $Q_j \in \Pi_n$ *i* $\deg Q_j = j < k$, *to* $(Q_j|P_k) = 0$.

[14] Tzn. $\forall k \in \mathbb{N} : f_1, \ldots, f_k$ są liniowo niezależne.

(iii) $\{P_j\}_j$ *jest wyznaczony jednoznacznie z dokładnością do czynnika skalarnego, to znaczy, jeżeli* $\{Q_j\}_j$ *jest innym ciągiem wielomianów ortogonalnych, to*

$$Q_j = \gamma_j P_j,\ \gamma_j \in \mathbb{R} \setminus \{0\}.$$

(iv) $P_k(x) = (\alpha_k x + \beta_k)P_{k-1}(x) + \gamma_k P_{k-2}(x),\ k \geq 2$ – *reguła trójczłonowa.*

(v) Miejsca zerowe wielomianów ortogonalnych są pojedyncze oraz zawarte w $[a, b]$.

Dowód.

ad*(i)* $P_j \in \Pi_n,\ j = 0, \ldots, n$ i są one liniowo niezależne.

ad*(ii)* Niech $Q_j \in \Pi_n$ będzie takie, że $\deg Q_j = j < k$, wówczas, na mocy *(i)*

$$Q_j = \sum_{s=0}^{j} c_s P_s,$$

stąd

$$(P_k|Q_j) = \Big(P_k\Big|\sum_{s=0}^{j} c_s P_s\Big) = \sum_{s=0}^{j} c_s (P_k|P_s) = 0,\ \text{bo } j < k.$$

ad*(iii)* Niech $\{Q_j\}_j$ będzie innym ciągiem wielomianów ortogonalnych. Wówczas

$$P_k = \sum_{s=0}^{n} \alpha_s Q_s,$$
$$(P_k|Q_i) = \sum_{s=0}^{n} \alpha_s (Q_s|Q_i) = \alpha_i \|Q_i\|^2.$$

Dzięki *(ii)*, dla $i < k$ mamy:

$$0 = (P_k|Q_i) = \alpha_i \|Q_i\|^2,$$

zatem $\alpha_i = 0$. Dla $i = k$

$$0 \neq (P_k|Q_k) = \alpha_k \|Q_k\|^2,$$

więc $\alpha_k \neq 0$ i $P_k = \alpha_k Q_k$.

ad*(iv)* Skoro $P_k \in \Pi_k$ i $\deg P_k = k$, to zapisać możemy

$$P_k(x) = a_k x^k + \text{ pozostałe},\ a_k \neq 0.$$

Definiujemy

$$\begin{aligned} W(x) &= P_k(x) - x\tfrac{a_k}{a_{k-1}} P_{k-1}(x) \\ &= (a_k x^k + \text{ pozostałe }) - x\tfrac{a_k}{a_{k-1}}(a_{k-1}x^{k-1} + \text{ pozostałe }) \\ &= \text{ pozostałe} \in \Pi_{k-1} = \operatorname{span}\{P_0, \ldots, P_{k-1}\} \end{aligned}$$

Stąd

$$W(x) = \sum_{s=0}^{k-1} b_s P_s(x).$$

Dla $j < k-2$, mnożąc W skalarnie przez P_j, mamy

$$\begin{aligned}(W|P_j) &= \sum_{s=0}^{k-1} b_s(P_s|P_j) = b_j\|P_j\|^2,\\ (W|P_j) &= \left(P_k - x\tfrac{a_k}{a_{k-1}}P_{k-1}|P_j\right) = (P_k|P_j) - \tfrac{a_k}{a_{k-1}}(xP_{k-1}|P_j)\\ &= (P_k|P_j) - \tfrac{a_k}{a_{k-1}}(P_{k-1}|xP_j) = 0.\end{aligned}$$

Wobec tego $b_j = 0$ dla $j < k-2$, zatem

$$W = b_{k-1}P_{k-1} + b_{k-2}P_{k-2},$$

czyli

$$P_k - x\frac{a_k}{a_{k-1}}P_{k-1} = b_{k-1}P_{k-1} + b_{k-2}P_{k-2}. \tag{87}$$

Wystarczy więc wyznaczyć b_{k-1} i b_{k-2}. Mnożąc skalarnie (87) przez P_{k-1} mamy

$$(P_k|P_{k-1}) - \frac{a_k}{a_{k-1}}(xP_{k-1}|P_{k-1}) = b_{k-1}\|P_{k-1}\|^2,$$

czyli

$$b_{k-1} = -\frac{a_k}{a_{k-1}} \cdot \frac{(xP_{k-1}|P_{k-1})}{\|P_{k-1}\|^2}.$$

Analogicznie (mnożąc tym razem (87) przez P_{k-2}) otrzymamy

$$b_{k-2} = -\frac{a_k}{a_{k-1}} \cdot \frac{(xP_{k-1}|P_{k-2})}{\|P_{k-2}\|^2}.$$

Zatem, podstawiając wyliczone b_{k-1} i b_{k-2} do (87), dostajemy tezę, dla $\alpha_k = \frac{a_k}{a_{k-1}}$, $\beta_k = b_{k-1}$ oraz $\gamma_k = b_{k-2}$ otrzymujemy tezę.

Zauważmy, że

$$(xP_{k-1}|P_{k-2}) = (P_{k-1}|xP_{k-2}),$$

oraz że

$$xP_{k-2} \in \Pi_{k-1} \Rightarrow xP_{k-2} = \sum_{s=0}^{k-1} c_s P_s,$$

Stąd

$$(P_{k-1}|xP_{k-2}) = \sum_{s=0}^{k-1} c_s(P_{k-1}|P_s) = c_{k-1}\|P_{k-1}\|^2.$$

Ale

$$\begin{aligned} xP_{k-2} &= x(a_{k-2}x^{k-2} + \text{ pozostałe}) \\ &= c_{k-1}(a_{k-1}x^{k-1} + \text{ pozostałe}) + \text{ pozostałe}, \end{aligned}$$

zatem

$$\begin{aligned} a_{k-2} &= c_{k-1}a_{k-1}, \\ c_{k-1} &= \frac{a_{k-2}}{a_{k-1}} = \frac{1}{\alpha_{k-1}}. \end{aligned}$$

W takim razie

$$(P_{k-1}|xP_{k-2}) = \frac{1}{\alpha_{k-1}}\|P_{k-1}\|^2.$$

Stąd

$$\gamma_k = b_{k-2} = -\frac{\alpha_k}{\alpha_{k-1}} \cdot \frac{\|P_{k-1}\|^2}{\|P_{k-2}\|^2}.$$

Zauważmy, że jeżeli $[a, b] = [-m, m]$, to $k_{k-1} = 0$ dla $p(x) = p(-x)$. Dzieje się tak, bo $xP_{k-1}P_{k-1}$ jest funkcją nieparzystą.

ad *(v)* Załóżmy, że P_k zmienia znak w otoczeniu punktów $z_s \in [a, b]$, $s = 1, \ldots, m$, $m \leq k$. Wystarczy wykazać, że $m = k$, bo wówczas oznaczać to będzie, że wielomian P_k (stopnia k) ma k różnych pierwiastków, zatem na mocy zasadniczego twierdzenia algebry muszą być one pojedyncze. Dla dowodu nie wprost niech $m < k$, oraz niech

$$w_m(x) = \prod_{s=1}^{m}(x - z_s) \in \Pi_m.$$

Możemy zapisać

$$P_k(x) = a(x)w_m(x),$$

gdzie $a(x)$ jest stałego znaku w (a, b). Skoro $m < k$, to

$$0 \overset{(ii)}{=} (P_k|w_m) = \int_a^b p(x)a(x)w_m^2(x)dx > 0$$

– sprzeczność.

□

Przykład 5.20.

1. **Wielomiany Legendre'a**:

$$\begin{aligned} P_0(x) &= 1, \\ P_k(x) &= \frac{1}{2^k}\frac{1}{k!}\frac{d^k}{dx^k}(x^2 - 1)^k, k = 1, 2, \ldots \end{aligned}$$

Ciąg wielomianów $\{P_k\}_k$ jest ortogonalny na $[-1,1]$ z wagą $p(x) = 1$, $P_k \in L^2([-1,1])$. Jest tak, bo

$$\gamma_k(x) = \tfrac{1}{2^k}\tfrac{1}{k!}(x^2-1)^k \Rightarrow P_k(x) = \tfrac{d^k}{dx^k}\gamma_k(x),$$
$$\gamma_k'(x) = \tfrac{1}{2^k}\tfrac{1}{k!}(x^2-1)^{k-1}2kx = x\gamma_{k-1}(x),$$
$$\gamma_k(\pm 1) = \gamma_k'(\pm 1) = \ldots = \gamma_k^{(k-1)}(\pm 1) = 0,$$
$$(P_k|P_l) = \int_{-1}^{1} P_k(x)P_l(x)dx = \int_{-1}^{1} \frac{d^k}{dx^k}(\gamma_k(x))P_l(x)dx = [\text{części}]$$
$$= (-1)^k \int_{-1}^{1} \gamma_k(x)P_l^{(k)}(x)dx = 0,\ l < k,$$

bo $\gamma_l \in \Pi_{2l}$ oraz

$$P_l^{(k)}(x) = \frac{d^{k+l}}{dx^{k+l}}\gamma_l(x) = \begin{cases} 0, & k > l \\ \frac{(2k)!}{2^k k!}, & k = l \end{cases}$$

Dla $l = k$

$$(P_k|P_k) = \int_{-1}^{1} P_k(x)P_k(x)dx = (-1)^k \int_{-1}^{1} \gamma_k(x)P_k^{(k)}(x)dx$$
$$= (-1)^k \int_{-1}^{1} \frac{1}{2^k k!}(x^2-1)^k \cdot \frac{(2k)!}{2^k k!}dx = \frac{2}{2k+1}.$$

Wynik tem otrzymujemy całkując k razy przez części wyrażenie

$$\int (x-1)^k(x+1)^k dx.$$

Zauważmy jeszcze, że na mocy reguły trójczłonowej

$$P_k(x) = \frac{2k-1}{k}xP_{k-1}(x) - \frac{k-1}{k}P_{k-2}(x).$$

($\beta_k = 0$ bo $xP_{k-1}P_{k-1}$ jest funkcją nieparzystą)

2. **Wielomiany Hermite'a**:

$$H_k(x) = (-1)^k e^{x^2}\tfrac{d^k}{dx^k}e^{-x^2}, x \in \mathbb{R},\ k = 0,1,\ldots$$

są wielomianami ortogonalnymi z wagą $p(x) = e^{-x^2}$. $H_k \in L^2_p(\mathbb{R})$, to znaczy

$$\int_{\mathbb{R}} e^{-x^2}H_k^2 dx < \infty.$$

$$\int_{\mathbb{R}} W(x)p(x)dx < \infty,\ W \in \Pi_n,\ n = 1, 2, \ldots,$$

$$\int_{\mathbb{R}} e^{-x^2}\,dx - \lim_{r\to\infty} \int_{-r}^{r} x^{-x^2}\,dx,$$

$$\int_{\mathbb{R}^2} e^{-(x^2+y^2)}dxdy = \left| \begin{array}{l} x = \rho\cos\varphi \\ y = \rho\sin\varphi \end{array} \right| = \int_0^{\infty}\int_0^{2\pi} e^{-\rho^2}\rho\,d\varphi\,d\rho = \pi.$$

$$H_0(x) = 1,$$
$$H_k(x) = 2xH_{k-1}(x) - (2k-2)H_{k-2}(x),\ k = 1, 2, \ldots,$$
$$\|H_k\|^2 = \sqrt{\pi}2^k k!.$$

3. **Wielomiany Czebyszewa**:

(88) $T_k(x) = \cos(k\arccos x),\ x \in [-1, 1],\ k = 0, 1, \ldots$

są wielomianami ortogonalnymi z wagą $p(x) = \frac{1}{\sqrt{1-x^2}}$. Wykażemy, że to są rzeczywiście wielomiany. Połóżmy $x = \cos t$, wtedy

$$\begin{aligned} T_k(x) = T_k(\cos t) = \cos kt &= \tfrac{1}{2}[(\cos kt + i\sin kt) + (\cos kt - i\sin kt)] \\ &= \tfrac{1}{2}[(\cos t + i\sin t)^k + (\cos t - i\sin t)^k] \\ &= \tfrac{1}{2}\left[\left(x + i\sqrt{1-x^2}\right)^k + \left(x - i\sqrt{1-x^2}\right)^k\right]. \end{aligned}$$

Otrzymaliśmy więc, że

(89) $T_k(x) = \frac{1}{2}\left[\left(x + i\sqrt{1-x^2}\right)^k + \left(x - i\sqrt{1-x^2}\right)^k\right],$

zatem $T_k(x)$ jest wielomianem stopnia k. Podobny wynik można otrzymać korzystając ze wzoru na sumę kosinusów:

$$\cos kt + \cos(k-2)t = 2\cos(k-1)t \cdot \cos t,$$

oraz reguły trójczłonowej:

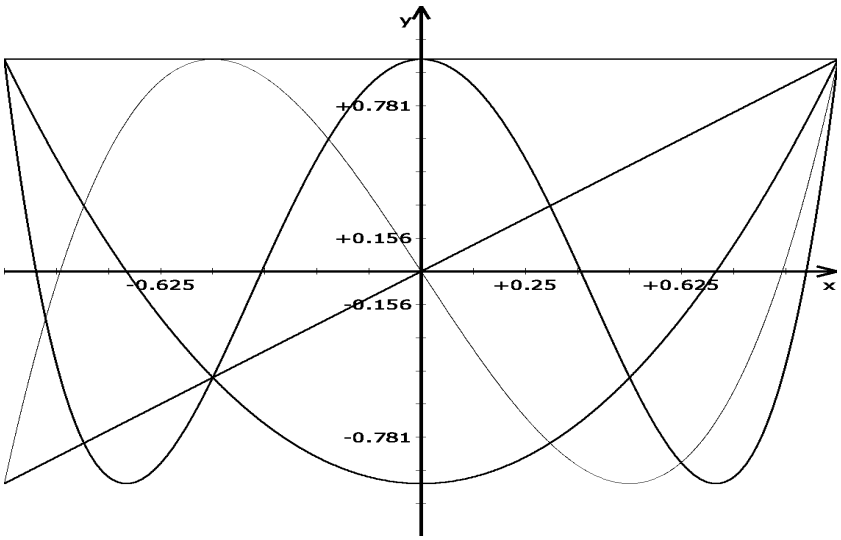
(90) $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x),\ k = 2, 3, \ldots,$

$$T_0(x) = 1,$$
$$T_1(x) = x.$$

Licząc dalej otrzymamy

$$T_2(x) = 2x^2 - 1,$$
$$T_3(x) = 4x^3 - 3x,$$
$$T_4(x) = 8x^4 - 8x^2 + 1.$$

Zapiszmy jeszcze pewne **własności wielomianów Czebyszewa**:



Rysunek 9: Wielomiany Czebyszewa

(a) $\{T_k\}_k$ jest ortogonalny w $[-1,1]$ z wagą $p(x) = \frac{1}{\sqrt{1-x^2}}$,

$$(T_k|T_l) = \int_{-1}^{1} \frac{1}{\sqrt{1-x^2}} \cos(k \arccos x) \cos(l \arccos x)\, dx = \begin{cases} 0, & k \neq l \\ \frac{\pi}{2}, & k = l \end{cases}$$

(b) na mocy (89) $T_k(-x) = (-1)^k T_k(x)$,

(c) na mocy (90) $T_k(x) = 2^{k-1}x^k +$ pozostałe,

(d) skoro T_k jest wielomianem ortogonalnym, to wszystkie zera leżą w $[-1,1]$, są one pojedyńcze, oraz $T_k(x_j) = 0 \Leftrightarrow \cos(k \arccos x_j) = 0 \Leftrightarrow k \arccos x_j = -\frac{\pi}{2} + j\pi,\ j = 0, \ldots, k-1 \Leftrightarrow x_j = \cos(\frac{2j-1}{2k}\pi),\ j = 0, \ldots, k-1$,

(e) punktami ekstremalnymi y_j, czyli takimi, że

$$T_k(y_j) = (-1)^j,$$

są:

$$k \arccos y_j = \pi j,\ j = 0, \ldots, k,$$
$$y_j = \cos \tfrac{\pi j}{k}.$$

5.2.1 Własności ekstremalne wielomianów Czebyszewa

Niech

$$\tilde{\Pi}_k = \{w \in \Pi_k |\ w(x) = x^k + a_{k-1}x^k + \ldots + a_0\},$$
$$\|u\|_c = \max\{|u(x)| :\ x \in [-1,1]\} \text{ – norma na } X = \mathcal{C}([-1,1], \mathbb{R}).$$

Zatem, na mocy własności (c) wielomianów Czebyszewa $\tilde{T}_k = \frac{1}{2^{k-1}} T_k \in \tilde{\Pi}_k$.

Twierdzenie 5.21. $\tilde{T}_k$ *jest wielomianem najmniej odchylającym się od zera, tzn.:*

$$\|\tilde{T}_k\|_c \leq \|w\|_c,\ \forall w \in \tilde{\Pi}_k.$$

Dowód. Dla dowodu nie wprost załóżmy, że istnieje $w^\star \in \tilde{\Pi}_k$ taki, że

$$\|w^\star\|_c < \|\tilde{T}_k\|_c = \tfrac{1}{2^{k-1}}.$$

Dla $x \in [-1,1]$

$$|w^\star(x)| < \tfrac{1}{2^{k-1}}.$$

Niech $y_0, \ldots, y_k$ będą takimi punktami, że $T_k(y_j) = (-1)^j,\ j = 0, \ldots, k$. Wówczas

$$\begin{aligned}
&\tilde{T}_k(y_0) = \tfrac{1}{2^{k-1}} > w^\star(y_0),\\
&\tilde{T}_k(y_1) = -\tfrac{1}{2^{k-1}} < w^\star(y_1),\\
&\tilde{T}_k(y_2) = \tfrac{1}{2^{k-1}} > w^\star(y_2)\ldots.
\end{aligned}$$

Wówczas $0 \neq \nu(x) = w^\star(x) - \tilde{T}_k(x)$ jest wielomianem stopnia $k-1$ (bo oba są unormowane), zerującym się w punktach przedziałów $(y_j, y_{j+1}),\ j = 0, \ldots, k-1$, zatem na mocy zasadniczego twierdzenia algebry $\nu \equiv 0$. □

Udowodnimy jeszcze jedno twierdzenie, które w następnej sekcji zostanie udowodnione ponownie, innymi metodami.

Twierdzenie 5.22. *(o wielomianach Czebyszewa)*
Niech a będzie takie, że $|a| > 1$, oraz $A \neq 0$. Wówczas dla dowolnego $w \in \Pi_k$ takiego, że $w(a) = A$ zachodzi

$$\left\| \tfrac{A}{T_k(a)} T_k \right\|_{[-1,1]} \leq \|w\|_{[-1,1]}.$$

Dowód. Przeprowadzimy dowód nie wprost. Przypuśćmy, że istnieje $w^\star \in \Pi_k$ taki, że

$$w^\star(a) = A \text{ oraz } \|w^\star\|_c < \left\| \tfrac{A}{T_k(a)} T_k \right\|_c.$$

Niech $y_0, \ldots, y_k$ będą takie, że $T_k(y_j) = (-1)^j$, wówczas

$$\begin{aligned}
&\left| \tfrac{A}{T_k(a)} \right| T_k(y_0) > w^\star(y_0),\\
&\left| \tfrac{A}{T_k(a)} \right| T_k(y_1) < w^\star(y_1),\\
&\ldots
\end{aligned}$$

Zatem

$$\nu(x) = w^\star(x) - \tfrac{A}{T_k(a)} T_k(x) \in \Pi_k$$

zeruje się w każdym z przedziałów $(y_i, y_{i+1}),\ j = 0, \ldots, k-1$. Ale $\nu(a) = 0$, zatem zeruje się on łącznie w $k+1$ punktach. Na mocy zasadniczego twierdzenia algebry $\nu \equiv 0$, co kończy dowód. □

5.3 Aproksymacja jednostajna

Dane niech będą $X = \mathcal{C}([a,b],\mathbb{R})$, $[a,b]$ zwarty ($K \subset \mathbb{R}^k$ zwarty), $u \in X$, $\|u\|_c = \max\{|u(x)| : x \in [a,b]\}$, $U \subseteq X$ – podprzestrzeń. Dla $u \in X$ szukać będziemy $v \in U$ takiego, że $\|u-v\|_c \leq \|u-y\|_c \ \forall y \in U$.

Definicja 5.23. *Mówimy, że $U \subset X$ – n-wymiarowa podprzestrzeń spełnia warunek Haara, jeśli $\forall f \in U \setminus \{0\} : f(x) = 0$ w co najwyżej $n-1$ punktach przedziału $[a,b]$.*

Uwaga 5.24. *Warunek Haara równoważny jest temu, że jeśli funkcja f zeruje się w n różnych punktach przedziału $[a,b]$ to musi być ona tożsamościowo równa zeru.*

Przykład 5.25.

1. $U = \Pi_{n-1} \subseteq \mathcal{C}([a,b],\mathbb{R})$ spełnia warunek Haara, co wynika z zasadniczego twierdzenia algebry.

2. $F = \left\{t_n : t_n(x) = \frac{1}{2}a_0 + \sum_{j=1}^{n}(a_j \cos jx + b_j \sin jx),\ a_j, b_j \in \mathbb{R}\right\}$ – przestrzeń wielomianów trygonometrycznych, $\dim F = 2n+1$, $F \subset \mathcal{C}([0, 2\pi - \varepsilon],\mathbb{R})$. W postaci zespolonej F przyjmuje postać $F = \left\{t_n : t_n(x) = \sum_{j=0}^{2n} c_n e^{ijx}\right\}$, oraz spełnia warunek Haara w $[0, 2\pi-\varepsilon]$ (na mocy twierdzenia 4.23).

Twierdzenie 5.26. *Jeżeli $U \subset X = \mathcal{C}([a,b],\mathbb{R})$ jest podprzestrzenią n-wymiarową, to następujące warunki są równoważne:*

(i) *U spełnia warunek Haara,*

(ii) *$\forall x_1,\ldots,x_n \in [a,b]$ parami różnych, $\forall y_1,\ldots,y_n \in \mathbb{R}\ \exists! w \in U : w(x_j) = y_j,\ j = 1,\ldots,n$ (własność interpolacji),*

(iii) *$\forall x_1,\ldots,x_n \in [a,b]$ parami różnych, $\forall f_1,\ldots,f_n \in U$ liniowo niezależnych $\det A \neq 0$, gdzie $A = (a_{ij})_{i,j}$, $a_{ij} = f_j(x_i)$.*

Dowód.
(i)⇒(iii)
Przypuśćmy, że istnieją $x_1,\ldots,x_n \in [a,b]$ parami różne, oraz $f_1,\ldots,f_n$ wektory bazowe U takie, że $\det A = 0$. Zatem istnieje $\alpha \in \mathbb{R}^n \setminus \{0\}$ taki, że $A\alpha = 0$. Niech

$$f = \sum_{j=1}^{n} \alpha_j f_j.$$

Oczywiście $f \in U$ oraz f zeruje się w n różnych punktach $x_1,\ldots,x_n$, zatem $f \equiv 0$, czyli

$$\sum_{j=1}^{n} \alpha_j f_j = 0.$$

Skoro $f_1,\ldots,f_n$ są liniowo niezależne, to $\alpha_1 = \ldots = \alpha_n = 0$ – sprzeczność.
(iii)⇒(ii)

Niech $x_1, \ldots, x_n \in [a, b]$ będą parami różne i niech $y_1, \ldots, y_n \in \mathbb{R}$. Skoro $\det A \neq 0$, to

$$\sum_{j=1}^{n} \alpha_j f_j(x_i) = y_i, \ i = 1, \ldots, n$$

ma dokładnie jedno rozwiązanie α. Jako w wystarczy przyjąć $\sum_{j=1}^{n} \alpha_j f_j$.
(ii)⇒(iii)
Niech $x_1, \ldots, x_n \in [a, b]$ będą parami różne, $f_1, \ldots, f_n$ liniowo niezależne. Skoro dla dowolnego $y = (y_1, \ldots, y_n)^T$ istnieje dokładnie jedena $w \in U$ taka, że $w(x_i) = y_i$, to

$$\forall y \in \mathbb{R}^n \ \exists! \alpha = (\alpha_1, \ldots, \alpha_n)^T : A\alpha = y.$$

Więc macierz A jest nieosobliwa.
(iii)⇒(i)
Niech $f \in U$ zeruje się w n różnych punktach $x_1, \ldots, x_n$. Skoro $f_1, \ldots, f_n$ jest bazą U, to istnieje $\alpha \in \mathbb{R}^n$ takie, że

$$f = \sum_{j=1}^{n} \alpha_j f_j.$$

Więc

$$\sum_{j=1}^{n} \alpha_j f_j(x_i) = 0, \ i = 1, \ldots, n.$$

Skoro $\det A \neq 0$, to jedynym rozwiązaniem jest $\alpha = 0$, czyli $f \equiv 0$. □

Twierdzenie 5.27. *(jednoznaczność rozwiązania optymalnego)*
Bez dowodu *Jeżeli $u \in X = \mathcal{C}([a, b], \mathbb{R})$ i $U \subset X$ jest n-wymiarową podprzestrzenią, to następujące warunki są równoważne:*

(i) $\exists! v \in U$ *element optymalny dla* u*,*

(ii) U *spełnia warunek Haara.*

Twierdzenie 5.28. *(o alternansie)*
Bez dowodu *Jeżeli $u \in X = \mathcal{C}([a, b], \mathbb{R})$ i $U \subset X$ jest n-wymiarową podprzestrzenią spełniającą warunek Haara, to warunkiem koniecznym i wystarczającym na to, by $v \in U \setminus \{0\}$ był elementem optymalnym dla $u \in X$ jest istnienie alternansu, tzn. ciągu $a \leq x_0 < x_1 < \ldots < x_n \leq b$ takiego, że*

(i) $|u(x_i) - v(x_i)| = \|u - v\|_c, \ i = 0, \ldots, n,$

(ii) $u(x_{i+1}) - v(x_{i+1}) = -(u(x_i) - v(x_i)), \ i = 0, \ldots, n-1.$

Uwaga 5.29. *Twierdzenia 5.26 – 5.28 są prawdziwe dla $X = \mathcal{C}(K, \mathbb{R})$, $K \subseteq \mathbb{R}^p$ zwartego.*

Uwaga 5.30. *Twierdzenie 5.28 jest prawdziwe dla $U = U_0 + g = \{\omega \in X : \omega = z + g, \ z \in U_0, g \in X\}$.*

Twierdzenie 5.31. *(o wielomianach Czebyszewa)*
Niech a *będzie takie, że* $|a| > 1$*, oraz* $A \neq 0$*. Wówczas dla dowolnego* $w \in \Pi_k$ *takiego, że* $w(a) = A$ *zachodzi*

$$\left\| \frac{A}{T_k(a)} T_k \right\|_{[-1,1]} \leq \|w\|_{[-1,1]}.$$

Dowód. Niech

$$U = \{w \in \Pi_k : w(a) = A\} = \{w \in \Pi_k : w(a) = 0\} + A = U_0 + A,$$

(warunek $w(a) = A$ „zabiera" jeden wymiar, zatem U_0 jest wymiaru k). U_0 spełnia warunek Haara na przedziale $[-1, 1]$ ponieważ $|a| > 1$. Niech $y_0, \ldots, y_k \in [-1, 1]$ będą takie, że $T_k(y_j) = (-1)^j$. Wówczas, dla $v(x) = \frac{A}{T_k(a)} T_k(x)$ spełnione są warunki twierdzenia o alternansie (dla $u \equiv 0$)

$$v(y_j) = \|v\|_{c([-1,1])}, \ j = 0, \ldots, k,$$
$$v(y_{j+1}) = -v(y_j), \ j = 0, \ldots, k-1.$$

Zatem v jest elementem optymalnym dla 0, co kończy dowód. □

Dane niech będą $X = \mathcal{C}([a, b], \mathbb{R})$, $U \subset X$ podprzestrzeń wymiaru n oraz $\{f_1, \ldots, f_n\}$ – baza U. Niech $x_0, \ldots, x_n$ będzie alternansem, a $u \in U$ – elementem optymalnym dla $f \in X$. Chcemy wyznaczyć $\alpha_1, \ldots, \alpha_n$ takie, że

$$u = \sum_{j=1}^{n} \alpha_j f_j.$$

Przez $e_u(f) = \|u - f\|_c$ rozumiemy błąd aproksymacji. Dzięki twierdzeniu 5.28 (o alternansie) mamy:

$$f(x_i) - u(x_i) = \varepsilon e_u(f) \cdot (-1)^i, \ i = 0, \ldots, n, \ \varepsilon = \pm 1,$$
$$f(x_i) = \sum_{j=1}^{n} \alpha_j f_j(x_i) + \varepsilon e_u(f)(-1)^i, \ i = 0, \ldots, n.$$

Mamy w takim razie $n + 1$ równań liniowych o niewiadomych $\alpha_1, \ldots, \alpha_n, \varepsilon e_u(f)$. Macierzą tego układu jest

$$\begin{pmatrix} f_1(x_0) & \ldots & f_n(x_0) & (-1)^0 \\ \vdots & \ddots & \vdots & \vdots \\ f_1(x_n) & \ldots & f_n(x_n) & (-1)^n \end{pmatrix}$$

Ponieważ jest ona nieosobliwa (dzięki jednoznaczności rozwiązania), to możemy wyznaczyć poszukiwane przez nas $\alpha_1, \ldots, \alpha_n$.

6 Całkowanie numeryczne.

W rozdziale tym będziemy szukać wartości całki

$$I_p(f) = \int_a^b p(x) f(x) dx,$$

mając dane:

– zwarty przedział $[a, b]$,

– sumowalną funkcję (wagę) $p : [a, b] \to [0, \infty)$, $p(x) > 0$ p.w.,

– funkcję $f \in \mathcal{C}([a, b], \mathbb{R})$.

Jeżeli $p \equiv 1$ to $I_1(f) = I(f)$.

Przez $Q(f)$ będziemy oznaczać przybliżoną wartość całki – **kwadraturę**, przez $R(f)$ – resztę kwadratury. Zatem

$$I_p(f) = Q(f) + R(f).$$

Definicja 6.1. *Kwadraturę $Q(f)$ nazywamy **kwadraturą rzędu** n, jeśli $I_p(f) = Q(f)$ dla dowolnej $f \in \Pi_{n-1}$ (Q jest dokładna) oraz istnieje $u \in \Pi_n$ taka, że $I_p(u) \neq Q(u)$.*

Definicja 6.2. *Kwadraturę nazywamy **kwadraturą liniową**, gdy*

$$\begin{aligned} Q(f) &= \sum_{j=0}^{n_0} A_{0j} f(x_{0j}) + \sum_{j=0}^{n_1} A_{1j} f'(x_{1j}) + \ldots + \sum_{j=0}^{n_k} A_{kj} f^{(k)}(x_{kj}) \\ &= \sum_{i,j} A_{ij} f^{(i)}(x_{ij}) [0 \leq j \leq n_i,\ 0 \leq i \leq k], \end{aligned}$$

to znaczy, gdy kwadratura zależy liniowo od $f, f', \ldots, f^{(k)}$ wyliczanych w węzłach $\{x_{0j}\}_{j=0}^{n_0}, \ldots, \{x_{kj}\}_{j=0}^{n_k}$.

*Współczynniki A_{ij} nazywane są **współczynnikami kwadratury**.*

Definicja 6.3. *Kwadraturę Q nazywamy **kwadraturą interpolacyjną**, jeśli*

$$Q(f) = I_p(H_N),$$

gdzie H_N jest wielomianem interpolacyjnym w postaci Hermite'a dla f z węzłami $x_0, \ldots, x_n$ o krotnościach odpowiednio $m_0, \ldots, m_n$, przy czym całkowita krotność $\sum_{j=0}^{n} m_j = N + 1$.

*Kwadratura interpolacyjna Q jest **liniowa**, jeśli*

$$Q(f) = \sum_{j=0}^{n} \sum_{i=0}^{m_j - 1} A_{ij} f^{(i)}(x_j).$$

Uwaga 6.4. *Kwadratura interpolacyjna oparta na węzłach o łącznej krotności $N + 1$ jest co najmniej rzędu $N + 1$.*

Dowód.

$$\begin{aligned} R(f) = I_p(f) - Q(f) &= \int_a^b p(x) f(x)\,dx - \int_a^b p(x) H_N(x)\,dx = \int_a^b p(x)[f(x) - H_N(x)]\,dx \\ &= \int_a^b p(x) r(x)\,dx, \end{aligned}$$

gdzie

$$r(x) = \tfrac{f^{(N+1)}(\xi)}{(N+1)!} p_{N+1}(x),\ p_{N+1}(x) = (x - x_0)^{m_0} \cdot \ldots \cdot (x - x_n)^{m_n}$$

jest resztą interpolacji Hermite'a. Jeśli więc $f \in \Pi_N$, to $f^{(N+1)}(x) = 0$, co kończy dowód. □

6.1 Kwadratury Newtona-Cotesa.

Definicja 6.5. *Kwadraturę* $Q(f) = I_p(L_n)$, *gdzie* L_n *jest wielomianem interpolacyjnym w postaci Lagrange'a opartym na równoodległych węzłach, nazywamy* ***kwadraturą Newtona-Cotesa.***

Wyprowadźmy wzór na tą kwadraturę. Wiemy, że

$$L_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x),$$

gdzie

$$l_j(x) = \frac{(x-x_0)\dots(\widehat{x-x_j})\dots(x-x_n)}{(x_j-x_0)\dots(\widehat{x_j-x_j})\dots(x_j-x_n)},\ x_i = a+ih,\ h = \tfrac{b-a}{n}.$$

Dla uproszczenia przyjmujemy $p(x) \equiv 1$. Wówczas

$$Q(f) = \int_a^b L_n(x)dx = \sum_{j=0}^{n} f(x_j) \int_a^b l_j(x)dx = \sum_{j=0}^{n} \tilde{\alpha}_j f(x_j),$$

gdzie

$$\tilde{\alpha}_j = \int_a^b l_j(x)dx = \int_a^b \prod_{i\neq j} \frac{x-x_i}{x_j-x_i}dx = \left| \begin{array}{l} x = a+hs \\ dx = hds \\ x-x_i = (s-i)h \\ x_j-x_i = (j-i)h \end{array} \right| = \int_0^n \prod_{i\neq j} \frac{(s-i)h}{(j-i)h} hds$$

$$= h\int_0^n \prod_{i=0,i\neq j}^{n} \frac{s-i}{j-i} ds.$$

Zatem

$$(91)\quad Q(f) = h\sum_{j=0}^{n} \alpha_j f(x_j),\ \alpha_j = \int_0^n \prod_{i=0,i\neq j}^{n} \frac{s-i}{j-i} ds.$$

Uwaga 6.6.

$$\sum_{j=0}^{n} \alpha_j = n.$$

Dowód. Niech $f \equiv 1$. Oczywiście $I(1) = Q(1)$, Ale

$$I(1) = \int_a^b dx = b-a \text{ oraz } Q(1) = h\sum_{j=0}^{n} \alpha_j,$$

zatem

$$\sum_{j=0}^{n} \alpha_j = \tfrac{b-a}{h} = n.$$

□

Wzór (91) można przedstawić jako

$$(92)\quad Q(f) = \frac{b-a}{sn}\sum_{j=0}^{n}\sigma_j f(x_j),$$

gdzie $\sigma_j = \alpha_j s$, przy czym s dobieramy tak, aby σ_j było całkowite.

n	σ_j	ns	$\lvert R\rvert$	nazwa
1	1 1	2	$h^3\frac{1}{12}f''(\xi)$	wzór trapezów
2	1 4 1	6	$h^5\frac{1}{90}f^{(4)}(\xi)$	wzór Simpsona
3	1 3 3 1	8	$h^5\frac{3}{80}f^{(4)}(\xi)$	wzór „trzech ósmych”
4	7 32 12 32 7	90	$h^7\frac{8}{945}f^{(6)}(\xi)$	wzór Milne'a

Dla $n \geq 7$ $(n \neq 9)$ σ_j zmieniają znaki i wzory stają się niepraktyczne.

6.1.1 Reszta kwadratur Newtona-Cotesa.

$p \equiv 1$

$$R(f) = I(f) - Q(f) = \int_a^b (f(x) - L_n(x))dx = \int_a^b r(x)dx = I(r),$$

gdzie $r(x) = f(x) - L_n(x)$ jest resztą interpolacyjną Lagrange'a.

Przyjrzyjmy się bliżej tym resztom, gdy $n = 1$ i $n = 2$. Dla pierwszego przypadku węzłami są $x_0 = a$ i $x_1 = b$, więc

$$r(x) = \frac{f''(\xi)}{2}(x-a)(x-b),$$

$(x-a)(x-b) < 0$ w (a,b) zatem, korzystając z twierdzenia o wartości średniej dla całki

$$R(f) = \int_a^b \frac{f''(\xi)}{2}(x-a)(x-b)dx = \frac{f''(\eta)}{2}\int_a^b (x-a)(x-b)dx$$
$$= -\frac{f''(\eta)}{2}\frac{(b-a)^3}{6} = \frac{h^3}{12}f''(\eta).$$

Dla $n = 2$ węzłami są $x_0 = a,\ x_1 = \frac{a+b}{2},\ x_2 = b$. Niech $H_3(x)$ będzie wielomianem interpolacyjnym Hermite'a dla f z węzłami x_0, x_1, x_2 o krotnościach 1,2,1. Możemy zapisać

$$H_3(x) = L_2(x) + K(x-x_0)(x-x_1)(x-x_2),$$

gdzie K dobieramy tak, aby $H_3'(x_1) = f'(x_1)$. Niech r oznacza resztę interpolacyjną

Hermite'a. Na mocy twierdzenia o reszcie

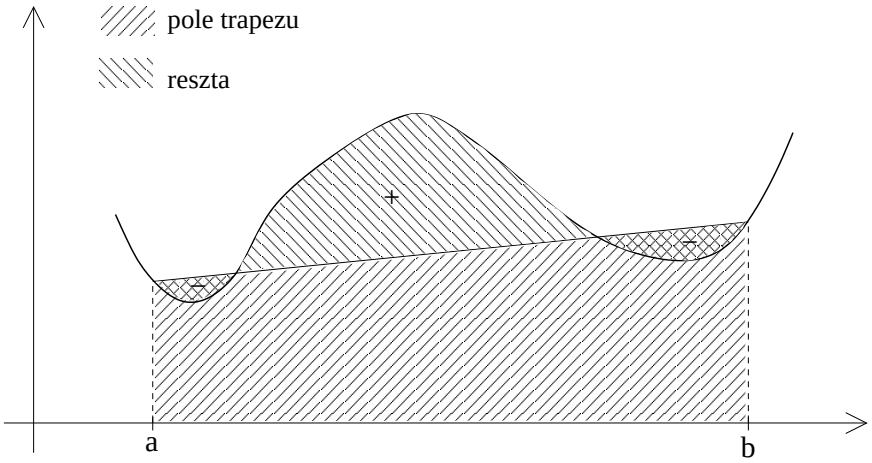
$$\begin{aligned} f(x) &= H_3(x) + r(x) = H_3(x) + \tfrac{f^{(4)}(\xi)}{4!}(x-x_0)(x-x_1)^2(x-x_2) \\ &= L_2(x) + K(x-x_0)(x-x_1)(x-x_2) + \tfrac{f^{(4)}(\xi)}{4!}(x-x_0)(x-x_1)^2(x-x_2), \\ R(f) &= \int_a^b (f(x) - L_2(x))dx \\ &= \int_a^b K(x-a)\big(x-\tfrac{a+b}{2}\big)(x-b)dx + \int_a^b \tfrac{f^{(4)}(\xi)}{4!}(x-x_0)(x-x_1)^2(x-x_2)dx \\ &= \tfrac{f^{(4)}(\eta)}{4!}\int_a^b (x-a)\big(x-\tfrac{a+b}{2}\big)^2(x-b)dx = \tfrac{(b-a)^5}{90}f^{(4)}(\eta) = \tfrac{h^5}{90}f^{(4)}(\eta), \end{aligned}$$

gdzie jedna z całek (pierwsza) zeruje się ze względu na parzystość względem środka przedziału całkowania.

Twierdzenie 6.7. *Dla dowolnego* $1 \le n \le 6$ *i funkcji* f *dostatecznie regularnej kwadratura Newtona-Cotesa oparta na* $n+1$ *węzłach ma rząd* $n+1$*, gdy* n *jest liczbą nieparzystą, albo też rząd* $n+2$*, gdy* n *jest liczbą parzystą.*

Dla $f \in \Pi_n$ mamy $R(f) = 0$, bo $Q(f) = I(L_n) - I(f)$. Więc kwadratury Newtona-Cotesa są rzędu przynajmniej $n+1$.

Przykład 6.8. Poniższe przykłady ilustrują reszty dla $n = 1, 2$.



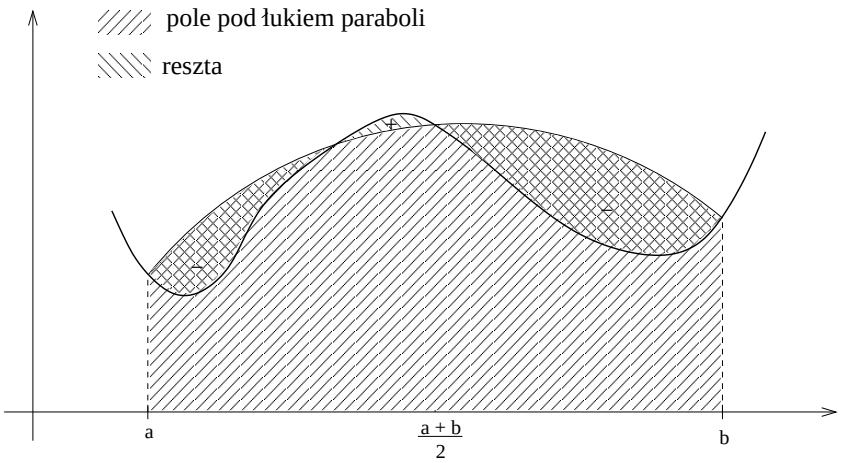
Rysunek 10: $n = 1$

6.2 Kwadratury złożone Newtona-Cotesa.

Dana niech będzie funkcja ciągła $f : [a, b] \to \mathbb{R}$. Przeprowadzamy następujące kroki

1. Dzielimy przedział $[a, b]$ na N przedziałów z węzłami

$$x_i = a + i\tfrac{b-a}{N},\ i = 0, \ldots, N.$$



Rysunek 11: $n = 2$

2. Przedziały $[x_i, x_{i+1}]$ $(i = 0, \ldots, N-1)$ dzielimy na n części
$$x_{ij} = x_i + j\frac{x_{i+1}-x_i}{n},\ j = 0, \ldots, n.$$

3. W każdym z przedziałów $[x_i, x_{i+1}]$ $(i = 0, \ldots, N-1)$ stosujemy kwadraturę Newtona-Cotesa opartą na $n+1$ węzłach, czyli
$$I^{(i)}(f) = \int_{x_i}^{x_{i+1}} f(x)dx = \frac{h}{n}\sum_{j=0}^{n} \alpha_j f(x_{ij}) + R_i(f) = Q_i(f) + R_i(f),\ h = \frac{b-a}{N},$$

$$(93)\quad Q_N(f) = \sum_{i=0}^{N-1} Q_i(f) = \frac{h}{n}\sum_{i=0}^{N-1}\sum_{j=0}^{n} \alpha_j f(x_{ij}),$$

$$(94)\quad R_N(f) = \sum_{i=0}^{N-1} R_i(f).$$

Uwaga 6.9. *Kwadratura złożona nie jest kwadraturą interpolacyjną.*

Przykład 6.10. Rozpatrzmy przykład dla $n = 1$, a więc złożoną kwadraturę trapezów. Wówczas

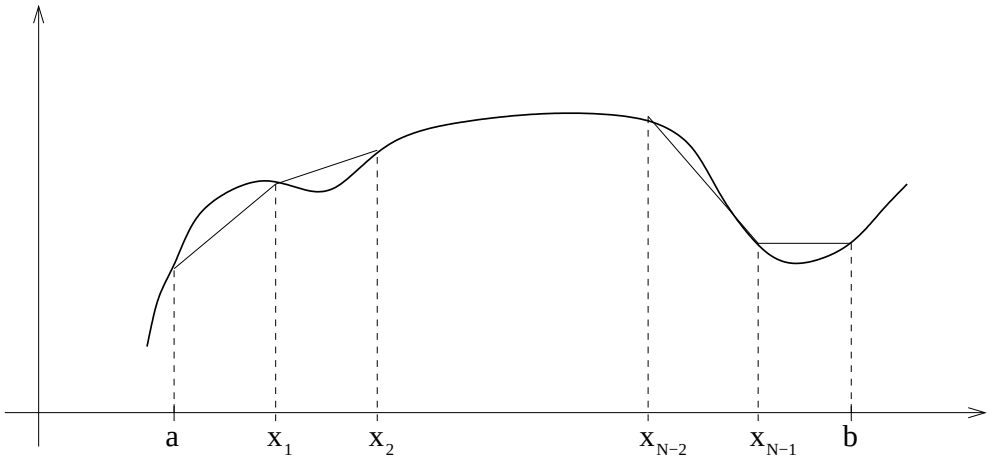
$$\alpha_0 = \alpha_1 = \frac{1}{2},$$
$$x_{ij} = x_{i+j},\ j = 0, 1,\ i = 0, \ldots, N-1.$$

Zatem, na mocy (93)

$$Q_N(f) = h\sum_{i=0}^{N-1} \frac{1}{2}(f(x_i) + f(x_{i+1})) = \frac{h}{2}\big(f(a) + 2\sum_{i=1}^{N} f(a+ih) + f(b)\big)$$

Wykażemy jeszcze, że reszta tej kwadratury zmierza do zera. Załóżmy, że $|f''(x)| \leq M$ dla $x \in [a, b]$, wówczas

$$|R_N(f)| \leq \sum_{i=0}^{N-1} |R_i(f)| = \sum_{i=0}^{N-1} \frac{h^3}{12}|f''(\xi_i)| \leq \frac{h^3}{12}NM = \frac{(b-a)^3 M}{N^2}.$$



Rysunek 12: złożona kwadratura trapezów

Zatem

$$\lim_{N\to\infty} R_N(f) = 0.$$

Twierdzenie 6.11. *Dla* $1 \le n \le 6$ *ciąg kwadratur* $(Q_N(f))_N$ *jest zbieżny do* $I(f)$.

Kiedy należy przerwać obliczenia? Rozpatrzmy przypadek $n = 1$ (dla $n > 1$ podobnie). Należy postępować zgodnie z następującym przepisem:

- ▷ wybieramy N_k, na przykład $N_k = m^k$, gdzie m jest ustaloną liczbą naturalną,
- ▷ ustalamy $\varepsilon > 0$,
- ▷ przerywamy obliczenia, gdy $\frac{|Q_{N_k}(f)-Q_{N_{k-1}}(f)|}{|Q_{N_k}(f)|} < \varepsilon$,
- ▷ wtedy $I(f) \approx Q_{N_k}(f)$.

6.3 Kwadratury Gaussa.

Zastanówny się, jaki może być maksymalny rząd kwadratury liniowe opartej na ustalonej liczbie węzłów. Niech więc Q będzie kwadraturą opartą na dowolnych węzłach $x_0, \dots, x_n$ o krotnościach $m_0, \dots, m_n$ (o łącznej sumie $N+1$). Zgodnie z definicją 6.2

$$Q(f) = \sum_{j=0}^{n} \sum_{i=0}^{m_j-1} A_{ij} f^{(i)}(x_j). \tag{95}$$

Na mocy uwagi 6.4 wiemy, że rząd kwadratury opartej na węzłach o łącznej krotności $N+1$ jest co najmniej równy $N+1$ (bo co najmniej taki jest rząd kwadratury interpolacyjnej opartej na węzłach o krotności $N+1$). Co więcej okazuje się, że dowolna kwadratura postaci (95), rzędu wyższego od $N+1$, jest kwadraturą interpolacyjną.

Lemat 6.12. *Jeśli kwadratura* Q *postaci (95) jest rzędu* $r > \sum_{i=0}^{n} m_i > N+1$*, to jest ona kwadraturą interpolacyjną.*

Dowód. Niech H_N będzie wielomianem interpolacyjnym w postaci Hermite'a dla funkcji f opartym na węzłach $x_0, \ldots, x_n$ o krotnościach $m_0, \ldots, m_n$. Zatem

$$H_N^{(j)}(x_i) = f^{(j)}(x_i),\ i = 0, \ldots, n,\ j = 0, \ldots, m_i - 1,$$

więc

$$Q(f) = Q(H_N).$$

Skoro kwadratura Q jest rzędu $r > N + 1$, czyli jest dokładna dla wielomianu H_N stopnia co najwyżej N, to

$$Q(f) = I(H_N),$$

co kończy dowód. □

Niech $x_0, \ldots, x_N$ będą zerami wielomianu P_{N+1} ortogonalnego na przedziale $[a, b]$ z iloczynem skalarnym

$$(u|v) = \int_a^b u(x)v(x)dx.$$

Wówczas

$$Q(f) = \sum_{i=0}^{N} A_i f(x_i) \tag{96}$$

nazywamy **kwadraturą Gaussa**.

Twierdzenie 6.13. *Kwadratura Gaussa jest rzędu* $2N + 2$. *Jej współczynniki dane są wzorem*

$$A_i = \int_a^b \prod_{j=0\ j\neq i}^{N} \frac{x - x_j}{x_i - x_j} dx.$$

Dowód. Niech $f \in \Pi_{2N-1}$, wówczas

$$f(x) = a(x)P_{N+1}(x) + r(x),$$

gdzie

$$\deg a \leq 2N - 1 - N + 1 = N,$$
$$\deg r < N + 1.$$

Wówczas

$$\int_a^b f(x)dx = \int_a^b r(x)dx.$$

Skoro

$$f(x_i) = r(x_i),\ i = 0, \ldots, N,$$

to r jest wielomianem interpolacyjnym dla funkcji f opartym na jednokrotnych węzłach $x_0, \ldots, x_N$. Możemy więc zapisać

$$r(x) = \sum_{i=0}^{N} f(x_i) l_i(x).$$

Oczywiście, podobnie jak w przypadku kwadratur Newtona-Cotesa

$$\int_a^b r(x)dx = \sum_{i=0}^{N} f(x_i) \int_a^b l_i(x)dx = \sum_{i=0}^{N} A_i f(x_i) = Q(f),$$

gdzie

$$A_i = \int_a^b \prod_{j=0\, j\neq i}^{N} \frac{x - x_j}{x_i - x_j} dx.$$

Wykazaliśmy więc, że kwadratura (96) jest rzędu co najmniej $2N + 2$. Aby pokazać że jest rzędu co najwyżej $2N+2$ wystarczy wskazać wielomian stopnia $2N+2$ dla którego kwadratura nie daje dokładnego wyniku. Niech więc

$$p(x) = (x - x_0) \ldots (x - x_N),$$

gdzie $x_0, \ldots, x_N$ są (pojedynczymi!) zerami wielomianu P_{N+1}. Wówczas p^2 jest stopnia $2N + 2$ oraz

$$I(p^2) = \int_a^b (x - x_0)^2 \ldots (x - x_N)^2 dx > 0,$$
$$Q(p^2) = \sum_{i=0}^{N} A_i p^2(x_i) = 0,$$

co kończy dowód. □

Uwaga 6.14. *Jeżeli rozpatrujemy całkę z wagą* p*, to iloczyn skalarny też należy wziąć z tą samą wagą, to znaczy*

$$(u|v) = \int_a^b p(x)u(x)v(x)dx,$$

Metoda ta ma dwa mankamenty. Pierwszym jest trudność wyliczenia zer wielomianu ortogonalnego, a drugim fakt, że jeśli dodamy dodatkowy węzeł, to wszystkie obliczenia należy powtórzyć.

Twierdzenie 6.15. $A_j > 0$.

Dowód. Dla $w_i(x) = \prod_{i \neq j=0}^{N} (x - x_j)^2 \in \Pi_{2N}$ mamy

$w_i > 0$

$$0 < \int_a^b w_i(x)dx = I(w_i) \overset{\text{tw. } 6.13}{=} Q(w_i) = A_i w_i(x_i),$$

zatem $A_i > 0$. □

6.3.1 Reszta kwadratur Gaussa

Twierdzenie 6.16. *(Reszta kwadratur Gaussa)*
Jeżeli $f \in \mathcal{C}^{2N+2}$*, to reszta kwadratury Gaussa dana jest wzorem*

$$R(f) = \frac{f^{(2N+2)}(\xi)}{(2N+2)!} \int_a^b p_{N+1}^2(x)dx, \ p_{N+1}(x) = (x - x_0)\dots(x - x_N). \tag{97}$$

Dowód. Dla $H_{2N+1}(x)$ – wielomianu interpolacyjnego w postaci Hermite'a opartego na węzłach $x_0, \dots, x_N$ o krotnościach 2 mamy (dla $i = 0, \dots, N$):

$$H_{2N+1}(x_i) = f(x_i)$$
$$H'_{2N+1}(x_i) = f'(x_i).$$

Oczywiście $H_{2N+1} \in \Pi_{2N+1}$. Na mocy twierdzenia o reszcie interpolacyjnej Hermite'a

$$r(x) = f(x) - H_{2N+1}(x) = \frac{f^{(2N+2)}(\xi)}{(2N+2)!} p_{N+1}^2(x).$$

Reszta kwadratury:

$$R(f) = I(f) - Q(f) = \int_a^b f(x)dx - Q(f) = \int_a^b H_{2N+1}(x)dx + \int_a^b r(x)dx - Q(f)$$

$f(x_i) = H_{2N+1}(x_i)$

$$= Q(H_{2N+1}) + \int_a^b r(x)dx - Q(H_{2N+1}) = \int_a^b \frac{f^{(2N+2)}(\xi)}{(2N+2)!} p_{N+1}^2(x)dx$$

ξ *zależy od* x

$$= \frac{f^{(2N+2)}(\eta)}{(2N+2)!} \int_a^b p_{N+1}^2(x)dx.$$

□

6.4 Zbieżność ciągu kwadratur

Twierdzenie 6.17. *(Banacha-Steinhausa)*
Niech $(E_1, \|\cdot\|_1)$ *będzie przestrzenią Banacha,* $(E_2, \|\cdot\|_2)$ *– przestrzenią unormowaną,*

$t_\alpha : E_1 \to E_2$ *($\alpha \in \mathcal{A}$) rodziną operatorów liniowych i ciągłych. Jeśli dla dowolnego* $x \in E_1$ *istnieje* m_x *takia, że*

$$\|t_\alpha(x)\|_2 \le m_x \ \forall\ \alpha \in \mathcal{A},$$

to istnieje m *takie, że*

$$\|t_\alpha(x)\|_2 \le m\|x\|_1 \ \forall\ \alpha \in \mathcal{A},\ \forall\ x \in E_1,$$

to znaczy

$$\|t_\alpha\| \le m \ \forall\ \alpha \in \mathcal{A}.$$

Twierdzenie 6.18. *Jeżeli* $(E_1, \|\cdot\|_1)$ *jest przestrzenią Banacha,* $(E_2, \|\cdot\|_2)$ *przestrzenią unormowaną,* $l_n : E_1 \to E_2,\ n \ge 1$ *rodziną operatorów liniowych i ciągłych,* $l : E_1 \to E_2$ *operatorem liniowym i ciągłym, a* $F \subseteq E_1$ *jest zbiorem gęstym, to warunkiem koniecznym i dostatecznym na to, by*

$$\lim_{n\to\infty} l_n(x) = l(x),\ \forall x \in E_1$$

jest, aby spełnione były następujące warunki

(i) $\lim_{n\to\infty} l_n(x) = l(x),\ \forall x \in F,$

(ii) $\exists m \in \mathbb{R} :\ \|l_n\| \le m$ *dla* $n \ge 1.$

Dowód. Dostateczność. Niech $x \in E_1$ i $\varepsilon > 0$ będą dowolne. Wybieramy $y \in F$ taki, by $\|y - x\|_1 < \varepsilon$. Wówczas

$$\begin{aligned}
\|l_n(x) - l(x)\|_2 &\le \|l_n(x) - l_n(y)\|_2 + \|l_n(y) - l(y)\|_2 + \|l(y) - l(x)\|_2 \\
&= \|l_n(x - y)\|_2 + \|l_n(y) - l(y)\|_2 + \|l(y - x)\|_2, \\
\|l(x - y)\|_2 &\le \|l\| \cdot \|x - y\|_1 < \varepsilon\|l\|, \\
\textit{(ii)} &\Rightarrow \|l_n(x - y)\|_2 \le m\|x - y\|_1 < \varepsilon m, \\
\textit{(i)} &\Rightarrow \exists n_0 \forall n \ge n_0 :\ \|l_n(y) - l(y)\|_2 < \varepsilon,
\end{aligned}$$

zatem, dla $n \ge n_0$

$$\|l_n(x) - l(x)\|_2 \le (m + \|l\| + 1)\varepsilon$$

Konieczność. Oczywiście warunek *(i)* jest spełniony, chcemy wykazać *(ii)*. Skoro

$$\lim_{n\to\infty} l_n(x) = l(x),\ \forall\ x \in E_1,$$

to $\{l_n(x)\}_n$ jest ograniczony dla dowolnego $x \in E_1$, zatem

$$\forall x \in E_1 \exists m_x :\ \|l_n(x)\|_2 \le m_x,\ n \ge 1.$$

Na mocy twierdzenia Banacha-Steinhausa

$$\exists m :\ \|l_n\| \le m,\ n \ge 1,$$

co kończy dowód. □

Niech

$$Q_n(f) = \sum_{i=0}^{n} A_i^{(n)} f(x_i^{(n)}),\ n = 1, 2, \dots$$

Twierdzenie 6.19. *(warunek konieczny i wystarczający zbieżności ciągu kwadratur) Dany niech będzie przedział zwarty* $[a, b]$*. Wówczas następujące warunki są równoważne:*

Π_∞ – *wielomiany dowolnego stopnia*

(i) $\lim_{n\to\infty} Q_n(f) = I(f),\ f \in \mathcal{C}([a, b], \mathbb{R})$

(ii) $(Q_n(f) \underset{n\to\infty}{\to} I(f)\ \forall f \in \Pi_\infty) \wedge \left(\exists K > 0 :\ \sum_{i=0}^{n} |A_i^{(n)}| \le K,\ n = 1, 2, \dots\right)$

Dowód. Sprawdzimy, że spełnione są warunki poprzedniego twierdzenia. Przyjmijmy $(E_1, \|\cdot\|_1) = (\mathcal{C}([a, b], \mathbb{R}), \|\cdot\|_c)$ z normą supremową, $(E_2, \|\cdot\|) = (\mathbb{R}, |\cdot|)$, $F = \Pi_\infty$ (przestrzeń wielomianów). Oczywiście $F \subset E_1$ i jest w nim gęsta (dzięki twierdzeniu Weierstrassa o aproksymacji funkcji ciągłych wielomianami). Przyjmijmy jeszcze $l_n = Q_n$ oraz $l = I$. Wystarczy sprawdzić, czy $\|Q_n\| \le \sum_{i=0}^{n} |A_i^{(n)}|$. Istotnie

$$|Q_n(f)| \le \sum_{i=0}^{n} |A_i^{(n)}| \cdot |f(x_i^{(n)})| \le \sum_{i=0}^{n} |A_i^{(n)}| \cdot \|f\|_c \le K\|f\|_c.$$

□

Twierdzenie 6.20. *Jeśli* $A_i^{(n)} \ge 0$ *dla* $n \ge 1$*, to następujące dwa warunki są równoważne*

(i) $\forall\ f \in \mathcal{C}([a, b], \mathbb{R}) : \lim_{n\to\infty} Q_n(f) = I(f)$,

(ii) $\forall\ g \in \Pi_\infty : \lim_{n\to\infty} Q_n(g) = I(g)$.

Dowód. Wystarczy wykazać implikację *(ii)* $\Rightarrow$ *(i)*, a więc że istnieje stała $K > 0$ taka, że $\sum_{i=0}^{n} |A_i^{(n)}| \le K$. Dla dowolnego n

$$Q_n(1) = I(1) = \int_a^b dx < \infty,$$

zatem ciąg $\{Q_n\}_n$ jest ograniczony od góry, czyli

$$\exists K : |Q_n(1)| \le K.$$

Skoro

$$Q_n(1) = \sum_{i=0}^{n} A_i^{(n)},$$

to

$$|Q_n(1)| = \Big|\sum_{i=0}^{n} A_i^{(n)}\Big| \overset{A_i^{(n)} \ge 0}{=} \sum_{i=0}^{n} A_i^{(n)},$$

więc

$$\sum_{i=0}^{n} A_i^{(n)} = \sum_{i=0}^{n} |A_i^{(n)}| \leq K.$$

Na mocy poprzedniego twierdzenia otrzymujemy tezę. □

Twierdzenie 6.21. *Ciąg kwadratur Gaussa jest zbieżny.*

Dowód. Twierdzenia 6.15 + 6.20. □

7 Rozwiązywanie równań nieliniowych.

Dana niech będzie funkcja $f : \mathbb{R} \to \mathbb{R}$ ciągła. Szukamy rozwiązania

(98) $f(x) = 0.$

Przez $\bar{x}$ oznaczamy rozwiązanie tego zagadnienia. Jeżeli zajdzie potrzeba, to zakładamy, że f jest odpowiednio regularna ($\mathcal{C}^1$ albo $\mathcal{C}^2$).

Będziemy chcieli znaleźć ciąg $\{x_k\}_k \subset \mathbb{R}$ taki, że

$$\lim_{k\to\infty} x_k = \bar{x}.$$

Dane niech będzie przybliżenie początkowe $x_0, \ldots, x_p \in \mathbb{R}$ i niech

(99) $x_{k+1} = F_k(x_k, x_{k-1}, \ldots, x_{k-p}),\ k = p, p+1, \ldots$

gdzie F_k jest zadaną funkcją określającą metodę iteracyjną ($(p+1)$-krokową, niestacjonarną).

Definicja 7.1. *Metodę nazywamy metodą* ***zbieżną globalnie*** *jeśli ciąg (99) jest zbieżny do $\bar{x}$ dla dowolnego wyboru wartości początkowych. Metoda jest* ***zbieżna lokalnie****, jeśli istnieje $\rho > 0$ takie, że ciąg (99) jest zbieżny do $\bar{x}$ dla dowolnych wartości początkowych $x_0, \ldots, x_p \in K(\bar{x}, \rho)$. Kulę $K(\bar{x}, \rho)$ nazywamy* ***kulą zbieżności*** *danej metody dla funkcji f.*

Przykład 7.2. Rozważmy układ równań

$$\begin{cases} y = ax^2 + bx + c \\ x = \alpha y^2 + \beta y + \gamma \end{cases}$$

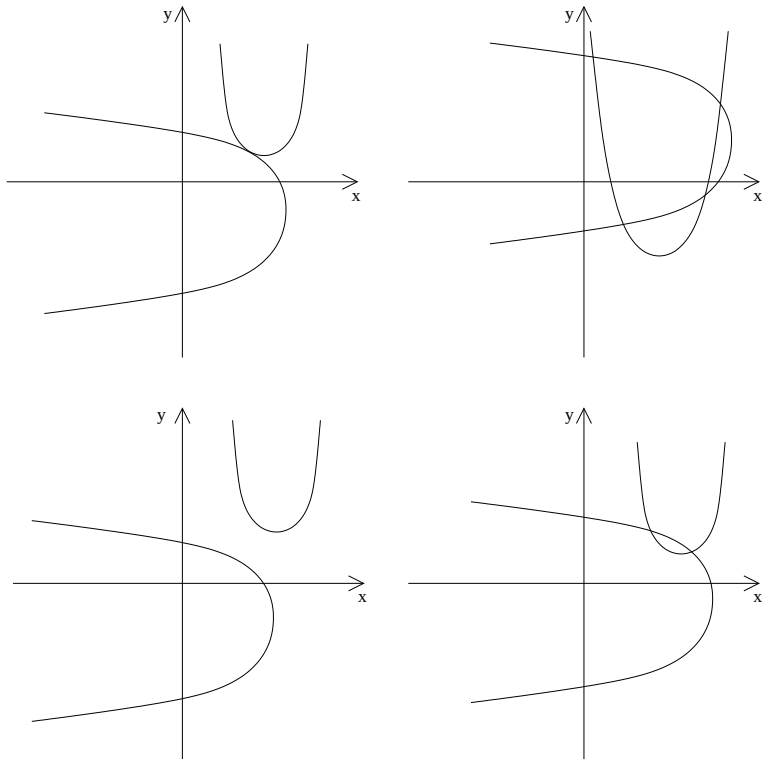
Możliwe rozwiązania dla tego układu przedstawia rysunek 13.

Uwaga 7.3. *W przypadku równań nieliniowych ciąg przybliżeń na ogół szybciej dąży do rozwiązania.*

Rozważmy metodę jednokrokową, stacjonarną

(100) $$x_{k+1} = F(x_k).$$

Niech $e_k = x_k - \bar{x}$ oznacza błąd k-tego przybliżenia rozwiązania dokładnego $\bar{x}$.



Rysunek 13:

Definicja 7.4. ***Wykładnikiem zbieżności*** *(rzędem zbieżności) metody nazywamy największą liczbę* p *spełniającą warunek*

$$\|e_{k+1}\| \leq A\|e_k\|^p, \ A > 0,$$

gdzie $\|\cdot\|$ *jest normą w* $\mathbb{R}^n$.

Jeśli $p = 1$ *to metoda jest zbieżna liniowo (z ilorazem* A*). Jeśli* $p > 1$*, to mówimy, że metoda jest zbieżna ponadliniowo (kwadratowo dla* $p = 2$*).*

Uwaga 7.5. *Przy metodach zbieżnych ponadliniowo błąd przybliżenia szybko maleje.*

Uwaga 7.6. *Jeżeli* $A < 1$ *to metoda zbieżna liniowo jest zbieżna globalnie.*

Twierdzenie 7.7. *Jeśli* $p > 1$ *to metoda (100) jest lokalnie zbieżna dla* $x_0 \in K(\bar{x}, r)$*, gdzie* $r \leq A^{\frac{-1}{p-1}}$.

Dowód.

$$|e_{k+1}| \leq A|e_k|^p \leq A(A|e_{k-1}|^p)^p = A^{p+1}|e_{k-1}|^{p^2} \leq \ldots \leq A^{1+p+\ldots+p^k}|e_0|^{p^{k+1}},$$

$$\sum_{i=0}^{k-1} p^i = \frac{p^k-1}{p-1},$$

$$|e_k| \leq A^{\frac{p^k-1}{p-1}}|e_0|^{p^k-1}|e_0| = (A^{\frac{1}{p-1}}|e_0|)^{p^k-1}|e_0|.$$

Zatem $e_k \underset{k\to\infty}{\to} 0$ o ile

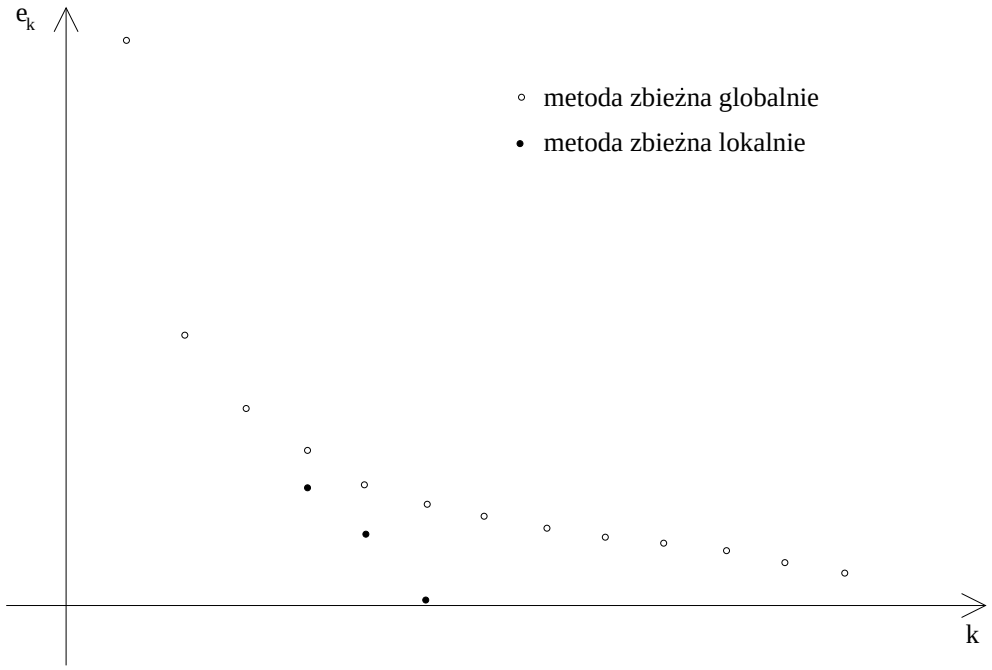
$$A^{\frac{1}{p-1}}|e_0| < 1,$$

a więc dla $|e_0| < A^{\frac{-1}{p-1}}$. Metoda jest więc zbieżna dla

$$|x_0 - \bar{x}| < A^{\frac{-1}{p-1}}.$$

□

Jak widać, aby wyznaczyć kulę zbieżności wymagana jest wiedza na temat $\bar{x}$, a więc na temat rozwiązania którego poszukujemy. Możemy coś o nim wiedzieć *a priori*, ale jak należy postąpić gdy tak nie jest? Jednym ze sposobów jest użycie metody zbieżnej globalnie w pierwszym etapie poszukiwania rozwiązania, a następnie skorzystać z metody zbieżnej lokalnie, która na ogół szybciej zmierza do rozwiązania (rysunek 14).



Rysunek 14:

Inną metodą jest zastąpienie równania (98) problemem znalezienia

$$\min\{f^2(x) : x \in \mathbb{R}\}.$$

7.1 Metoda bisekcji.

Dana niech będzie funkcja ciągła $f : \mathbb{R} \to \mathbb{R}$. Metoda ta opiera się na twierdzeniu Darboux.[15] Mając zadaną funkcję spełniającą założenia tego twierdzenia, przyjmujemy $x_0 = a$, $y_0 = b$. Przyjmijmy też, że $f(a) < 0 < f(b)$. Ciągi $\{x_k\}_k$, $\{y_k\}_k$ tworzymy w następujący sposób:

- przyjmujemy $z_k = \frac{x_k + y_k}{2}$, $k = 0, 1, \ldots$,
- jeśli $f(z_k) < 0$, to $x_{k+1} = z_k$, $y_{k+1} = y_k$,
- jeśli $f(z_k) > 0$, to $x_{k+1} = x_k$, $y_{k+1} = z_k$.

[15] Jeśli $f : \mathbb{R} \to \mathbb{R}$ jest ciągła, $a < b$, $f(a)f(b) < 0$, to istnieje $\xi \in (a, b)$ takie, że $f(\xi) = 0$.

Otrzymane w ten sposób ciągi są monotoniczne i ograniczone:

- ciąg $\{x_k\}_k$ rosnący i ograniczony od góry (przez b),
- ciąg $\{y_k\}_k$ malejący i ograniczony od dołu (przez a),

zatem są one zbieżne. Niech $\tilde{x} = \lim_{k\to\infty} x_k$, $\tilde{y} = \lim_{k\to\infty} y_k$. Zauważmy, że

$$y_{k+1} - x_{k+1} = \begin{cases} z_k - x_k, & f(z_k) > 0 \\ y_k - z_k, & f(z_k) < 0 \end{cases} = \tfrac{1}{2}(y_k - x_k) = \ldots = \left(\tfrac{1}{2}\right)^k (b - a),$$

zatem $\tilde{x} = \tilde{y}$. Ale

$$f(x_k)f(y_k) < 0 \Rightarrow f^2(\tilde{x}) \leq 0 \Rightarrow f(\tilde{x}) = 0,$$

zatem $\tilde{x}$ jest rozwiązaniem (globalnym) problemu (98).

Zauważmy, że w maszynie cyfrowej nie możemy przeprowadzać obliczeń nieskończonych. W związku z tym wprowadzamy kryterium stopu. Wybieramy $\varepsilon > 0$. Obliczenia przerywamy w dwóch przypadkach

- ▷ gdy $f(z_k) < \varepsilon$, wówczas przyjmujemy $\bar{x} = z_k$,
- ▷ gdy $y_k - x_k < \varepsilon$, wówczas przyjmujemy $\bar{x} = \frac{y_k - x_k}{2}$.

Uwaga 7.8. *Metoda bisekcji jest metodą niestacjonarną, wybór* $x_{k+1} = z_k$ *lub* $y_{k+1} = z_k$ *zależy od* k.

7.2 Kontrakcje.

Załóżmy teraz, że daną mamy metodę stacjonarną jednokrokową F oraz ciąg $\{x_k\}_k$

$$x_{k+1} = F(x_k),$$

zbieżny do rozwiązania $\bar{x}$. Wówczas

$$\bar{x} = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} F(x_k) = F(\bar{x}).$$

Zatem, jeśli $\bar{x}$ jest rozwiązaniem (98), to jest on punktem stałym metody F. Zastanówmy się jakie są warunki istnienia punktu stałego.

Definicja 7.9. *Mówimy, że funkcja* $f : \mathbb{R}^n \to \mathbb{R}^n$ *spełnia warunek Lipschitza ze stałą* L*, gdy*

$$\forall\, x, y \in \mathbb{R}^n : \ \|f(x) - f(y)\| \leq L\|x - y\|.$$

Jeśli $0 \leq L < 1$ *to funkcję* f *nazywamy kontrakcją lub odwzorowaniem zwężającym.*

Uwaga 7.10. *Jeżeli funkcja* f *spełnia warunek Lipschitza, to jest ciągła.*

Twierdzenie 7.11. *(Banacha o kontrakcjach)*
Niech $D \subset \mathbb{R}^n$ *będzie zbiorem domkniętym. Jeśli* $F : D \to D$ *jest kontrakcją w* D*, to istnieje dokładnie jeden punkt stały* $\bar{x} \in D$ *funkcji* F.

Dowód. Dla $x_0 \in D$ definiujemy ciąg

$$x_{k+1} = F(x_k),\ k = 0, 1, \ldots,\ \{x_k\}_k \subset D$$

Wówczas

$$\|x_{k+1} - x_k\| = \|F(x_k) - F(x_{k-1})\| \leq L\|x_k - x_{k-1}\| \leq \ldots \leq L^k\|x_1 - x_0\|,$$

więc ciąg $\{x_k\}_k$ spełnia warunek Cauchy'ego:

$$\forall \varepsilon > 0\ \exists k_0\ \forall l > k \geq k_0 :\ \|x_l - x_k\| < \varepsilon,$$

bo

$$\|x_l - x_k\| \leq \|x_l - x_{m-1}\| + \ldots + \|x_{k+1} - x_k\| \leq \sum_{j=k}^{l-1} L^j\|x_1 - x_0\|.$$

Skoro $D \subseteq \mathbb{R}^n$ jest domknięty, to dzięki zupełności $\mathbb{R}^n$ zbiór D jest zupełny. Istnieje więc $\bar{x} \in D$ taki, że

Bo F – kontrakcja.

$$\bar{x} = \lim_{k\to\infty} x_{k+1} = \lim_{k\to\infty} F(x_k) = F(\bar{x}).$$

Pokażemy jeszcze jedyność punktu $\bar{x}$. Dla dowodu nie wprost załóżmy, że istnieją dwa różne punkty stałe $\bar{x}, \bar{\bar{x}} \in D$. Wówczas

$$\|\bar{x} - \bar{\bar{x}}\| = \|F(\bar{x}) - F(\bar{\bar{x}})\| = L\|\bar{x} - \bar{\bar{x}}\|.$$

Skoro $L < 1$ to $\bar{x} = \bar{\bar{x}}$ – sprzeczność. □

Twierdzenie 7.12. *(warunek wystarczający istnienia punktu stałego)*
Jeśli F *jest kontrakcją w kuli* $\overline{K}(x_0, r)$ *ze stałą* $L \in [0, 1)$ *oraz* $\|x_0 - F(x_0)\| \leq r(1 - L)$, *to* F *ma dokładnie jeden punkt stały w* $\overline{K}(x_0, r)$.

Dowód. Na mocy twierdzenia Banacha o kontrakcjach wystarczy wykazać, że $F(\overline{K}(x_0, r)) \subseteq \overline{K}(x_0, r)$. Niech $x \in \overline{K}(x_0, r)$, wówczas

$$\|x_0 - F(x)\| \leq \|x_0 - F(x_0)\| + \|F(x_0) - F(x)\| \leq r(1-L) + L\|x_0 - x\| \leq r(1-L) + Lr = r,$$

czyli $F(x) \in \overline{K}(x_0, r)$. □

Twierdzenie 7.13. *Jeżeli* F *jest kontrakcją na* $\overline{K}(\bar{x}, r)$, *gdzie* $F(\bar{x}) = \bar{x}$, *to ciąg* $\{x_k\}_k$ *określony wzorem*

$$x_{k+1} = F(x_k),\ k \geq 0,$$

jest zbieżny do $\bar{x}$ *dla dowolnego* $x_0 \in \overline{K}(\bar{x}, r)$.

Dowód. Skoro F jest kontrakcją na $\overline{K}(\bar{x}, r)$, to istnieje stała $L < 1$ taka, że

$$\|F(x) - F(y)\| \leq L\|x - y\|, \ \forall x, y \in \overline{K}(\bar{x}, r).$$

Ustalmy dowolny punkt $x_0 \in \overline{K}(\bar{x}, r)$. Wówczas

$L < 1$

$$\|x_1 - \bar{x}\| = \|F(x_0) - F(\bar{x})\| < \|x_0 - \bar{x}\| < r.$$

Zatem $\{x_k\}_k \subset \overline{K}(\bar{x}, r)$ oraz

$$\|x_{k+1} - \bar{x}\| = \|F(x_k) - F(\bar{x})\| \leq L\|x_k - \bar{x}\| \leq \ldots \leq L^{k+1}\|x_0 - \bar{x}\| \underset{k\to\infty}{\to} 0,$$

co należało dowieść. □

Lemat 7.14. *Jeżeli* F *jest klasy* $\mathcal{C}^1$ *oraz*

$$|F'(x)| \leq M < 1, \ x \in [a, b],$$

to F *jest kontrakcją w* $[a, b]$.

Dowód. Niech $x, y \in [a, b]$. Na mocy twierdzenia o wartości średniej, istnieje $\xi \in I(x, y)$ taki, że

$$|F(x) - F(y)| = |f'(\xi)| \cdot |x - y| \leq M|x - y|.$$

□

7.3 Metoda siecznych

Ustalmy dwa punkty $x_{k-1}, x_k \in \mathbb{R}$. Szukać będziemy rozwiązania zagadnienia

(101)
$$P_{k-1,k}(x) = 0,$$

gdzie $P_{k-1,k}$ jest wielomianem interpolacyjnym w postaci Lagrange'a dla funkcji f opartym o węzły x_{k-1}, x_k. Wielomian ten jest oczywiście prostą przechodzącą przez punkty $(x_{k-1}, f(x_{k-1}))$ i $(x_k, f(x_k))$. Jako x_{k+1} bierzemy punkt przecięcia się tej prostej z osią x (jak pokazano na rysunku 15).

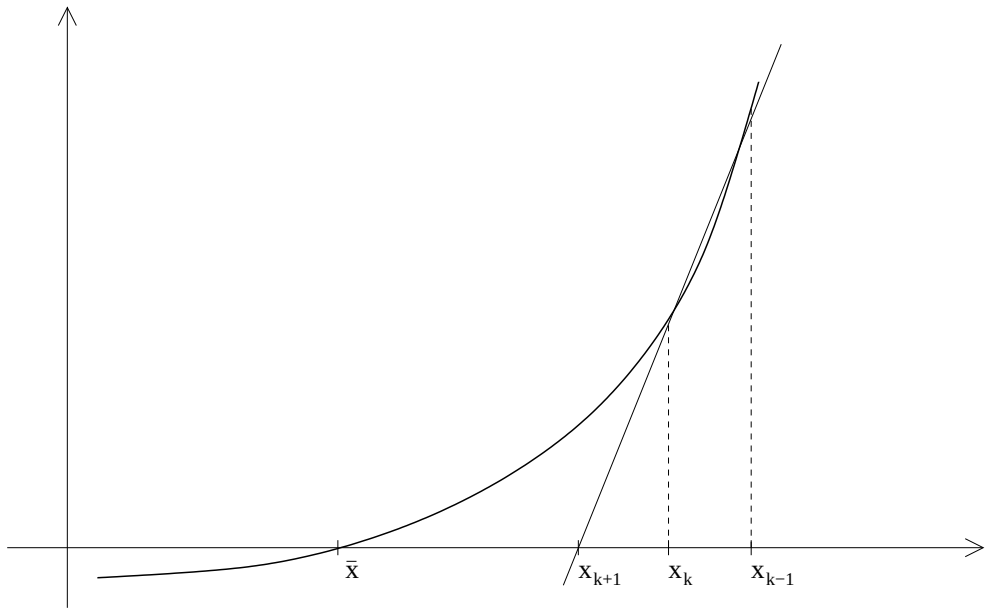
Oczywiście

$$P_{k-1,k}(x) = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}(x - x_k) + f(x_k),$$

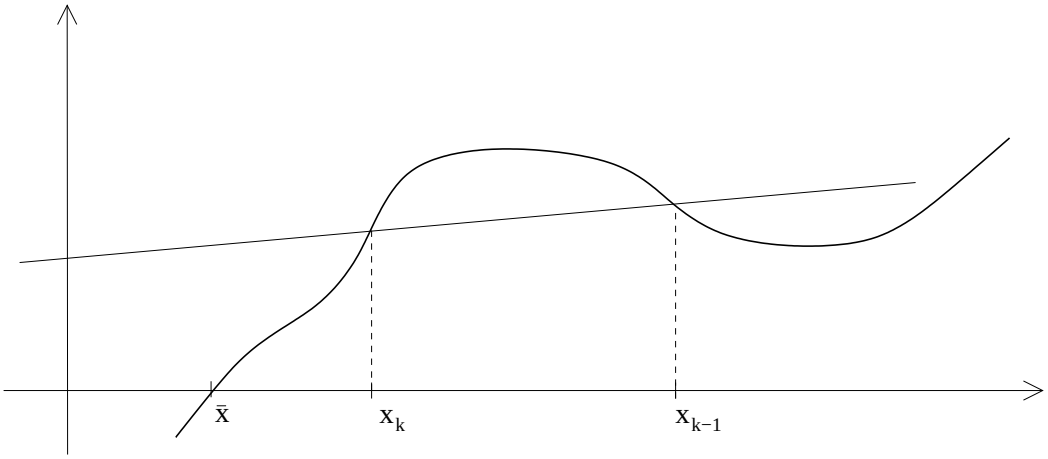
zatem

(102)
$$x_{k+1} = x_k - f(x_k)\frac{x_k - x_{k-1}}{f(x_k) - f(x_{k-1})}.$$

Widać więc, że metoda ta wymaga założenia, aby $f(x_k) \neq f(x_{k-1})$. Założyć należy też, że rozwiązanie $\bar{x}$ jest miejscem zerowym pojedynczym, to znaczy $f'(\bar{x}) \neq 0$.



Rysunek 15: Metoda siecznych



Rysunek 16: Metoda siecznych

Uwaga 7.15. *Metoda nie jest optymalna, gdy punkty* x_k *i* x_{k-1} *wybrane zostały tak, że* $f(x_k) \approx f(x_{k-1})$. *Sytuację taką przedstawia rysunek 16*

Zajmiemy się teraz błędem przybliżenia dla tej metody. Niech $e_k = x_k - \bar{x}$. Z (102) wynika, że

Bez dowodu

(103)
$$e_{k+1} = \frac{f''(\bar{x})}{2f'(\bar{x})}(1 + \mathcal{O}(e_{k-1} + e_k))e_k e_{k-1}.$$

Wynika z niego, że

(104)
$$|e_{k+1}| \approx B|e_k||e_{k-1}|, \; B = \left|\tfrac{f''(\bar{x})}{2f'(\bar{x})}\right|.$$

Mnożąc obustronnie ostatnie równanie przez B i przyjmując $u_k = \ln B|e_k|$ otrzymujemy

Ciąg Fibonacciego

$$u_{k+1} = u_k + u_{k-1},\ k = 1, 2, \ldots,$$

czyli równanie różnicowe

(105)
$$u_{k+1} - u_k - u_{k-1} = 0,\ k \geq 1.$$

Rozwiązanie tego równania jest postaci

(106)
$$u_k = c\lambda_1^k + d\lambda_2^k,$$

gdzie λ_1, λ_2 są pierwiastkami równania

$$\lambda^2 - \lambda - 1 = 0,$$

a c, d stałymi (zależnymi od warunku początkowego u_0).

Dowód wzoru (106). Oczywiście, jeśli w_k i z_k są rozwiązaniami (105), to ich kombinacja też. Niech więc $\lambda \neq 0$ i niech

$$w_k = c\lambda^k.$$

Wstawiając tą wartość do (105) mamy

$$c\lambda^{k+1} - c\lambda^k - c\lambda^{k-1} = 0,$$
$$c\lambda^{k-1}(\lambda^2 - \lambda - 1) = 0.$$

□

Rozwiązaniami $\lambda^2 - \lambda - 1 = 0$ są

$$\lambda_1 = \tfrac{1+\sqrt{5}}{2}$$
$$\lambda_2 = \tfrac{1-\sqrt{5}}{2}$$

i skoro $|\lambda_2| < 1$ to

$$u_k \approx c\lambda_1^k.$$

Otrzymujemy więc

$$u_k = \ln B|e_k|,$$
$$c\lambda_1^k = \ln B|e_k|,$$
$$B|e_k| = \exp(c\lambda_1^k),$$
$$B|e_{k+1}| = (\exp(c\lambda_1^k))^{\lambda_1} = (B|e_k|)^{\lambda_1},$$

ostatecznie

$$|e_{k+1}| \approx B^{\lambda_1 - 1}|e_k|^{\lambda_1}.$$

Zatem wykładnik zbieżności metody siecznych

$$p = \lambda_1 = \tfrac{1+\sqrt{5}}{2} > 1,$$

czyli metoda jest zbieżna ponadliniowo.

7.4 Metoda „reguła falsi".

Metoda ta jest modyfikacją poprzedniej. Przyjmujemy, że $x_{k-1} = a$ jest ustalone. Wówczas iteracja (102) przyjmuje postać

(107)
$$x_{k+1} = x_k - f(x_k)\frac{x_k - a}{f(x_k) - f(a)},\ f(a) \neq f(x_k).$$

Jest to metoda 1-krokowa, możemy więc przypuszczać, że jest zbieżna słabiej od poprzedniej (która była 2-krokowa). Niech f będzie klasy $\mathcal{C}^2$ i niech

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

wówczas (107) równoważne jest

(108)
$$x_{k+1} = F(x_k).$$

Zatem

$$F(x) = \frac{x(f(x) - f(a)) - f(x)(x - a)}{f(x) - f(a)} = \frac{af(x) - xf(a)}{f(x) - f(a)}.$$

Załóżmy, podobnie jak przy metodzie siecznych, że $f'(\bar{x}) \neq 0$, wówczas

$$F'(x) = \frac{(af'(x) - f(a))(f(x) - f(a)) - f'(x)(af(x) - xf(a))}{(f(x) - f(a))^2},$$
$$F'(\bar{x}) = \frac{f(a) - f'(\bar{x})(a - \bar{x})}{f(a)}.$$

Ze wzoru Taylora

$$f(a) = f(\bar{x}) + f'(\bar{x})(a - \bar{x}) + \tfrac{1}{2}f''(\xi)(a - \bar{x})^2,\ \xi \in I(a, \bar{x}),$$
$$f(a) - f'(\bar{x})(a - \bar{x}) = \tfrac{1}{2}f''(\xi)(a - \bar{x})^2,$$

zatem

$$F'(\bar{x}) = \frac{\frac{1}{2}f''(\xi)(a - \bar{x})^2}{f(a)} \overset{f(\bar{x})=0}{=} -\frac{1}{2}\frac{f''(\xi)}{\frac{f(\bar{x})-f(a)}{\bar{x}-a}}(\bar{x} - a).$$

Dla $|a - \bar{x}|$ dostatecznie małych

$$|F'(\bar{x})| \overset{\xi\in I(\bar{x},a)}{\approx} \frac{|f''(\bar{x})|}{2|f'(\bar{x})|}|\bar{x} - a|,$$

bo $\frac{f(\bar{x})-f(a)}{\bar{x}-a} \approx f'(\bar{x})$. Skoro $f'(\bar{x}) \neq 0$, to

$$|f'(x)| > 0,\ \text{dla } |\bar{x} - x| \text{ małych.}$$

Więc

$$|F'(\bar{x})| \approx C|\bar{x} - a|,\ C = \tfrac{1}{2}\left|\tfrac{f''(\bar{x})}{f'(\bar{x})}\right|,$$

czyli $|F'(\bar{x})| < 1$ dla $|\bar{x} - a|$ małych. Ponadto F' jest ciągła (bo f jest klasy $\mathcal{C}^2$), zatem F jest kontrakcją. Na mocy twierdzenia 7.13

$$x_k \underset{k\to\infty}{\to} \bar{x},$$

o ile $x_0 \in \overline{K}(\bar{x}, r)$ dla pewnego promienia zbieżności r. Wykazaliśmy więc twierdzenie

Twierdzenie 7.16. *Jeśli $f'(\bar{x}) \neq 0$ oraz f jest klasy $\mathcal{C}^2$, to istnieje $r > 0$ takie, że dla dowolnych $a, x_0 \in \overline{K}(\bar{x}, r)$ ciąg (107) jest zbieżny do rozwiązania $\bar{x}$ zagadnienia (98).*

7.5 Metoda stycznych (Newtona).

Niech $H_1(x)$ będzie wielomianem interpolacyjnym w postaci Hermite'a dla funkcji f opartym na dwukrotnym węźle x_k. Przez x_{k+1} oznaczmy rozwiązanie problemu

(109)
$$H_1(x) = 0.$$

Na mocy definicji wielomianu interpolacyjnego Hermite'a

$$H_1(x) = b_0 + b_1(x - x_k),$$
$$H_1'(x) = b_1.$$

Na mocy warunków interpolacyjnych

$$H_1(x_k) = f(x_k),$$
$$H_1'(x_k) = f'(x_k),$$

zatem

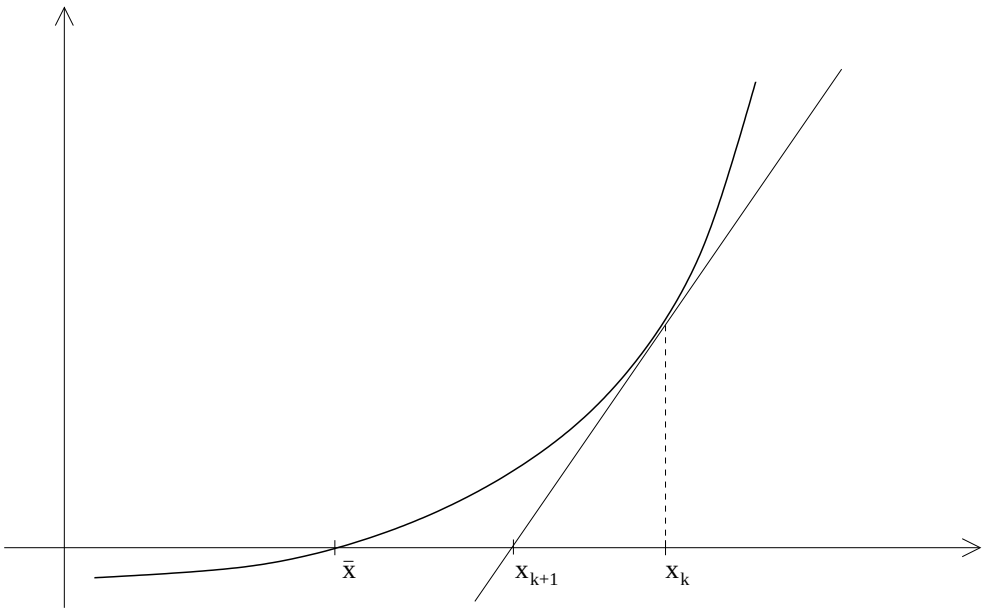
$$H_1(x) = f(x) + f'(x_k)(x - x_k).$$

Skoro x_{k+1} ma być miejscem zerowym H_1, to

(110)
$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)},\ f'(x_k) \neq 0.$$

Jak widać $y = H_1(x)$ jest równaniem stycznej do wykresu $y = f(x)$ w punkcie $(x_k, f(x_k))$ (rysunek 17)

Twierdzenie 7.17. *Jeśli f jest klasy $\mathcal{C}^2$ oraz $\bar{x}$ jest jej miejscem zerowym pojedynczym, to istnieje $r > 0$ taka, że dla dowolnego $x_0 \in \overline{K}(\bar{x}, r)$ ciąg (110) jest zbieżny do $\bar{x}$.*



Rysunek 17: Metoda stycznych

Dowód. Niech

$$F(x) = x - \frac{f(x)}{f'(x)},$$

wówczas iteracja (110) równoważna jest

$$x_{k+1} = F(x_k).$$

Oczywiście

$$F'(x) = 1 - \frac{(f'(x))^2 - f(x)f''(x)}{(f'(x))^2} = \frac{f(x)f''(x)}{(f'(x))^2},$$
$$F'(\bar{x}) = 0.$$

Skoro F' jest ciągłe, to istnieje $r > 0$ takie, że

$$|F'(x)| < 1, \text{ dla } |x - \bar{x}| \leq r,$$

zatem F jest kontrakcją w kuli $K(\bar{x}, r)$. Twierdzenie 7.13 kończy dowód. □

Określmy jeszcze rząd tej metody. Niech $e_k = x_k - \bar{x}$. Wówczas, na mocy (110)

(111)
$$e_{k+1} = e_k - \frac{f(x_k)}{f'(x_k)}.$$

Ze wzoru Taylora

$$0 = f(\bar{x}) = f(x_k) + f'(x_k)(\bar{x} - x_k) + \frac{1}{2}f''(\xi)(\bar{x} - x_k)^2, \ \xi \in I(\bar{x}, x_k),$$
$$x_k - \bar{x} = \frac{f(x_k)}{f'(x_k)} + \frac{f''(\xi)}{2f'(x_k)}(\bar{x} - x_k)^2,$$
$$e_k = x_k - \bar{x} = \frac{f(x_k)}{f'(x_k)} + \frac{f''(\xi)}{2f'(x_k)}e_k^2,$$

zatem, z (110)

$$e_{k+1} = \frac{f''(x_k)}{2f'(x_k)} e_k^2,$$

więc metoda stycznych jest rzędu 2.

Index