Krzysztof Moszyński

METODY NUMERYCZNE DLA INFORMATYKÓW

Rok akademicki 2004/2005

Rozdział 1
APROKSYMACJA.

Ogólne zagadnienie aproksymacji w przestrzeni liniowej

$(X, \|\cdot\|)$ -przestrzeń liniowa unormowana, P -podzbiór przestrzeni X.

- Dla $x \in X$ poszukujemy elementu $p \in P$ takiego, że $\|x - p\|$ jest *wystarczająco małe*: p aproksymuje x.
- Dla $x \in X$ poszukujemy elementu $p \in P$ takiego, że
$$\forall q \in P \quad \|p - x\| \le \|q - x\|;$$
p nazywa się wtedy *elementem najlepszej aproksymacji* $x \in X$ *przez elementy podzbioru* P.

Własności elementu aproksymującego (w szczególności elementu najlepszej aproksymacji) zależą od X, P i $\|\cdot\|$. Dla tego, jeśli mówimy o aproksymacji, to musimy być świadomi tego

- **skąd bierzemy element aproksymowany (x),**
- **gdzie szukamy elementu aproksymującego (p),**
- **w jaki sposób mierzymy jakość aproksymacji ($\|\cdot\|$).**

Istnienie elementu najlepszej aproksymacji

Twierdzenie 1.1

- $(X, \|\cdot\|)$ *- przestrzeń liniowa unormowana,*
- $P \subset X$ *- podprzestrzeń skończonego wymiaru.*

Wtedy, dla każdego $x \in X$ *istnieje element* $p \in P$*, najlepszej aproksymacji dla* x*.*

Dowód

1. Jeśli $x \in P$, to bierzemy $p = x$.

2. Jeśli $x \notin P$, to $\rho(x,P) = \inf_{q\in P} \|x-q\| = r > 0$, gdyż P jest skończonego wymiaru. Niech $Q = P \cap \{q \in P \quad | \quad \|q-x\| \le r+\epsilon\}$, gdzie $\epsilon > 0$ jest ustaloną liczbą. Wtedy Q jest zbiorem *zwartym* (dlaczego?). Połóżmy $f(q) = \|q-x\|$ dla $q \in Q$; funkcja f jest ciągła i jest określona na zbiorze zwartym Q, a więc na Q osiąga swój kres dolny. To znaczy, że istnieje $p \in Q$ spełniające warunek $\|p-x\| = f(p) = \inf_{q\in Q} f(q)$. To oznacza, że p jest elementem najlepszej aproksymacji dla x. $\Box$

W sytuacji, o której mówi Twierdzenie 1.1, element najlepszej aproksymacji dla $x \in X$ może być jedyny lub nie, w zależności od własności normy $\|\cdot\|$.

Przykład
Niech $X = \mathbf{R}^2 = \{(\xi_1,\xi_2) \quad | \quad \xi_1,\xi_2 \in \mathbf{R}\}$; $P = \{(\xi_1,\xi_2)| \quad \xi_2 = 0\}$, $x = (0,1)$.

1. Jeśli w przestrzeni X przyjmiemy normę *euklidesową*, $\|y\| = \sqrt{(\xi_1^2+\xi_2^2)}$ dla $y = (\xi_1,\xi_2)$, to *jedynym* elementem najlepszej aproksymacji dla x będzie $p = (1,0)$.

2. Jeśli zaś normę określimy tak: $\|y\| = \max\{|\xi_1|,|\xi_2|\}$, to zbiorem wszystkich elementów najlepszej aproksymacji dla x w P będzie odcinek otwarty $((-1,0),(1,0))$.

3. Jeśli (na przykład przy definicji normy z punktu 1.), jako zbiór P przyjmiemy
$$P = \{(\xi_1,\xi_2) \quad | \quad \xi_2 < 1\},$$
to okaże się, że **w P nie ma elementu najlepszej aproksymacji dla x.** (Dlaczego?).$\Box$

Obiektami, które najczęściej musimy aproksymować są *funkcje*. Chodzi nam zwykle o to, abyśmy mogli zastąpić funkcję

bardzo skomplikowaną

lub

taką, o której wiemy zbyt mało

przez inną funkcję, z którą łatwo potrafimy sobie radzić. Takimi stosunkowo łatwymi funkcjami są, na przykład, *wielomiany*. Ich wartości potrafimy łatwo obliczać (patrz - ćwiczenia: *schemat Hornera*).

Najczęściej będą nas interesować *funkcje ciągłe* określone na pewnym ustalonym zbiorze zwartym $\Omega \in \mathbf{R}^d$, mające wartości rzeczywiste. (Gdy $d = 1$, najczęściej będzie $\Omega = [a, b]$.) Niech więc naszym zbiorem X będzie *zbiór wszystkich funkcji ciągłych określonych na* Ω. W X łatwo określimy, w sposób naturalny, operację + - dodawania elementów, oraz operację mnożenia ich przez liczby. W ten sposób w zbiorze X zbudujemy strukturę *przestrzeni liniowej*. Mamy już *przestrzeń liniową* X. Jeśli Ω jest zbiorem o nieskończonej mocy, to wymiar (algebraiczny) X jest **nieskończony**.

W naszej przestrzeni liniowej X możemy teraz określić *normę* na różne sposoby. Nasza przestrzeń X stanie się w ten sposób *przestrzenią liniową unormowaną*.

Najczęściej w X używa się *normy "sup"*; dla $f \in X$

$$\|f\|_{\infty,\Omega} = \sup_{t\in\Omega} |f(t)|.$$

Jeśli nie będzie wątpliwości co do zbioru Ω, będziemy pisać krócej $\|f\|_\infty$. Zbieżność w sensie normy $\|\cdot\|_{\infty,\Omega}$, to *zbieżność jednostajna w* Ω. Inną normą, z którą będziemy mieć do czynienia to *norma* $L^2(\Omega)$

$$\|f\|_2 = (\int_\Omega |f(t)|^2 d\Omega)^{\frac{1}{2}}.$$

Aproksymacja w sensie każdej z tych norm ma inne własności.

INTERPOLACJA LAGRANGE'A

Niech X będzie przestrzenią liniową wszystkich funkcji ciągłych, określonych na skończonym przedziale domkniętym $[a, b] \subset \mathbf{R}$; niech P będzie zbiorem wszystkich wielomianów jednej zminnej rzeczywistej. Szczególnym rodzajem aproksymacji elementów przestrzeni X przez elementy *jej podprzestrzeni* P jest *interpolacja w sensie Lagrange'a*

(1.1) Zadanie interpolacji wielomianowej, globalnej w sensie Lagrange'a

W przedziale $[a, b]$ *dany jest układ* $n+1$ *różnych punktów zwanych* **węzłami***:*

$$a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b.$$

Dla $f \in X$ poszukuemy wielomianu $P_n \in P$, stopnia $\leq n$, o tej własności, że

$$f(x_j) = P_n(x_j) \ \textit{dla} \ j = 0, 1, 2, \cdots, n.$$

Wielomian P_n spełniający powyższe warunki to **wielonian interpolacyjny Lagrange'a dla funkcji f, i węzłów $x_0, x_1, \cdots, x_n$.**

Ten sposób aproksymacji pozwala *prybliżać* przy pomocy wielomianu P_n stopnia $\leq n$ dowolną funkcję (nawet nie koniecznie ciągłą!), określoną jedynie w zadanych węzłach. Funkcję f, której wartości znamy jedynie w węzłach wymienionych w sformułowaniu zadania (1.1), (mogą to być na przykład wielkości otrzymane z pomiarów eksperymentalnych), zastępujemy wielomianem P_n.

Wielomian interpolacyjny Lagrange'a **nie jest na ogół elementem najlepszej aproksymacji!**.

Twierdzenie 1.2
Zadanie interpolacji Lagrange'a (1.1) ma jednoznaczne rozwiązanie

Dowód
1. Istnienie. Podamy konstrukcję rozwiązania, używając tak zwanych *wielomianów bazowych Lagrange'a*, związanych z węzłami $x_0, x_1, \cdots, x_n$. Każdemu węzłowi przyporządkowany jest wielomian stopnia n:

$$l_j(x) = \frac{(x-x_0)(x-x_1)\cdots(x-x_{j-1})(x-x_{j+1})\cdots(x-x_n)}{(x_j-x_0)(x_j-x_1)\cdots(x_j-x_{j-1})(x_j-x_{j+1})\cdots(x_j-x_n)}, \tag{1.2}$$

dla $j = 0, 1, \cdots, n$. Zauważmy, że

$$l_j(x_k) = \delta_{jk} \text{ dla } j, k = 0, 1, \cdots, n,$$

oraz że każda z funkcji l_j jest wielomianem stopnia n. Stąd natychmiast wynika, że

$$P_n(x) = \sum_{j=0}^{n} f(x_j) l_j(x), \tag{1.3}$$

jest wielomianem stopnia $\leq n$, oraz że

$$P_n(x_k) = \sum_{j=0}^{n} \delta_{jk} f(x_j) = f(x_k),$$

co oznacza, że P_n jest wielomianem interpolacyjnym Lagrange'a, o węzłach $x_0, x_1, \cdots, x_n$ dla funkcji f.

2. Jednoznaczność. Jeśli poszukiwany wielomian P_n zapiszemy w postaci *naturalnej*,

$$P_n(x) = \sum_{j=0}^{n} a_j x^j,$$

to jest w postaci jego rozwinięcia względem bazy wielomianów $1, x, x^2, \cdots, x^n$, to widzimy, że zadanie (1.1) sprowadza się do znalezienia współczynników

$$a_0, a_1, a_2, \cdots, a_n,$$

spełniających *układ* $n+1$ *równań algebraicznych liniowych*

$$\sum_{j=0}^{n} x_k^j a_j = f(x_k) \text{ dla } k = 0, 1, \cdots n. \tag{1.4}$$

Macierzą tego układu jest *macierz Vandermonda*:

$$V = \begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & x_1^3 & \cdots & x_1^n \\ 1 & x_2 & x_2^2 & x_2^3 & \cdots & x_2^n \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & x_n^2 & x_n^3 & \cdots & x_n^n \end{bmatrix}. \tag{1.5}$$

Wiadomo, że macierz taka jest nieosobliwa, jeśli węzły są różne. Zatem układ (1.4) ma jednoznaczne rozwiązanie. □

Zauważmy, że dowód Twierdzenia 1.2 zawiera *dwa różne algorytmy wyznaczania wielomianu* P_n. Jeden z nich określony jest wzorem (1.3), zaś drugi wzorem (1.4). Każdy z tych algorytmów wyznacza ten sam wielomian P_n w postaci rozwinięcia *względem innej bazy podprzestrzeni wielomianów stopnia* $\leq n$.

Chwilowo zwróćmy uwagę na to, że układ równań (1.4) o macierzy Vandermonda (1.5) jest na ogól, przy dużych wartościach n, *bardzo źle uwarunkowany*. To też dla n dużych unika się wyznaczania P_n przy pomocy układu (1.4).

Algorytm różnic dzielonych, to jeszcze jeden sposób wyznaczania wielomianu interpolacyjnego Lagrange'a P_n.

Zdefiniujemy najpierw **różnice dzielone** dla funkcji f, określonej w węzłach $x_0, x_1, x_2, \cdots, x_n$. Symbolem

$$f[x_0, x_1, \cdots, x_k]$$

oznaczamy **k-tą różnicę dzieloną funkcji f dla węzłów** $x_0, x_1, x_2, \cdots, x_k$**.** Różnice dzielone definiujemy rekurencyjnie:

- $f[x_j] = f(x_j)$ **- zerowa różnica dzielona dla węzła** x_j**,**
- $f[x_0, x_1] = \frac{f(x_1)-f(x_0)}{x_1-x_0}$ **- pierwsza różnica dzielona dla węzłów** x_0 **i** x_1**,**
- $f[x_0, x_1, \cdots, x_{k+1}] = \frac{f[x_1,x_2,\cdots,x_{k+1}]-f[x_0,x_1,\cdots,x_k]}{x_{k+1}-x_0}$ **- k-ta różnica dzielona dla węzłów** $x_0, x_1, \cdots, x_{k+1}$**.**

Twierdzenie 1.3

$$f[x_0, x_1, x_2, \cdots, x_k] =$$

$$= \sum_{j=0}^{k} \frac{f(x_j)}{(x_j - x_0)(x_j - x_1)\cdots(x_j - x_{j-1})(x_j - x_{j+1})\cdots(x_j - x_k)}.$$

Wniosek 1.3
Wartość różnicy dzielonej $f[x_0, x_1, x_2, \cdots, x_k]$ *nie zależy od porządku argumentów* $x_0, x_1 \cdots, x_k$. $\square$

Zadanie 1.1
Udowodnić Twierdzenie 1.3. Można zastosować indukcję względem k.

Twierdzenie 1.4
Wielomian interpolacyjny Lagrange'a dla funkcji $f : [a, b] \to \mathbf{R}$*, oraz węzłów* $x_0, x_1, x_2, \cdots, x_n$ *da się zapisać w postaci Newtona:*

$$P_n(x) = f[x_0] + f[x_0, x_1](x - x_0) + f[x_0, x_1, x_2](x - x_0)(x - x_1)+$$

$$+ \cdots + f[x_0, x_1, \cdots, x_n](x - x_0)(x - x_1)\cdots(x - x_{n-1}). \tag{1.6}$$

Uwaga. Mamy tu jeszcze jedno przedstawienie wielomianu P_n przy pomocy *bazy newtonowskiej* podprzestrzeni wielomianów stopnia $\leq n$:

$$1,$$

$$x - x_0,$$
$$(x - x_0)(x - x_1),$$
$$\cdots\cdots\cdots\cdots\cdots$$
$$(x - x_0)(x - x_1) \cdots (x - x_{n-1}).$$

Współczynnikami rozwinięcia są w tym przypadku, *różnice dzielone.*
Dowód. (Indukcja względem n.)
Sprawdźmy najpierw, że wzór (1.6) wyznacza wielomian interpolacyjny Lagrange'a dla $n = 1$.

$$P_1(x) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x - x_0).$$

Stąd

$$P_1(x_0) = f(x_0),$$
$$P_1(x_1) = f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0}(x_1 - x_0) = f(x_1).$$

Ponieważ P_1 jest stopnia ≤ 1, jest to zatem wielomian interpolacyjny dla węzłów x_0 i x_1.

Zadanie 1.2
Wykonać krok indukcyjny. Wskazówka: Zakładamy, że wzór (1.6) zachodzi dla *dowolnego układu k węzłów* $x_{i_0}, x_{i_1} \cdots, x_{i_{k-1}}$. Udowodnić, że wzór ten przedstawia też wielomian interpolacyjny dla węzłów $x_0, x_1, \cdots, x_k$. Trzeba zauważyć najpierw że

$$P_k(x) = P_{k-1}(x) + f[x_0, x_1, \cdots, x_{k-1}, x_k](x - x_0)(x - x_1) \cdots (x - x_{k-1}),$$

i następnie sprawdzać, że $P_k(x_j) = f(x_j)$, najpierw dla $j = 0, 1, 2, \cdots, k-1$, w końcu dla $j = k$, wykorzystując to, że różnice dzielone nie zależą od porządku argumentów.

Tablica różnic dzielonych.

Kolejne różnice dzielone otrzymamy wypełniając poniższą *tablicę różnic dzielonych.* (Tablica dla $n = 4$.)

$$
\begin{array}{llllll}
x_0 & f[x_0] & & & & \\
 & & f[x_0,x_1] & & & \\
x_1 & f[x_1] & & f[x_0,x_1,x_2] & & \\
 & & f[x_1,x_2] & & f[x_0,x_1,x_2,x_3] & \\
x_2 & f[x_2] & & f[x_1,x_2,x_3] & & f[x_0,x_1,x_2,x_3,x_4] \\
 & & f[x_2,x_3] & & f[x_1,x_2,x_3,x_4] & \\
x_3 & f[x_3] & & f[x_2,x_3,x_4] & & \\
 & & f[x_3,x_4] & & & \\
x_4 & f[x_4] & & & &
\end{array}
$$

Tablicę tworzymy posługując się definicją rekurencyjną różnic dzielonych.

Zauważmy, że **dla wyznaczenia wielomianu interpolacyjnego Lagrange'a w postaci Newtona potrzebujemy tylko górnej diagonali tablicy.**

Zadanie 1.3 Napisz program obliczający wartość w zadanym punkcie x wielomianu interpolacyjnego Lagrange'a, stosując wzór (1.6) i tablicę różnic dzielonych.

Zadanie 1.4 Różnice dzielone nie zależą od porządku argumentów. Wyciągnij z tego wnioski dotyczące wzoru (1.6) i tablicy różnic dzielonych.

Oszacowanie błędu dla wielomianu interpolacyjnego Lagrange'a.

Niech P_n będzie wielomianem interpolacyjnym Lagrange'a dla funkcji $f : [a, b] \to \mathbf{R}$, o węzłach

$$a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b.$$

Twierdzenie 1.4 *Jeśli $f \in C^{n+1}([a, b])$, to dla każdego $x \in [a, b]$, w przedziale otwartym*

$$(\min\{x, x_0, \cdots, x_n\}, \max\{x, x_0, \cdots, x_n\})$$

istnieje punkt $\xi(x)$, taki że

$$f(x) - P_n(x) = \frac{f^{n+1}(\xi(x))}{(n+1)!}\omega(x),$$

gdzie $\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$.

Uwaga: To twierdzenie podaje błąd interpolacji w każdym punkcie $x \in [a, b]$. Zauważmy, że błąd zależy jedynie od własności funkcji aproksymowanej f, oraz od węzłów interpolacji ($\omega(x)$).

Dowód
Niech

$$K(x) = \begin{cases} \frac{f(x)-P_n(x)}{\omega(x)} & \text{gdy } x \neq x_j, \quad j = 0, 1, \cdots, n \\ 0 & \text{gdy } x = x_j, \quad j = 0, 1, \cdots, n \end{cases}$$

oraz

$$F(t, x) = f(t) - P_n(t) - K(x)\omega(t).$$

Potraktujemy t jako zmienną, zaś x jako ustalony parametr. Zauważmy, że $F(t, x)$ jest funkcją różniczkowalną $n + 1$ razy w sposób ciągły jako funkcja zmiennej $t \in [a, b]$. Ponad to

$$F(x, x) = 0,$$

$$F(x_j, x) = 0,\ j = 0, 1, \cdots, n.$$

Jeśli $x \neq x_j$ $j = 0, 1, \cdots, n$, to $F(t, x)$ traktowana jako funkcja zmiennej t, zeruje się w $n + 2$ *różnych* punktach przedziału $[a, b]$

$$x, x_0, x_1, \cdots, x_n.$$

Stosując $n + 1$ razy *twierdzenie Rolle'a* do kolejnych pochodnych funkcji F względem t, widzimy że

- $\frac{\partial}{\partial t}F(t, x)$ znika w n punktach między kolejnymi węzłami $x, x_1, \cdots, x_n$, a więc n razy w różnych punktach przedziału otwartego (a, b),
- $\frac{\partial^2}{\partial t^2}F(t, x)$ znika w $n - 1$ różnych punktach przedziału (a, b)

  ....

- $\frac{\partial^{n+1}}{\partial t^{n+1}}F(t, x)$ znika przynajmniej w jednym punkcie przedziału (a, b). Oznaczmy ten punkt przez $\xi(x)$.

Zgodnie z definicją funkcji $F(t,x)$, mamy

$$\frac{\partial^{n+1}}{\partial t^{n+1}}F(t,x) = f^{(n+1)}(\xi(x)) - K(x)(n+1)! = 0,$$

gdyż $\omega^{(n+1)}(t) = (n+1)!$.

Stąd, gdy $x \neq x_j \; j = 0,1,\cdots,n$

$$K(x) = \frac{f(x) - P_n(x)}{\omega(x)} = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}$$

lub inaczej

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega(x).$$

Wzór ten pozostaje prawdziwy, również gdy $x = x_j \; j = 0,1,\cdots,n$. $\square$

Wnioski

1. Z twierdzenia 1.5 wynika, że jeśli $f \in C^{n+1}([a,b])$, to dla błędu interpolacji Lagrange'a mamy następujące oszacowanie w normie w normie "sup" na przedziale $[a,b]$:

   (1.10). $$\|f - P_n\|_{\infty,[a,b]} \leq \frac{\|f^{(n+1)}\|_{\infty,[a,b]}}{(n+1)!}\|\omega\|_{\infty,[a,b]}$$

   We wzorze tym błąd interpolacji jest szacowany z góry przez wyrażenie zależne od normy "sup" $n+1$-szej pochodnej funkcji f.

2. **Zadanie 1.5** Udowodnić, że jeśli

   $$a = x_0 < x_1 < x_2 < \cdots < x_n = b,$$

   to

   (1.11) $$\|\omega\|_{\infty,[a,b]} \leq \frac{n!h^{n+1}}{4},$$

   gdzie $h = \max_j(x_{j+1} - x_j)$.

   Wskazówka. Zastosować indukcję względem n.

Ze wzorów (1.10) i (1.11) wynika następujące oszacowanie błędu interpolacji Lagrange'a w zależności od h:

$$\|f - P_n\|_{\infty,[a,b]} \le \frac{\|f^{n+1}\|_{\infty,[a,b]}}{4(n+1)} h^{n+1}.$$

To oszacwanie ma następującą *wadę:* liczba węzłów jest związana z regularnością funkcji f. Zatem rząd pochodnej dąży do ∞, gdy liczba węzłów dąży do ∞.

3. Niech L_n będzie operatorem (funkcją) przyporządkowującym funkcji $f \in C([a,b])$ jej wielomian interpolacyjny Lagrange'a dla danego, ustalonego układu $n+1$ węzłów:

   $$L_n : f \to P_n.$$

   Łatwo zauważyć, że jest to operator liniowy (dlaczego?). Przypuśćmy, że rozważamy ciąg układów $n+1$ węzłów:

   $$a \le x_0^n < x_1^n < \cdots < x_n^n \le b.$$

   Pytanie Czy dla każdego $f \in C([a,b])$

   $$\|L_n f - f\|_{\infty,[a,b]} \to 0 \tag{1.12}$$

   gdy $n \to \infty$, przy dowolnym ciągu układów węzłów.

   Jest to pytnie o zbieżność interpolacji Lagrange'a dla dowolnej funkcji ciągłej f. Okazuje się, że w przestrzeni $C([a,b])$ zależność (1.12) nie zachodzi przy dowolnie wybranym ciągu układów węzłów, bez dodatkowych założeń o funkcji f. Inaczej mówiąc interpolacja Lagrange'a *nie jest aproksymacją zbieżną w przestrzeni* $C([a,b])$.

Wzór (1.9) i oszacowanie (1.10) zakładają, że funkcja f ma tyle pochodnych ciągłych, ile wynosi liczba węzłów interpolacji. Nasuwa się naturalne pytanie, *co można powiedzieć o błędzie interpolacji Lagrange'a jeśli f ma mniej pochodnych ciągłych niż liczba węzłów interpolacji.* Można na nie odpowiedzieć wykorzystując *Twierdzenie Jacksona.* To twierdzenie podaje oszacowanie błędu dla *wielomianu najlepszej aproksymacji* stopnia $\le n$ w sensie normy "sup" na przedziale $[a, b]$. Jak wiemy, taki wielomian zawsze istnieje (dlaczego?). Oszacowanie w Twierdzeniu Jacksona zależy od *stopnia regularności* funkcji f.

Twierdzenie Jacksona. *Niech $f \in C^s([a, b])$, oraz niech Q_n będzie wielomianem stopnia $\leq n$, najlepszej aproksymacji dla f w sensie normy $\|\cdot\|_{\infty,[a,b]}$. Wtedy*

$$\|f - Q_n\|_{\infty,[a,b]} \leq$$

$$\leq \begin{cases} 6\omega(f, \frac{b-a}{2n}), & gdy\ f \in C([a, b]), \\ 3\frac{b-a}{n}\|f'\|_{\infty,[a,b]}, & gdy\ f \in C^1([a, b]), \\ 6\frac{(s-1)^{s-1}}{(s-1)!}s(\frac{b-a}{n})^s\|f^{(s)}\|_{\infty,[a,b]}, & gdy\ f \in C^s([a, b]), \quad s \geq 2. \end{cases}$$

Tutaj

$$\omega(f, \tau) = \sup_{|\Delta t| \leq \tau, t, t+\tau \in [a,b]} |f(t + \Delta t) - f(t)|$$

jest tak zwanym modułem ciągłości funkcji f na $[a, b]$.

Mając Twierdzenie Jacksona potrafimy oszacować błąd interpolacji Lagrange'a także, gdy $f \in C^s([a, b])$, $s \leq n$.

Istotnie, niech P_n będzie wielomianem interpolacyjnym Lagrange'a dla f o węzłach

$$a \leq x_0, x_1 \cdots, x_n \leq b.$$

Oznaczając przez Q_n wielomian najlepszej aproksymacji, mamy

$$f - P_n = f - Q_n + Q_n - P_n.$$

Zauważmy, że wielomian iterpolacyjny Lagrange'a dla Q_n o podanych węzłach jest poprostu równy Q_n (odpowiedz dla czego?). Możemy więc napisać, używając funkcji bazowych Lagrange'a l_j (patrz (1.2) i (1.3))

$$Q_n = \sum_{j=0}^{n} l_j Q_n(x_j),$$

$$P_n = \sum_{j=0}^{n} l_j f(x_j).$$

Stąd

$$f - P_n = f - Q_n + \sum_{j=0}^{n} l_j(Q_n(x_j) - f(x_j)).$$

Teraz szacując

$$\|f - Q_n\|_{\infty,[a,b]} \le \|f - Q_n\|_{\infty,[a,b]} + \sum_{j=0}^{n} \|l_j\|_{\infty,[a,b]} \sup_{x\in[a,b]} |Q_n(x) - f(x)| =$$

$$= (1 + \sum_{j=0}^{n} \|l_j\|_{\infty,[a,b]})\|f - Q_n\|_{\infty,[a,b]}.$$

Zadanie 1.6. Udowodnij, że jeśli

$$a = x_0 < x_2 < \cdots < x_n = b,$$

to

$$\|l_j\|_{\infty,[a,b]} \le \frac{n!}{j!(n-j)!}(\frac{h}{\bar{h}})^n \tag{1.13}$$

gdzie $h = \max_j(x_{j+1} - x_j)$, $\bar{h} = \min_j(x_{j+1} - x_j)$.

Ostatecznie, wykorzystując wzór (1.13), otrzymamy oszacowanie błędu dla wielomianu interpolacyjnego P_n dla węzłów $a = x_0 < x_1, \cdots < x_n = b$:

jeśli $f \in C^s([a,b])$, $0 \le s \le n$, to

$$\|f - P_n\|_{\infty,[a,b]} \le (1 + 2^n(\frac{h}{\bar{h}})^n)\|f - Q_n\|_{\infty,[a,b]}. \tag{1.14}$$

INTERPOLACJA HERMITE'A

Załóżmy jak poprzednio, że dane są różne węzły w przedziale $[a,b]$:

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

Ponadto przypuśćmy, że każdemu z węzłów przyporządkowana jest liczba naturalna $m_j \ge 1$, zwana *krotnością węzła* x_j. Niech $f \in C^{(\max_j m_j)-1}([a,b])$.

(1.15) Zadanie interpolacji (wielomianowej, globalnej) Hermite'a

Dla danej funkcji f, oraz danej tablicy węzłów i krotności

$$\begin{array}{ccccc} x_0 & x_1 & x_2 & \cdots & x_n \\ m_0 & m_1 & m_2 & \cdots & m_n \end{array}$$

znaleźć wielomian P_M stopnia $\leq M = (\sum_{j=0}^{n} m_j) - 1$ taki, że

$$P_M^{(s)}(x_j) = f^{(s)}(x_j) \; j = 0, 1, \cdots, n \; ; \; s = 0, 1, \cdots, m_j - 1 \tag{1.15.}$$

Twierdzenie 1.5 *Zadanie interpolacyjne Hermite'a (1.15) dla funkcji f dostatecznie regularnej ma jednoznaczne rozwiązanie.*

Dowód. Zadanie 1.7 Udowodnić Twierdzenie 1.4.
Wskazówka: Zapiszmy: $P_M(x) = \sum_{j=0}^{M} a_j x^j$. Teraz widać, że zadanie (1.15) polega na rozwiązaniu układu równań liniowych algebraicznych, z którego należy wyznaczyć współczynniki $a_0, a_1, \cdots, a_M$. Wypisz postać macierzy tego układu, oraz udowodnij, że przy przyjętych założeniach jest ona nieosobliwa.
□

Uwaga

- Z podobnych względów jak w przypadku interpolacji Lagrange'a, układ równań z zadania interpolacyjnego (1.15) przy większych wartościach n, nie jest na ogól używany do numerycznego wyznaczania wielomianu interpolacyjnego P_M.
- Interpolacja Hermite'a może być uważana za graniczny przypadek interpolacji Lagrange'a, *gdy pewne węzły interpolacji w granicy sklejają się*. Stąd można łatwo wyprowadzić wnioski co do szacowania błędu tego rodzaju interpolacji.

Dość wygodny algortm realizujący zadanie interpolacji Hermite'a jest oparty na różnicach dzielonych. Aby go opisać musimy zdefiniować *różnice dzielone z powtórzeniami.*

Różnicę dzieloną o różnych węzłach $x_0, x_1, \cdots, x_n$ z powtórzeniami odpowiednio $k_0, k_1, \cdots, k_n$ razy oznaczamy symbolem:

$$f[x_0 k_0, x_1 k_1, \cdots, x_n k_n].$$

Jeśli $k_0 = k_1 = \cdots = k_n = 1$, jest to zwykła różnica dzielona $f[x_0, x_1, \cdots, x_n]$; jeśli którąś z liczb $k_j = 0$, to oznacza że węzeł x_j nie występuje. Z definicji przyjmiemy:

$$f[xk] = \frac{f^{(k-1}(x)}{(k-1)!},$$

oraz dla $k_j \leq 1$, $j = 0, 1, 2, \cdots, n$:

$$f[x_0k_0, x_1k_1, \cdots, x_nk_n] =$$

$$= \frac{f[x_0k_0 - 1, x_1k_1, \cdots, x_nk_n] - f[x_0k_0, x_1k_1, \cdots, x_nk_n - 1]}{x_n - x_0} \tag{1.16.}$$

Wzory (1.8) pozwalają tworzyć i wykorzystywać do budowy wielomianu interpolacyjnego Hermite'a *tablicę różnic dzielonych*, w podobny sposób, jak w przypadku interpolacji Lagrange'a.

Przykład. Chcemy zbudować wielomian interpolacyjny Hermite'a o dwóch węzłach $x_0 < x_1$ ikrotnościach 4 i 3 odpowiednio. Wielomian będzie stopnia $\leq 4 + 3 - 1 = 6$.

$$\begin{array}{rcl}
P_6(x_0) & = & f(x_0) \\
P_6^{(1)}(x_0) & = & f^{(1)}(x_0) \\
P_6^{(2)}(x_0) & = & f^{(2)}(x_0) \\
P_6^{(3)}(x_0) & = & f^{(3)}(x_0) \\
P_6(x_1) & = & f(x_1) \\
P_6^{(1)}(x_1) & = & f^{(1)}(x_1) \\
P_6^{(2)}(x_1) & = & f^{(2)}(x_1)
\end{array}$$

Zbudujemy najpierw tablicę różnic dzielonych z powtórzeniami. W tej tablicy węzeł o krotności k pojawi się k- razy i odpowiadać mu będą wartości funkcji f i jej $k-1$ pochodnych, *jako dane zadania.* Startując od danych zadania, uzupełniamy tablicę wykorzystując wzór (1.8).

$$\begin{array}{cccccccc}
x_0 & f[x_0] & & & & & & \\
 & & f[x_0 2] & & & & & \\
x_0 & f[x_0] & & f[x_0 3] & & & & \\
 & & f[x_0 2] & & f[x_0 4] & & & \\
x_0 & f[x_0] & & f[x_0 3] & & f[x_0 4, x_1] & & \\
 & & f[x_0 2] & & f[x_0 3, x_1] & & f[x_0 4, x_1 2] & \\
x_0 & f[x_0] & & f[x_0 2, x_1] & & f[x_0 3, x_1 2] & & f[x_0 4, x_1 3] \\
 & & f[x_0, x_1] & & f[x_0 2, x_1 2] & & f[x_0 3, x_1 3] & \\
x_1 & f[x_1] & & f[x_0, x_1 2] & & f[x_0 2, x_1 3] & & \\
 & & f[x_1 2] & & f[x_0, x_1 3] & & & \\
x_1 & f[x_1] & & f[x_1 3] & & & & \\
 & & f[x_1 2] & & & & & \\
x_1 & f[x_1] & & & & & &
\end{array}$$

Wielomian interpolacyjny Hermite'a P_6 budujemy w oparciu o wzór analogiczny do wzoru (1.6). Aby wypisac prawidłowo jego poszczególne elementy najlepiej krotne węzły *rozmnożyć* zastępując węzeł l-krotny x_k, l- różnymi węzłami, na przykład

$$x_k^1, x_k^2, \cdots, x_k^l.$$

Wypisać wielomian interpolacyjny Lagrange'a wykorzystując wzór (1.6), a następnie spowrotem zidentyfikować węzły $x_k^1, x_k^2, \cdots, x_k^l$, jako x_k. Nasz wielomian P_6 jest następującej postaci:

$$P_6(x) = f[x_0] + f[x_0 2](x - x_0) + f[x_0 3](x - x_0)^2 + f[x_0 4](x - x_0)^3 +$$

$$+f[x_0 4, x_1](x-x_0)^4+f[x_0 4, x_1 2](x-x_0)^4(x-x_1)+f[x_0 4, x_1 3](x-x_0)^4(x-x_0)^2.$$

INTERPOLACJA TRYGONOMETRYCZNA

Często zachodzi potrzeba aproksymacji funkcji nie przy pomocy *zwykłych wielomianów*, ale przy pomocy *wielomianów trygonometrycznych*.

Funkcję (zmiennej rzeczywistej), mającą wartości *zespolone* postaci

$$T_n(x) = \sum_{j=0}^{n} c_j e^{ixj},$$

gdzie c_j są zespolonymi współczynnikami zaś $i = \sqrt{-1}$, nazywamy *wielomianem trygonometrycznym* stopnia $\leq n$. Nazwa *trygonometryczny* bierze się stąd, że

$$e^{ixj} = (e^{ix})^j = cos(jx) + isin(jx).$$

Będziemy rozpatrywać funkcje $f : [0, 2\pi] \to \mathbf{C}$, które są okresowe z okresem równym 2π. Oznacza to, ze $f(0) = f(2\pi)$. Takie funkcje można *przedłużyć* na całą prostą rzeczywistą, i wtedy, po przedłużeniu, spełniają one warunek $f(x) = f(x + 2\pi)$.

Bedziemy omawiać tu jedynie interpolację przy pomocy wielomianów trygonometrycznych - *interpolację trygonometryczną* - dla następującego układu węzłów równoodległych, leżących w przedziale $[0, 2\pi]$

$$x_k = \frac{2\pi}{n+1} k, \quad k = 0, 1, \cdots, n.$$

(1.17) Zadanie interpolacji trygonometrycznej.

Poszukujemy wielomianu trygonometrycznego stopnia $\leq n$

$$T_n(x) = \sum_{j=0}^{n} c_j e^{ixj},$$

spełniającego warunki

$$T_n(x_k) = f(x_k) \quad k = 0, 1, \cdots, n \tag{1.17}$$

dla układu równoodległych węzłów $x_k = \frac{2\pi}{n+1} k$.

Twierdzenie 1.6 *Zadanie interpolacji trygonometrycznej (1.17) ma jednoznaczne rozwiązanie.*

Dowód. Aby wyznaczyć wielomian T_n, możemy rozwiązać układ równań liniowych algebraicznych, z którego wyliczymy współczynniki

$$c_j, \quad j = 0, 1, \cdots, n.$$

Łatwo zauważyć, że macierzą tego układu jest, podobnie jak poprzednio, *macierz Vandermonda* utworzona dla $n+1$ różnych liczb $z_k = e^{ix_k}, \quad k = 0, 1, \cdots, n$, a więc jest to macierz odwracalna. □

Wygodnie będzie oznaczyć *funkcje bazowe* rozwinięcia wielomianu T_n

$$\phi_j(x) = e^{ixj} \quad j = 0, 1, \cdots, n.$$

Zdefiniujemy również, dla funkcji f, g określonych w rozważanych tu węzłach, *iloczyn skalarny*

$$(f, g) = \sum_{k=0}^{n} f(x_k) \bar{g}(x_k).$$

Zauważmy od razu, że nasze funkcje bazowe $\phi_j \quad j = 0, 1, \cdots, n$ stanowią *układ ortogonalny* w sensie tego iloczynu skalarnego. Istotnie:

$$(\phi_r, \phi_s) = \sum_{k=0}^{n} \phi_r(x_k)\bar{\phi}_s(x_k) =$$

$$= \sum_{k=0}^{n} (e^{i\frac{2\pi}{n+1}(r-s)})^k.$$

Oznaczmy $q = e^{i\frac{2\pi}{n+1}(r-s)}$. Wtedy

$$(\phi_r, \phi_s) = \sum_{j=0}^{n} q^j = \begin{cases} n+1 & \text{gdy} \quad q = 1 \\ \frac{1-q^{n+1}}{1-q} & \text{gdy} \quad q \neq 1 \end{cases}.$$

Ponieważ $r - s$ jest liczbą całkowitą, to

$$q^{n+1} = (e^{i\frac{2\pi}{n+1}(r-s)})^{n+1} = e^{i2\pi(r-s)} = 1.$$

Stąd

$$(\phi_r, \phi_s) = \delta_{r,s}(n+1).$$

Fakt ortogonalności układu funkcji bazowych $\phi_k \quad k = 0, 1, \cdots, n$ pozwala w prosty sposób wyrazić współczynniki c_j wielomianu interpolacyjnego T_n.

Mnożąc stronami wzór (1.17) z prawej strony przez $\bar{\phi}_r(x_k)$, oraz sumując dla $k = 0, 1, \cdots, n$ otrzymamy

$$(T_n, \phi_r) = \sum_{j=0}^{n} c_j(\phi_j, \phi_r) = (f, \phi_r).$$

Ale $(\phi_j, \phi_r) = \delta_{j,k}(n+1)$, więc

$$c_r = \frac{1}{n+1}(f, \phi_r), \quad r = 0, 1, \cdots, n. \tag{1.18}$$

Współczynniki $c_j = \frac{(f,\phi_j)}{n+1}$ noszą nazwę *współczynników Fouriera* funkcji f względem układu ortogonalnego $\phi_k, \quad k = 0, 1, \cdots, n$. Zadanie obliczania współczynnikóq Fouriera nazywa się *analizą fourierowską* zaś zadanie obliczania wartości wielomianu interpolacyjnego $T_n(x) = \sum_{j=0}^{n} c_j\phi_j(x)$ - *syntezą fourierowską.*

Zauważmy, że

$$c_k = \frac{1}{n+1} \sum_{j=0}^{n} f(x_j) e^{-ix_j k},$$

zaś

$$T_n(x) = \sum_{j=0}^{n} c_j e^{ixj}.$$

Można więc powiedzieć że analiza i synteza fourierowska sprowadzają się do *obliczania kombinacji liniowych funkcji wykładniczych.*

Na przykład, wyliczanie $c_0, c_1, \cdots, c_n$ przy użyciu powyższych wzorów wymaga liczby działań rzędu $(n+1)^2$ (mnożenie przez e^{ix_j} uważamy za jedno działanie). Istnieje jednak algorytm bardziej oszczędny: **FFT (Fast Fourier Transform - Szybkie Przekształcenie Fouriera).**

FFT FAST FOURIER TRANSFORM - SZYBKIE PRZEKSZTAŁCENIE FOURIERA

Algorytm FFT przedstawimy w szczególnym przypadku, gdy liczba węzłów interpolacji spełnia równość $N = n+1 = 2^r$, dla pewnego całkowitego r. Zajmiemy się przypadkiem *analizy fourierowskiej*, czyli wyliczeniem wartości współczynnika fourierowskiego, to jest wyrażenia

$$c_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-i\frac{2\pi qj}{N}},$$

gdzie $N = n + 1 = 2^r$ dla pewnego całkowitego r, i dla ustalonego q spośród $q = 0, 1, 2, \cdots, N-1$. Przypadek *syntezy fourierowskiej* nie rózni się od analizy w sposób istotny.

Pomysł polega na tym, żeby nie wykonywać zbędnych obliczeń: w tym wypadku żeby *nie wykonywać mnożeń przez* 1.

Przeanalizujemy dokładnie wzór dla współczynnika c_q. Zapiszemy najpierw q i j w systemie binarnym:

$$q = \sum_{k=1}^{r} q_k 2^{k-1} = q_1 2^0 + q_2 2^1 + \cdots + q_r 2^{r-1},$$

$$j = \sum_{m=1}^{r} j_{r-m+1} 2^{m-1} = j_r 2^0 + j_{r-1} 2^1 + \cdots + j_1 2^{r-1}.$$

W rozwinięciu binarnym liczby j rozmyślnie użyliśmy numeracji cyfr binarnych *w odwrotną stronę*. Stąd, wyodrębniając część całkowitą wyrażenia, którą oznaczamy przez s, mamy

$$\frac{qj}{N} = \frac{qj}{2^r} = \sum_{m=1}^{r} \sum_{k=1}^{r} q_k j_{r-m+1} 2^{m+k-r-2} = s + \sum_{m=1}^{r} j_{r-m+1} \sum_{k=1}^{r-m+1} 2^{m+k-r-2} q_k,$$

ponieważ $m + k - r - 2 < 0$ dla części ułamkowej. Biorąc pod uwagę to, że $N = 2^r$ i że $e^{-i2\pi s} = 1$, możemy napisać używając zapisu binarnego wskaźnika j przy f_j, $j = j_1 j_2 j_3 \cdots j_r$

$$c_q = \frac{1}{N} \sum_{j=0}^{N-1} f_j e^{-i2\pi \sum_{m=1}^{r} j_{r-m+1} \sum_{k=1}^{r-m+1} 2^{m+k-r-2} q_k} =$$

$$= \frac{1}{2}\frac{1}{2} \cdots \frac{1}{2} \sum_{j_r=0}^{1} (\sum_{j_{r-1}=0}^{1} \cdots (\sum_{j_1=0}^{1} f_{j_1 j_2 j_3 \cdots j_r} \cdot$$

$$\cdot e^{-2\pi i [j_r 2^{-r} \sum_{k=1}^{r} 2^{k-1} q_k]} \cdot e^{-2\pi i [j_{r-1} 2^{1-r} \sum_{k=1}^{r-1} 2^{k-1} q_k]} \cdot \cdots \cdot e^{-2\pi i [j_1 2^{-1} q_1]}) \cdots).$$

Porządkując ten wzór otrzymamy ostatecznie

$$c_q =$$

$$= \frac{1}{2} \sum_{j_r=0}^{1} (e^{-2\pi i [j_r 2^{-r} \sum_{k=1}^{r} 2^{k-1} q_k]} \cdot \frac{1}{2} \sum_{j_{r-1}=0}^{1} (e^{-2\pi i [j_{r-1} 2^{1-r} \sum_{k=1}^{r-1} 2^{k-1} q_k]} \cdots$$

$$\cdots \frac{1}{2} \sum_{j_2=0}^{1} (e^{-2\pi i [j_2 2^{-2} \sum_{k=1}^{2} 2^{k-1} q_k]} \cdot \frac{1}{2} \sum_{j_1=0}^{1} (e^{-2\pi i [j_1 2^{-1} q_1]} f_{j_1 j_2 \cdots j_r})) \cdots)) \tag{1.19}$$

Oznaczmy

$$c^0(j_1 j_2 \cdots j_r) = f_{j_1 j_2 \cdots j_r},$$

oraz określimy rekurencyjnie

$$c^1(q_1 j_2 \cdots j_r) = \frac{1}{2} \sum_{j_1=0}^{1} e^{-2\pi i [j_1 2^{-1} q_1]} c^0(j_1 j_2 \cdots j_r),$$

$$c^2(q_1q_2j_3\cdots j_r)=\frac{1}{2}\sum_{j_2=0}^{1}e^{-2\pi i[j_2 2^{-2}(q_1 2^0+q_2 2^1)]}c^1(q_1j_2\cdots j_r),$$

$$\cdots\cdots\cdots\cdots\cdots$$

$$c^l(q_1q_2\cdots q_lj_{l+1}\cdots j_r)=$$

$$=\frac{1}{2}\sum_{j_l=0}^{1}e^{-2\pi i[j_l 2^{-l}(q_1 2^0+q_2 2^1+\cdots q_l 2^{l-1})]}c^{l-1}(q_1q_2\cdots q_{l-1}j_l\cdots j_r).$$

Ze wzoru (1.19) wynika, że

$$c_q=c^r(q_1q_2\cdots q_r)=\frac{1}{2}\sum_{j_r=0}^{1}e^{-2\pi i[j_r 2^{-r}(q_1 2^0+q_2 2^1+\cdots q_r 2^{r-1})]}c^{r-1}(q_1q_2\cdots q_{r_1}j_r).$$

Oznacza to, że po r krokach tego algorytmu rekurencyjnego wyliczymy współczynnik fourierowski c_q. Zauważmy teraz, że gdybyśmy wyliczali wszystkie współczynniki $c_0, c_1, \cdots, c_{N-1}$, to na każdym kroku rekursji musielibyśmy wykonać liczbę operacji rzędu $O(N)$. Zatem wyliczenie wszystkich współczynników kosztowałoby liczbę operacji rzędu $O(Nr) = O(N \log_2 N)$ (zamiast $O(N^2)$, w przypadku bezpośredniego stosowania wzorów (1.18) definiujących te współczynniki).

Przykład

Niech $N = n + 1 = 8 = 2^3$, zatem $r = 3$. W tym przypadku algorytm **FFT** wykonuje $r = 3$ kroki. Oto poszczególne etapy wypisane dla obliczenia współczynnika $c_q = c_{q_3q_2q_1}$:

$$\begin{array}{llll}
c^0(000) & & & \\
c^0(001) & & & \\
c^0(010) & c^1(q_100) & & \\
c^0(011) & c^1(q_101) & c^2(q_1q_20) & \\
\quad\cdots & \quad\cdots & \quad\cdots & c^3(q_1q_2q_3)=c_q \\
c^0(100) & c^1(q_110) & c^2(q_1q_21) & \\
c^0(101) & c^1(q_111) & & \\
c^0(110) & & & \\
c^0(111) & & &
\end{array}$$

INTERPOLACJA SPLAJNOWA

Przykład. Niech $f : [a, b] \to \mathbf{R}$ będzie funkcją ciągłą. Przedział $[a, b]$ podzielimy na N równych części przy pomocy punktów

$$a = x_0 < x_1 < x_2 < \cdots < x_N = b$$

gdzie $x_j = x_0 + jh$, $h = \frac{b-a}{N}$, $j = 0, 1, \cdots N$. Dla każdego podprzedziału $[x_j, x_{j+1}]$ zbudujemy wielomian interpolacyjny Lagrange'a funkcji f o węzłach x_j i x_{j+1}. Otrzymamy w ten sposób *łamaną* - funkcję przedziałami liniową, interpolującą w sensie Lagrange'a funkcję f na przedziale $[a, b]$. Oznaczmy przez s_N tak otrzymaną funkcję przedziałami liniową.

Zadanie 1.8 Używając wiadomości dotyczących oszacowania błędu interpolacji Lagrange'a

1. Udowodnij, że $\|f - s_N\|_{\infty,[a,b]} \to 0$ gdy $N \to \infty$,
2. Oszacuj błąd $f - s_N$ gdy $f \in C^1([a, b])$, oraz gdy $f \in C^2([a, b])$.

Widzimy więc, że funkcja interpolująca s_N zbiega jednostajnie do f gdy $N \to \infty$ nawet przy założeniu, że f jest tylko funkcją ciągłą. W tym przypadku sytuacja jest zupełnie inna niż w przypadku interpolacji *globalnej* jednym wielomianem interpolacyjnym Lagrange'a dla węzłów $x_0 < x_1 < \cdots < x_N$. Dla interpolacji globalnej nie było zbieżności, gdy $h \to 0$.

Funkcje s_N zdefiniowane wyżej są szczególnym przypadkiem *splajnów wielomianowych.*

Definicja. *Niech π będzie podziałem odcinka $[a, b]$ dokonanym przy pomocy węzłów $a = x_0 < x_1 < \cdots < x_N = b$.* $\mathbf{S_n^m}(\pi)$ *jest przestrzenią liniową (z działaniami $+$ i $\cdot$ określonymi w sposób naturalny) wszystkich funkcji s_N, które na każdym z przedziałów $[x_j, x_{j+1}]$, $j = 0, 1, \cdots, N-1$ są wielomianami stopnia $\leq n$, połączonymi w ten sposób, że $s_N \in C^m([a, b])$. Te przestrzenie liniowe noszą nazwę* **przestrzeni splajnów**.

W omówionym przykładzie występuje *zadanie interpolacji przy pomocy splajnów z przestrzeni* $\mathbf{S_1^0}(\pi)$. Zadanie tam omówione wskazuje na to, że interpolacja splajnowa może być zbieżna już dla funkcji ciągłych, a *szybkość zbieżności zależy od gładkości funkcji interpolowanej.*

Szczególną rolę odgrywa interpolacja przy pomocy elementów przestrzeni $\mathbf{S_{2n+1}^{2n}}(\pi)$. Zadanie interpolacyjne w tym przypadku formułuje się wyjątkowo prosto. Rozpatrzymy tu przypadek tak zwanych B-splajnów kubicznych; wtedy $n = 1$, a więc przestrzeń splajnów, to $\mathbf{S_3^2}(\pi)$. Są to *przedziałami wielomiany stopnia* ≤ 3, które są funkcjami klasy $C^2([a, b])$.

Sformułowanie zadania interpolacji przy pomocy splajnów kubicznych z przestrzeni $\mathbf{S_3^2}(\pi)$.

Przypuśćmy, że podzial π odcinka $[a, b]$ definiuje następujący układ węzłów:

$$a = x_0 < x_1 < \cdots < x_N = b,$$

gdzie $x_j = x_0 + jh$, $h = \frac{b-a}{N}$. Określimy najpierw tak zwane *B-splajny kubiczne*, związane z podziałem π odcinka $[a, b]$. W tym celu rozszerzymy przedział $[a, b]$, oraz zbiór punktów π dodając punkty x_{-2}, x_{-1}, oraz x_{N+1} i x_{N+2}. Teaz

$$\pi : \{x_{-2} < x_{-1} < x_0 < \cdots < x_N < x_{N+1} < x_{N+2}\}.$$

Z każdym z punktów $x_{-1}, x_0, x_1, \cdots, x_N, x_{N+1}$ zwiążemy funkcję B_j, $j = -1, 0, 1, \cdots, N, N + 1$, należącą do przestrzeni $\mathbf{S_3^2}(\pi)$, tak zwany *B-splajn kubiczny*, określony w sposób następujący:

$$B_j(x) =$$

$$= \frac{1}{h^3}\begin{cases} (x - x_{j-2})^3 & x \in [x_{j-2}, x_{j-1}] \\ h^3 + 3h^2(x - x_{j-1}) + 3h(x - x_{j-1})^2 - 3(x - x_{j_1})^3 & x \in [x_{j-1}, \quad x_j] \\ h^3 + 3h^2(x_{j+1} - x) + 3h(x_{j+1} - x)^2 - 3(x_{j+1} - x)^3 & x \in [x_j, \quad x_{j+1}] \\ (x_{j+2} - x)^3 & x \in [x_{j+1}, x_{j+2}] \\ 0 & x \notin [x_{j-2}, x_{j+2}] \end{cases}$$

Zadanie 1.9 Udowodnij, że funkcje B_j, $j = -1, 0, 1, \cdots, N, N + 1$ należą do przestrzeni $\mathbf{S_3^2}(\pi)$.

Można udowodnić [1], *że funkcje*

$$B_{-1}, B_0, B_1, \cdots, B_N, B_{N+1}$$

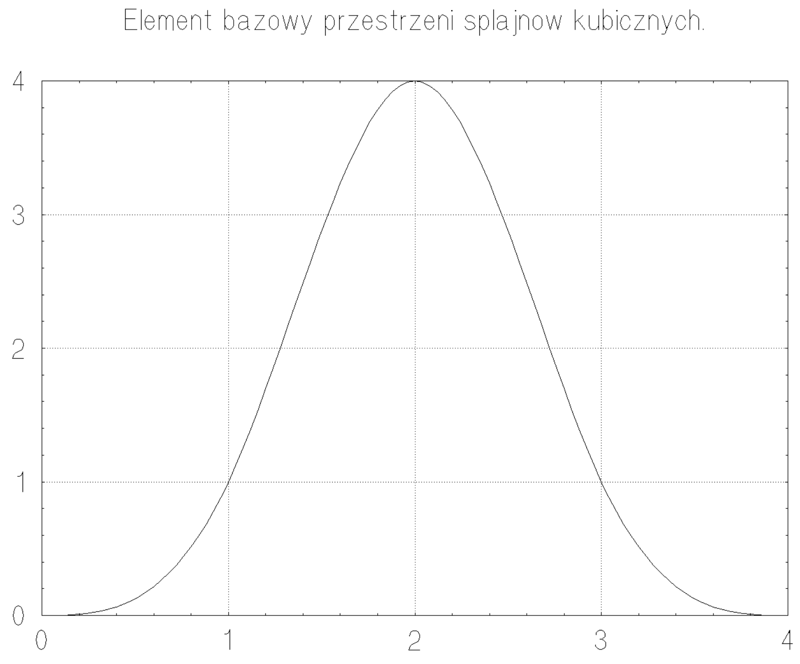
stanowią **bazę przestrzeni** $\mathbf{S_3^2}(\pi)$, *gdzie* π *jest równomiernym podziałem odcinka* $[a, b]$ *przy pomocy węzłów*

$$a = x_0 < x_1 < \cdots < x_N = b,$$

$$x_j = x_0 + jh, \quad h = \frac{b-a}{N}, \quad j = 0, 1, \cdots, N.$$

Zatem, w tym przypadku, przestrzeń $\mathbf{S_3^2}(\pi)$ *ma wymiar* $N + 3$.

Poniższy wykres przedtawia fragment wykresu funkcji B_j, ograniczony do jej *nośnika*, to jest do zbioru $[x_{j-2}, x_{j+2}]$. Na osi poziomej wykresu, punkty 0, 1, 2, 3, 4 są przyporządkowane odpowiednio punktom x_{j-2}, x_{j-1}, x_j, x_{j+1}, x_{j+2}.



[1] Patrz P.M.Prenter "Splines and Variational Methods". Książka jest w bibliotece WMIM.

Wielkość nośnika funkcji bazowej B_j ma istotne znaczenie przy różych operacjach obliczeniowych z użyciem splajnów z przestrzeni $\mathbf{S_3^2}(\pi)$. Zauważmy, że jedynie funkcje B_{j-2}, B_{j-1}, B_j, B_{j+1}, B_{j+2} mają nośniki o nie rozłącznym wnętrzu z nośnikiem funkcji B_j.

Przy wykorzystywaniu funkcji z przestrzeni $\mathbf{S_3^2}(\pi)$, pomocna może być następująca tablica wartości funkcji B_j, B_j' i B_j'':

	x_{j-2}	x_{j-1}	x_j	x_{j+1}	x_{j+2}
B_j	0	1	4	1	0
B_j'	0	$3/h$	0	$-3/h$	0
B_j''	0	$6/h^2$	$-12/h^2$	$6/h^2$	0

(1.20) Zadanie interpolacji typu Lagrange'a przy pomocy splajnów z przestrzeni $\mathbf{S_3^2}(\pi)$.

Najprostszym zadaniem interpolacyjnym dla przestrzeni $\mathbf{S_3^2}(\pi)$ jest następujące zadanie typu Lagrange'a:

Niech będzie dany równomierny podział π odcinka $[a, b]$:

$$\pi : \qquad a = x_0 < x_1 < \cdots < x_N = b$$

$$x_j = x_0 + jh, \quad h = \frac{b-a}{N}, \quad j = 0, 1, \cdots, N.$$

Dla danej funkcji $f \in C([a, b])$, posiadającej pierwsze pochodne (jednostronne) określone w punktach a i b, poszukujemy **splajnu interpolacyjnego** $s \in \mathbf{S_3^2}(\pi)$ *spełniającego następujące warunki:*

- $s'(x_0) = f'(x_0)$ *- warunek brzegowy,*
- $s(x_j) = f(x_j)$ *dla* $j = 0, 1, \cdots, N$ *- warunki interpolacji,*
- $s'(x_N) = f'(x_N)$ *- warunek brzegowy.*

Komentarz. Warunków interpolacji jest tylko $N + 1$, zaś

$$\dim(\mathbf{S_3^2}(\pi)) = N + 3.$$

Zatem samych warunków interpolacji nie wystarcza do jednoznacznego wyznaczenia splajnu interpolacyjnego s. Dlatego dodane są dwa warunki brzegowe.

Poniższe twierdzenie o istnieniu i jednoznaczności **definiuje jednocześnie dobry algorytm wyznaczania splajnu interpolacyjnego**.

Twierdzenie 1.7 *Zadanie interpolacyjne (1.20) ma zawsze jednoznaczne rozwiązanie.*

Dowód. (Uwaga: dowód zawiera dobry numerycznie algorytm wyznaczania splajnu interpolacyjnego $s \in \mathbf{S_3^2}(\pi)$).

Ponieważ

$$s(x_k) = \sum_{j=-1}^{N+1} c_j B_j(x_k) \quad k = 0, 1, \cdots, N,$$

$$s'(x_0) = \sum_{j=-1}^{N+1} c_j B_j'(x_0),$$

$$s'(x_N) = \sum_{j=1}^{N+1} c_j B_j'(x_N),$$

to wykorzystując tablicę wartości funkcji B_j i B_j', otrzymamy następujący układ równań algebraicznych liniowych, z którego możemy wyznaczyć współczynniki

$$c_{-1}, c_0, c_1, \cdots, c_N, c_{N+1}.$$

(1.21)
$$Ac = f,$$

gdzie

$$c = [c_{-1}, c_0, c_1, \cdots, c_N, c_{N+1}]^T,$$

$$f = [f'(x_0), f(x_0), f(x_1), \cdots, f(x_N), f'(x_N)]^T,$$

(1.22)
$$A = \begin{bmatrix} -3/h & 0 & 3/h & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & 0 & 0 & \cdots & 0 & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & -3/h & 0 & 3/h \end{bmatrix},$$

Układ równań (1.21) może z powodzeniem służyć do wyznaczania splajnu interpolacyjnego $s \in \mathbf{S_3^2}(\pi)$. Zauważmy od razu, że macierz A jest zupełnie inna niż w przypadku interpolacji wielomianowej.

Zadanie 1.10 Wykorzystując podane niżej **Twierdzenie Gershgorina** udowodnij, że macierz A jest nieosobliwa.
Zadanie 1.11 Oszacuj współzynnik uwarunkowania A.
Zadanie 1.12 Napisz program znajdujący splajn interpolacyjny, oraz wyliczający jego wartości w zadanych punktach. Zadbaj o optymalność.

Istnienie i jednoznaczność rozwiązania układu (1.21) jest równoznaczne z istniniem jedynego splajnu interpolacyjnego. □
Twierdzenie Gershgorina [2] *Niech $A = (a_{ij})_{i,j=1,2,\cdots,n}$ będzie macierzą kwadratową o elementach zespolonych.*

Wszystkie wartości własne macierzy A mieszczą się w zbiorze

$$\Lambda = \cup_{j=1}^{n} K_j \subset \mathbf{C}$$

leżącym na płaszczyźnie zespolonej $\mathbf{C}$, przyczym

$$K_j = \{z \in \mathbf{C}| \quad |z - a_{jj}| \leq \sum_{i=1, i\neq j}^{n} |a_{ij}|, j = 1, 2, \cdots, n\}$$

Jeśli zbiór Λ jest niespójny, to każda z jego składowych zawiera wartości własne macierzy A.

Na zakończenie podamy pewne oszacowania błędu interpolacji splajnowej.

1. Jeśli $f \in C^2([a, b])$ i s jest splajnem interpolacyjnym, to

$$\|f - s\|_2 \leq 8h^2 \|f''\|_2,$$

$$\|f' - s'\|_2 \leq 4h \|f''\|_2,$$

$$\|f'' - s''\|_2 \leq \|f''\|_2.$$

[2] Patrz książka Gantmachera "Matrix theory". Oryginał rosyjski jest w bibliotece WMIM

2. Jeśli $f \in C^4([a,b])$ i s jest splajnem interpolacyjnym, to

$$\|f - s\|_2 \leq 64h^4 \|f''''\|_2,$$

$$\|f' - s'\|_2 \leq 32h^3 \|f''''\|_2,$$

$$\|f'' - s''\|_2 \leq 8h^2 \|f''''\|_2.$$

Wszystkie normy w powyższych wzorach są normami z przestrzeni $L^2([a,b])$. Dowód jest w cytowanej juz książce: P.M. Prenter "Splines and Variational Methods".

DFT DYSKRETNA TRANSFORMATA FOURIERA (Discrete Fourier Transform)

Dyskretna transformata Fouriera - w skrócie DFT - jest ważnym narzędziem mającym liczne zastosowania. Typowym przykładem zastosowania DFT jest przetwarzanie sygnałów. DFT jest blisko "spokrewniona" z interpolacją trygonometryczną. DFT przekształca ciągi liczb zespolonych na inne takie ciągi.

Niech będzie dany ciąg skończony

$$\underline{u} = \{u_0, u_1, \cdots, u_{N-1}\}.$$

Będziemy zawsze zakładać, że nasz ciąg jest *przedłużony w obie strony* w sposób periodyczny, to znaczy, że dla każdego k

$$u_k = u_{N+k}.$$

W wyniku zastosowania DFT do tego ciągu otrzymamy inny ciąg

$$\underline{\hat{u}} = \{\hat{u}_0, \hat{u}_1, \cdots, \hat{u}_{N-1}\}$$

gdzie

$$\hat{u}_k = \frac{1}{N} \sum_{j=0}^{N-1} e^{-i\frac{2\pi}{N}kj} u_j.$$

Czasem jest wygodnie używać takiego oznaczenia:

$$\hat{u}_k = (\hat{\underline{u}})_k.$$

Jak interpretować wynik transformaty? Kolejnym funkcjom wykładniczym zmiennej całkowitej j

$$\phi_k(j) = e^{-i\frac{2\pi}{N}jk} \quad k = 0, 1, 2 \cdots, N-1$$

możemy przyporządkować kolejne *częstotliwości* które one ze sobą niosą. Każdej z rozważanych funkcji wykładniczych przyporządkujemy *częstotliwość* reprezentowaną przez liczbę okresów tej funkcji, które mieszczą się w zakresie indeksów (argumentów) $0 \leq j \leq N$.

Aby zorientować się w sytuacji, rozpatrzmy przypadek, gdy $N = 4$. Wartości funkcji $\phi_k(j)$ dla różnych k i j podaje poniższa tablica.

	j=0	j=1	j=2	j=3	j=4	liczba okresów
k=0	1	1	1	1	1	constans
k=1	1	i	-1	-i	1	1
k=2	1	-1	1	-1	1	2
k=3	1	-i	-1	i	1	1
k=4	1	1	1	1	1	constans

Widać stąd, że maksymalną częstotliwość niesie funkcja

$$\phi_2(j) = \phi_{\frac{N}{2}}(j).$$

Ogólnie można powiedzieć, że maksymalne częstotliwości znajdują się *w okolicy* $\frac{N}{2}$, gdyż N nie zawsze jest parzyste. Kolejne elementy transformaty DFT $\hat{u}_k$ są przyporządkowane kolejnym funkcjom ϕ_k i mówią o *udziale* odpowiadających im częstotliwości w ciągu $\{u_0, u_1, \cdots, u_{N-1}\}$, gdyż są to współczynniki Fouriera dla tego ciągu.

Transformatą odwrotną ciągu

$$\underline{u} = \{u_0, u_1, \cdots, u_{N-1}\}.$$

jest ciąg

$$\underline{\check{u}} = \{\check{u}_0, \check{u}_1, \cdots, \check{u}_{N-1}\},$$

gdzie

$$\check{u}_k = \sum_{j=0}^{N-1} e^{i\frac{2\pi}{N}kj} u_j.$$

Zadanie 1.13. Udowodnij, że $\check{\hat{\underline{u}}} = \underline{u}$.
Wskazówka: udowodnij najpierw, że $\sum_{s=0}^{N-1} e^{-i\frac{2\pi}{N}(k-j)s} = \begin{cases} 0 & \text{gdy} \quad k \neq j \\ N & \text{gdy} \quad k = j \end{cases}$

Przesunięcie. Niech dla całkowitego p

$$\underline{u_{\cdot\ +p}} = \{u_p, u_{1+p}, u_{2+p}, \cdots, u_{N-1+p}\}.$$

Jest to ciąg $\underline{u}$ *przesunięty o p*.

Zadanie 1.14. Udowodnij, że

$$\left(\hat{\underline{u_{\cdot\ +p}}}\right)_k = e^{i\frac{2\pi}{N}pk}\hat{u}_k.$$

Norma. Niech $\|\underline{u}\|_0^2 = \frac{1}{N}\sum_{j=0}^{N-1} |u_j|^2$.

Zadanie 1.15. Udowodnij, że

$$\|\hat{\underline{u}}\|_0 = \frac{1}{\sqrt{N}}\|\underline{u}\|_0.$$

Splot. Splotem dwóch ciągów

$$\underline{u} = \{u_0, u_1, \cdots, u_{N-1}\}.$$

$$\underline{v} = \{v_0, v_1, \cdots, v_{N-1}\}.$$

nazywamy ciąg

$$(\underline{u} \star \underline{v})_k = \sum_{j=0}^{N-1} u_{k-j} v_j.$$

Zadanie 1.16. Udowodnij następujące własności splotu:

1. $\underline{u} \star \underline{v} = \underline{v} \star \underline{u}$.
2. Niech $\underline{u} \cdot \underline{v} = \{u_0 v_0, u_1 v_1, \cdots, u_{N-1} v_{N-1}\}$. Wtedy $\underline{u} \,\hat{\cdot}\, \underline{v} = \underline{\hat{u}} \star \underline{\hat{v}}$.
3. Udowodnij, że $(\underline{\hat{u}} \,\check{\cdot}\, \underline{\hat{v}}) = \frac{1}{N}(\underline{u} \star \underline{v})$.

FILTRY. Zadanie 1.16 p.3 można wykorzystać do budowy *filtrów*. Na przykład filtr wycinający najwyższe częstotliwości można zbudować tak. Oznaczmy

$$\underline{\hat{H}} = \{\hat{H}_0, \hat{H}_1, \cdots, \hat{H}_{N-1}\},$$

gdzie

$$\hat{H}_s = \begin{cases} 1 & \text{gdy} \quad s = 1, 2, \cdots, p-1 \\ 0 & \text{gdy} \quad s = p, p+1, \cdots, N-p-1 \\ 1 & \text{gdy} \quad s = N-p, N-p+1, \cdots, N-1 \end{cases}.$$

Ciąg

$$\underline{\hat{u}} \cdot \underline{\hat{H}}$$

to ciąg $\underline{\hat{u}}$ pozbawiony wyrazów o *wysokich częstotliwościach,* które mieszczą się w przedziale indeksów $[p, N-p+1]$, (trzeba tu założyć, że $0 \leq p < \frac{N+1}{2}$). Odfiltrowany ciąg oryginalny, to

$$\frac{1}{N}(\underline{u} \star \underline{H}).$$

Łatwo znaleźć $\underline{H}$:

$$H_k = \sum_{s=0}^{p-1} e^{i\frac{2\pi}{N}sk} + \sum_{s=N-p}^{N-1} e^{i\frac{2\pi}{N}sk}.$$

Oczywiście można budować różne inne filtry, bardziej wyrafinowane niż filtr pokazany powyżej.

Zadanie 1.17. Znajdź odfiltrowany ciąg oryginalny

$$\frac{1}{N}(\underline{u} \star \underline{H}).$$

Znajdź odfiltrowany ciąg innym sposobem, jako $(\underline{\hat{u}} \,\check{\cdot}\, \underline{\hat{H}})$.

Rozdział 2
METODY PRZESTRZENI HILBERTA.

Aproksymacja w przestrzeni unitarnej.

Zajmiemy się teraz zagadnieniem aproksymacji w *przestrzeniach unitarnych.* **Przestrzeń unitarna** to taka przestrzeń liniowa H nad ciałem $\mathbf{R}$, (przestrzeń unitarna rzeczywista), lub nad ciałem $\mathbf{C}$, (przestrzeń unitarna zespolona), w której jest określony *iloczyn skalarny*:

$$(\cdot,\cdot) : H \times H \to \mathbf{R},$$

gdy przestrzestrzeń jest rzeczywista,

$$(\cdot,\cdot) : H \times H \to \mathbf{C},$$

gdy przestrzeń jest zespolona. Iloczyn skalarny jest funkcją dwóch zmiennych, liniową względem pierwszego argumentu:

$$(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z),$$

i antysymetryczną:

$$(x, y) = \overline{(y, x)}.^3$$

Ponadto $(x, x) \geq 0$ dla każdego elementu x przestrzeni H, zaś $(x, x) = 0$ jedynie, gdy $x = 0$. Te ostatnie warunki pozwalają określić *normę* $\|x\| = \sqrt{(x, x)}$. Przestrzeń unitarna, która jest *zupełna* nazywa się *przestrzenią Hilberta.*

Działając w przestrzeni unitarnej, gdzie norma jest indukowana przez iloczyn skalarny, **otrzymujemy dodatkowe narzędzie, którego nie mieliśmy dotychczas: iloczyn skalarny, a co za tym idzie, pojęcie ortogonalności.**

Zadanie 2.1 Udowodnij, że każda przestrzeń unitarna jest *silnie unormowana*, to znaczy, że warunek $\|x + y\| = \|x\| + \|y\|$ zachodzi wtedy i tylko wtedy, gdy istnieje stała $\alpha \geq 0$ taka, że $y = \alpha x$.

Wiemy już, że w dowolnej przestrzeni unormowanej, w jej podprzestrzeni skończonego wymiaru istnieje co najmniej jeden element najlepszej aproksymacji dla dowolnego punktu tej przestrzeni.

[3] Jeśli H jest przestrzenią rzeczywistą - to jest to symetria: $(x, y) = (y, x)$.

Twierdzenie 2.1 *Jeśli przestrzeń H jest silnie unormowana - na przykład, gdy jest przestrzenią unitarną, element najlepszej aproksymacji w dowolnej podprzestrzeni $V \subset H$ jest jednoznacznie wyznaczony.*

Dowód. Przypuśćmy, że tak nie jest, i że dla elementu $x \in H$ w podprzestrzeni V, istnieją *dwa różne* elementy najlepszej aproksymacji v_1 i v_2. Odrazu zauważmy, że wtedy napewno $x \notin V$. Niech $\|x - v_1\| = \|x - v_2\| = e$. Wtedy

$$\|x - \frac{v_1 + v_2}{2}\| = \frac{1}{2}\|(x - v_1) + (x - v_2)\| \leq \frac{1}{2}(\|x - v_1\| + \|x - v_2\|) = e,$$

i ponieważ odległość x i żadnego elementu V *nie może być mniejsza od* e, widzimy, że

$$\|(x - v_1) + (x - v_2)\| = \|x - v_1\| + \|x - v_2\| = 2e.$$

Ponieważ przestrzeń H jest *silnie unormowana*, to istnieje $\alpha \geq 0$, że $x - v_2 = \alpha(x - v_1)$. Zauważmy odrazu, że $\alpha \neq 1$, bo w przeciwnym wypadku musiałoby być $v_1 = v_2$. Stąd $x = \frac{v_1 - \alpha v_2}{1-\alpha}$, co oznacza że x jest kombinacją liniową elementów z V, a więc $x \in V$, co nie jest możliwe. □

Niech znów $V \subset H$, będzie podprzestrzenią H i niech $x \in H$. Element $v_0 \in V$ nazywa się *rzutem ortogonalnym x na V* jeśli

$$(x - v_0, v) = 0 \text{ dla każdego } v \in V.$$

Wiadomo, że jeśli H jest przestrzenią Hilberta i $V = \bar{V}$ (podprzestrzeń V jest domknięta), to dla każdego $x \in H$ istnieje rzut ortogonalny na V. My *skonstruujemy rzut ortogonalny dla x*, w przypadku, gdy $\dim(V) < \infty$, $V = span\{\phi_1, \phi_2, \cdots, \phi_n\}$, gdzie układ $\{\phi_1, \cdots, \phi_n\}$ jest liniowo niezależny.

Niech v_0 będzie szukanym rzutem ortogonalnym elementu $x \in H$ na podprzestrzeń V. Z warunku ortogonalności otrzymamy następujące równania:

$$(x - v_0, \phi_k) = 0 \quad \text{dla} \quad k = 1, 2, \cdots, n.$$

Ponieważ $v_0 \in V$, to $v_0 = \sum_{j=1}^{n} \phi_j c_j$, to ostatecznie

$$\sum_{j=1}^{n} (\phi_j, \phi_k) c_j = (x, \phi_k) \qquad k = 1, 2, \cdots, n. \tag{2.1}$$

Układ równań liniowych algebraicznych (2.1) zapiszemy w postaci macierzowej:

$$G\underline{c} = \underline{x}, \tag{2.2}$$

gdzie $G = (g_{k,j})_{k,j=1,2\cdots,n}$, $g_{k,j} = (\phi_j, \phi_k)$ nazywa się *macierzą Gramma*, $\underline{c} = [c_1, c_2, \cdots, c_n]^T$ jest szukanym wektorem, zaś $\underline{x} = [(x, \phi_1), \cdots, (x, \phi_n)]^T$.

Zadanie 2.2 Udowodnij, że macierz Gramma jest *nieosobliwa* i dodatnio określona, jeśli układ $\{\phi_1, \phi_2, \cdots, \phi_n\}$ jest liniowo niezależny. Ponadto $G = G^*$.

Układ (2.2) nazywa się *układem równań normalnych*, i ma jednoznaczne rozwiązanie. Nie zawsze jednak rozwiązywanie tego układu jest dobrym algorytmem wyznaczania *rzutu ortogonalnego*. Dlaczego tak może być - wyjaśnimy dalej.

Na szczególną uwagę zasługuje przypadek, gdy *baza* $\phi_1, \phi_2, \cdots, \phi_n$ *jest ortogonalna*. Wtedy macierz G jest diagonalna i na diagonali ma kolejno elementy $\|\phi_1\|^2, \|\phi_2\|^2, \cdots, \|\phi_n\|^2$, zaś rozwiązanie jest postaci

$$c_k = \frac{(x, \phi_k)}{\|\phi_k\|^2}, \quad \text{dla} \quad k = 1, 2, \cdots, n. \tag{2.3}$$

Współczynniki c_k, to *współczynniki Fouriera elementu x względem bazy*

$$\phi_1, \phi_2, \cdots, \phi_n.$$

Przedstawienie rzutu ortogonalnego v_0 jako

$$v_0 = \sum_{j=1}^{n} c_j \phi_j = \sum_{j=1}^{n} \frac{(x, \phi_j)}{\|\phi_j\|^2} \phi_j \tag{2.4}$$

nazywamy *rozwinięciem Fouriera elementu x, względem bazy ortogonalnej* $\phi_1, \phi_2, \cdots, \phi_n$. Przypomnijmy, że z takim rozwinięciem spotkaliśmy się już przy omawianiu *interpolacji trygonometrycznej*.

Twierdzenie 2.2 *Rzut ortogonalny elementu $x \in H$ na podprzestrzeń V (jeśli istnieje), jest elementem najlepszej aproksymacji dla x w V.*

Dowód. Niech $v \in V$ będzie dowolnym elementem, zaś v_0, rzutem ortogonalnym x na V. Wtedy możemy napisać $v = v_0 + w$, gdzie $w \in V$, i

$$\|x-v\|^2 = (x-v_0-w, x-v_0-w) = \|x-v_0\|^2 + (x-v_0, w) + (w, x-v_0) + \|w\|^2 =$$

$$= \|x - v_0\|^2 + \|w\|^2,$$

gdyż v_0 jest rzutem ortogonalnym x. Stąd, oczywiście $\|x-v\| \geq \|x-v_0\|$, co oznacza, że v_0 jest elementem najlepszej aproksymacji dla x w V, ponieważ $v \in V$ jest dowolny. □

Z tego twierdzenia wynika, że rzut ortogonalny, jeśli istnieje, to jest wyznaczony jednoznacznie.

Przykład. Niech A będzie macierzą *prostokątną* o m-wierszach i n-kolumnach, gdzie $m > n$.

$$A = [a_1, a_2, \cdots, a_n],$$

gdzie

$$a_j = [a_{1j}, a_{2j}, \cdots, a_{mj}]^T$$

jest $j-$tą kolumną macierzy A. Niech $b = [b_1, b_2, \cdots, b_m]^T$ będzie wektorem, $b \in \mathbf{R^m}$. Poszukujemy wektora $x = [x_1, x_2, \cdots, x_n]^T$, takiego aby

$$\|b - Ax\|^2 = \min_{x \in \mathbf{R^n}} . \tag{2.5}$$

Zadanie (2.5), to *liniowe zadanie najmniejszych kwadratów - w skrócie LZNK.* Zadanie to możemy interpretować jako *poszukiwanie elementu najlepszej aproksymacji w podprzestrzeni span*$\{a_1, a_2, \cdots, a_n\}$, *dla wektora* $b \in \mathbf{R^m}$. Użyta tu norma, to norma *euklidesowa* $\|b\|^2 = \sum_{j=1}^m b_j^2$. Wypiszmy *układ równań normalnych* dla tego zadania:

$$A^T Ax = A^T b. \tag{2.6}$$

Jest to układ n równań liniowych z n niewiadomymi. Jeśli macierz A jest rzędu n ($rank(A) = n$, n - maksymalny możliwy rząd!), to macierz $A^T A$ jest nieosobliwa, i układ jest jednoznacznie rozwiązalny. Zauważmy, że warunek $rank(A) = n$ oznacza, że wektory $a_1, a_2, \cdots, a_n$ stanowią układ liniowo niezależny. Wyobraźmy sobie teraz, że $n = m$. Wtedy maierz A jest kwadratowa, i przy założeniu, że $rank(A) = n$, jest nieosobliwa. Załóżmy dodatkowo, że $A^T = A$, i weźmy pod uwagę dwa układy:

$$Ax = b,$$

(teraz ten układ jest jednoznacznie rozwiązalny - nie ma zatem potrzeby odwoływania się do zadania LZNK!). Drugi układ, to (2.6):

$$A^T A x = A^T b.$$

Nie trudno zauważyć, że *współczynnik uwarunkowania* dla naszej normy, dla macierzy A wynosi

$$cond(A) = \|A\|\|A^{-1}\| = \frac{|\lambda_{max}|}{|\lambda_{min}|},$$

gdzie λ_{max} i λ_{min} to odpoweiednio, wartości własne A o maksymalnym i minimalnym module. (Zastanów się - dlaczego tak jest!) Dla drugiego układu otrzymujemy natomiast

$$cond(A^T A) = cond(A^2) = (\frac{\lambda_{max}}{\lambda_{min}})^2.$$

Oba układy są równoważne, zaś *współczynnik uwarunkowania drugiegiego z nich,* **jest kwadratem współczynnika uwarunkowania pierwszego**. Gdy współczynnik uwarunkowania A jest duży - to współczynnik uwarunkowania $A^T A$ może okazać się *ogromny*, co może, w najlepszym razie poważnie utrudnić rozwiązywanie numeryczne tego drugiego zadania. Te wszystkie rozważania nie dotyczą oczywiście maleńkich zadań, gdzie wynik możemy wyliczyć "odręcznie, na papierze". Widać stąd potrzebę znalezienia innego wyjścia dla zagadnień LZNK, (a ogólnie, dla poszukiwania rzutu ortogonalnego), nie opartego na rozwiązywaniu układu normalnego. Dla niektórych zadań LZNK stosuje się często, *algorytm tak zwanego rozkładu "QR" macierzy A*. O tym algorytmie będzie jeszcze mowa w dalszej części tego rozdziału.

Operator rzutu ortogonalnego.

Niech $V \subset H$ będzie podprzestrzenią przestrzeni H. Załóżmy, że Dla każdego $x \in H$ istnieje rzut ortogonalny na V. Wtedy operator P

$$P : H \to V,$$

przyporządkowujący *elementom H ich rzuty ortogonalne na V* jest dobrze określony. Nie trudno sprawdzić, że P jest operatorem liniowym na H i że

$$PP = P. \tag{2.7}$$

Niech teraz x i y będą dwoma dowolnymi elementami H. Mamy:

$$(Px, y) = (Px, Py + y - Py) = (Px, Py),$$

gdyż $(Px, y - Py) = 0$, bo Py jest rzutem ortogonalnym elementu y. Dalej:

$$(Px, y) = (Px, Py) = (Px - x + x, Py) = (x, Py),$$

ponieważ $(Px - x, Py) = 0$, gdyż Px jest rzutem ortogonalnym elementu x, oraz $Py \in V$. Udowodniliśmy więc, że

$$(Px, y) = (x, Py). \tag{2.8}$$

Równość (2.8) oznacza, że P jest *operatorem samosprzężonym*, czyli *jest równy swojemu operatorowi sprzężonemu:*

$$P = P^*.$$

Ostatecznie możemy napisać, ze *operator rzutu ortogonalnego, to taki operator liniowy $P : H \to H$, że*

$$P = PP = P^*.$$

Zadanie 2.3

- Udowodnij, że warunki $P = PP = P^*$ charakteryzują operator rzutu ortogonalnego z H na PH.
- Niech H będzie przestrzenią Hilberta, zaś

  $$V = span\{\phi_1, \phi_2, \cdots, \phi_n\},$$

  gdzie elementy ϕ_j, $j = 1, 2, \cdots, n$ są liniowo niezależne. Udowodnij, że:

  1. Każdy operator liniowy $P : H \to_{na} V$ jest postaci $Px = \sum_{j=1}^{n} \phi_j(x, \psi_j)$, gdzie ψ_j, $j = 1, 2, \cdots, n$ jest pewnym układem liniowo niezależnym w H.
  2. Operator P^*, sprzężony do P, jest postaci $P^*(x) = \sum_{j=1}^{n} \psi_j(x, \phi_j)$.

3. P jest rzutem ($PP = P$) wtedy i tylko wtedy, gdy bazy $\{\phi_1, \cdots, \phi_n\}$ i $\{\psi_1, \cdots, \psi_n\}$ są względem siebie *biortonormalne* - to znaczy, że

$$(\phi_k, \psi_l) = \delta_{k,l}.$$

4. Rzut P jest rzutem ortogonalnym na V, wtedy i tylko wtedy, gdy

$$span\{\phi_1, \cdots, \phi_n\} = span\{\psi_1, \cdots, \psi_n\}.$$

Zadanie 2.4 Skonstruuj rzut ortogonalny $P : H \to_{na} V = span\{\phi\}$.

Algorytm Gramma-Schmidt'a

Ten dobrze znany algorytm wykonuje następujące zadanie:

Dany jest w przestrzeni *rzeczywistej* Hilberta H układ liniowo niezależny

$$x_1, x_2, \cdots, x_n.$$

Należy skonstruować układ *ortonormalny*

$$q_1, q_2, \cdots, q_n$$

taki, że dla każdego k, $k = 1, 2, \cdots, n$

$$span\{x_1, x_2, \cdots, x_k\} = span\{q_1, q_2, \cdots, q_k\}.$$

Przypomnimy najpierw *wersję klasyczną tego algorytmu.*

Algorytm G-S K

- Definiujemy

$$p_1 = x_1,$$
$$q_1 = \frac{p_1}{\|p_1\|}$$

stąd

$$x_1 = \alpha_{1,1} q_1, \quad \text{gdzie} \quad \alpha_{1,1} = \|x_1\|.$$

- Mamy już $q_1, q_2, \cdots, q_{k-1}$, o żądanych własnościach. Określimy:

$$p_k = x_k - \sum_{j=1}^{k-1} \alpha_{k,j} q_j, \tag{2.9}$$

gdzie $(p_k, q_j) = 0$ dla $j = 1, 2, \cdots, k-1$. Z tych warunków wynika, że

$$\alpha_{k,j} = (x_k, q_j) \text{ dla } j = 1, 2, \cdots, k-1.$$

Teraz określamy

$$q_k = \frac{p_k}{\|p_k\|}.$$

Stąd

$$x_k = \sum_{j=1}^{k} \alpha_{k,j} q_j,$$

gdzie

$$\alpha_{k,j} = (x_k, q_j) \text{ dla } j = 1, 2, \cdots, k-1,$$

zaś

$$\alpha_{k,k} = \|p_k\| = (\|x_k\|^2 - \sum_{j=1}^{k-1} \alpha_{k,j}^2)^{\frac{1}{2}}.$$

Zadanie 2.5 Udowodnij, że jeśli układ $x_1, x_2, \cdots, x_n$ jest liniowo niezależny, to algorytm G-S K generuje ciąg $q_1, q_2, \cdots, q_n$ o żądanych własnościach.

Zadanie 2.6 Niech $H = \mathbf{R^n}$ i oznaczmy przez A macierz, której kolumnami są liniowo niezależne wektory $x_1, x_2, \cdots, x_n$. Udowodnij, że algorytm G-S K można zapisać tak:

$$A = QR,$$

gdzie Q jest macierzą ortogonalną, $Q = [q_1, q_2, \cdots, q_n]$, zaś

$$R^T = \begin{bmatrix} \alpha_{1,1} & 0 & 0 & 0 & \cdots & 0 \\ \alpha_{2,1} & \alpha_{2,2} & 0 & 0 & \cdots & 0 \\ \alpha_{3,1} & \alpha_{3,2} & \alpha_{3,3} & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{n,1} & \alpha_{n,2} & \alpha_{n,3} & \cdots & \cdots & \alpha_{n,n} \end{bmatrix}$$

Jest to tak zwany rozkład QR macierzy A - rozkład na na iloczyn macierzy ortogonalnej i trójkątnej górnej.

Okazuje się, że algorytm G-S K jest bardzo niedobry pod względem numerycznym: błędy zaokrągleń mogą po nawet nie wielkiej liczbie kroków sprawić, że obliczone wektory $q_1, q_2, \cdots, q_k$ zatracą ortogonalność. Można tę wadę w znacznej mierze wyelimiować, stosując *Poprawiony Algorytm Gramma - Schmid'a* G-S P. Założymy teraz, że $H = \mathbf{R}^{\mathbf{m}}$, $m \geq n$.

Aby zdefiniować algorytm G-S P zapiszemy najpierw wzór (2.9) w nieco innej, równoważnej postaci

$$p_k = x_k - \sum_{j=1}^{k-1}(x_k, q_j)q_j = x_k - \sum_{j=1}^{k-1} q_j q_j^T x_k = x_k - \sum_{j=1}^{k-1} Q_j x_k = (I - \sum_{j=1}^{k-1} Q_j)x_k,$$

gdzie $Q_j = q_j q_j^T$ jest macierzą kwadratową wymiaru $n \times n$. [4]

Zadanie 2.7 Uwaga! zrobienie tego zadania jest ważne dla zrozumienia algorytmu G-S P! Sprawdź, że:

- Macierze $Q_1, Q_2, \cdots, Q_n$ stanowią *układ rzutów ortogonalnych, i wzajemnie do siebie ortogonalnych.* To znaczy, że
  1. $Q_i Q_j = Q_j Q_i = \delta_{i,j} Q_j$,
  2. $Q_j = Q_j^*$ dla $\quad j = 1, 2, \cdots, n$,
  3. $Q_j : H \to span\{q_j\}$ jest to rzut ortogonalny na podprzestrzeń jednowymiarową!
- $I - \sum_{j=1}^{k} Q_j = (I - Q_1)(I - Q_2) \cdots (I - Q_k)$ dla $k = 1, 2, \cdots, n$; ponadto poszczególne czynniki komutują.

Wykorzystując powyższe zadanie wnioskujemy, że

$$p_k = (I - Q_{k-1})(I - Q_{k-2}) \cdots (I - Q_1)x_k.$$

Teraz określimy nowe wektory:

$$p_{k,1} = x_k,$$

[4] Pamiętamy, że wektory, to macierze o jednej kolumnie, i że stosujemy tu reguły mnożenia macierzy!

$$p_{k,j+1} = (I - Q_j)p_{k,j}, \quad \text{dla} \quad j = 1, 2, \cdots, k-1,$$

$$p_k = p_{k,k}.$$

Zaważmy, że:

- $(p_{k,j}, q_j) = ((I - Q_{j-1})p_{k,j-1}, q_j) = (p_{k,j-1}, q_j) - (Q_{j-1}p_{k,j-1}, q_j) = (p_{k,j-1}, q_j) - (p_{k,j-1}, Q_{j-1}q_j) = (p_{k,j-1}, q_j)$,

  więc stąd wynika, że

  $(p_{k,j}, q_j) = (p_{k,j-1}, q_j) = \cdots = (p_{k,1}, q_j) = (x_k, q_j) = \alpha_{k,j}$,

- $p_{k,j+1} = p_{k,j} - Q_j p_{k,j} = p_{k,j} - q_j(p_{k,j}, q_j) = p_{k,j} - q_j \alpha_{k,j}$ dla $j = 1, 2, \cdots, k-1$.

Możemy teraz zdefiniować poprawiony algorytm Gramma-Schmidt'a G-S P.

Algorytm G-S P.

- Określamy

  $$p_{1,1} = x_1$$

  oraz

  $$q_1 = \frac{p_{1,1}}{\|p_{1,1}\|}.$$

- Już mamy:

  $$q_1, q_2, \cdots, q_{k-1},$$

  oraz

  $$\begin{array}{cccc} \alpha_{1,1} & & & \\ \alpha_{2,1} & \alpha_{2,2} & & \\ \cdots & \cdots & \cdots & \\ \alpha_{k-1,1} & \alpha_{k-1,2} & \cdots & \alpha_{k-1,k-1} \end{array}$$

  Obliczamy współczynniki $\alpha_{k,j}$ i wektory $p_{k,j}$ dla $j = 1, 2, \cdots, k$:

  $$p_{k,1} = x_k, \quad \alpha_{k,1} = (p_{k,1}, q_1),$$

  $$\cdots$$

  $$\alpha_{k,j} = (p_{k,j}, q_j), \quad p_{k,j+1} = p_{k,j} - \alpha_{k,j} q_j,$$

$$\cdots$$

$$\alpha_{k,k-1} = (p_{k,k-1}, q_{k-1}), \quad p_{k,k} = p_{k,k-1} - \alpha_{k,k-1} q_{k-1}.$$

Wyliczamy teraz kolejny wektor q_k:

$$q_k = \frac{p_{k,k}}{\|p_{k,k}\|} \text{ i } \alpha_{k,k} = \|p_{k,k}\|.$$

Sprobujmy odpowiedzieć, dlaczego ta wersja algorytmu Grammma - Schmidt'a jest numerycznie lepsza od G-S K. Przyczyna leży w sposobie liczenia współczynników $\alpha_{k,j}$, $j = 1, 2, \cdots, k$.

W wersji klasycznej (G-S K)	W wersji poprawionej (G-S P)
$\alpha_{k,j} = (x_k, q_j) = (p_{1,1}, q_j)$	$\alpha_{k,j} = (p_{k,j}, q_j)$

Gdybyśmy mogli wykonywać obliczenia, używając arytmetyki "prawdziwej", obie wersje niczym by się nie różniły. Błąd numeryczny przy obliczaniu iloczynu skalarnego $\alpha_{k,j} = (x_k, q_j)$ jest *tym większy, im większe normy mają czynniki.* Najłatwiej to wyjaśnić obserwując błąd iloczynu dwóch liczb a i b. Ich reprezentacje w arytmetyce komputerowej to $a(1 + \epsilon_a)$ i $b(1 + \epsilon_b)$. Stąd mamy błąd iloczynu $\Delta = |a(1 + \epsilon_a)b(1 + \epsilon_b) - ab| = |a||b||(\epsilon_a + \epsilon_b + \epsilon_a\epsilon_b)|$. Jest on proporcjonalny do $|a||b|$. Iloczyn skalarny zachowuje się analogicznie. Czynnik q_j ma normę równą 1, zatem wszystko zależy od normy $x_k = p_{1,1}$ lub $p_{k,j}$. W algorytmie G-S K mamy zawsze x_k, podczas, gdy w G-S P występują wektory $p_{k,j}$. Obliczymy kwadrat normy $\|p_{k,j}\|^2$. Mamy:

$$p_{k,j} = p_{k,j-1} - q_{j-1}\alpha_{k,j-1},$$

$$\begin{aligned}\|p_{k,j}\|^2 &= (p_{k,j-1} - q_{j-1}\alpha_{k,j-1}, p_{k,j-1} - q_{j-1}\alpha_{k,j-1}) = \\ &= \|p_{k,j-1}\|^2 - \alpha^2_{k,j-1} = \|p_{k,j-2}\|^2 - \alpha^2_{k,j-2} - \alpha^2_{k,j-1} = \\ &= \cdots = \|p_{k,1}\|^2 - \alpha^2_{k,1} - \alpha^2_{k,2} - \cdots - \alpha^2_{k,j-1} < \|p_{k,1}\|^2 = \|x_k\|^2.\end{aligned}$$

Zatem zawsze, gdy $j > 1$ jest $\|p_{k,j}\|^2 < \|x_k\|^2$.

Powróćmy jeszcze na chwilę do zadania LZNK

$$\|Ax - b\|^2 = Min.$$

Zauważyliśmy już, że rozwiązywanie *układu normalnego* może nie być najlepszym sposobem. Pokażemy tu inny sposób nie odwołujący się do macierzy $A^T A$. Przypuśćmy, że kolumnami macierzy A są liniowo niezależne wektory

$$a_1, a_2, \cdots, a_n,$$

należące do przestrzeni $\mathbf{R^m}$, $m \geq n$. Mówimy wtedy, że zadanie LZNK jest *regularne*. Dokonajmy rozkładu "QR" macierzy A. Można to zrobić przy pomocy algorytmu G-S P, zastosowanego do kolumn macierzy A. Otrzymamy zadanie

$$\|QRx - b\|^2 = Min,$$

gdzie Q jest macierzą o n kolumnach *ortonormalnych*, zaś R jest macierzą *trójkątną górną* wymiaru $n \times n$. Oznaczmy teraz $y = Rx$. W ten sposób nasze zadanie sprowadziło się do

$$\|Qy - b\| = Min,$$

czyli do wyznaczenia rzutu ortogonalnego wektora b na podprzestrzeń generowaną przez n ortonormalnych kolumn macierzy Q. Współrzędnymi wektora y są więc *współczynniki fourierowskie* wektora b względem bazy kolumn macierzy Q. To znaczy:

$$y = Q^T b.$$

Ponieważ jednak szukamy wektora x, nie wektora y, to ostatecznie musimy rozwiązać układ z macierzą trójkątną

$$Rx = Q^T b.$$

Zadanie 2.8 Dopasowanie krzywej o równaniu wielomianowym do zadanego układu puktów.

Przypuśćmy, że mamy dany układ m punktów na płaszczyźnie

$$(x_k, y_k) \quad k = 1, 2, \cdots, m.$$

Poszukujemy krzywej, o równaniu

$$y = \sum_{j=0}^{n} a_j x^j \quad n \leq m,$$

która *najlepiej pasuje* do zadanego układu punktów.

- Sformułuj powyższe zadanie, jako *zadanie LZNK.*
- Sformułuj warunki na to, aby zadanie było *regularne*.
- Zbuduj algorytm typu *"równania normalne"*.
- Zbuduj algorytm typu *"rozkład QR "*.
- Rozważ szczególny przypadek $n = 2$.

Zadanie 2.9 Dana jest funkcja $f \in L^2(a, b)$ i układ liniowo niezależny

$$\{\phi_1, \phi_2, \cdots, \phi_n\} \subset L^2(a, b).$$

Znajdź *element najlepszej aproksymacji* dla f w podprzestrzeni

$$span\{\phi_1, \phi_2, \cdots, \phi_n\} \subset L^2(a, b).$$

- Wykorzystaj metody opisane wyżej.
- Oznacz:
$$F(c_1, c_2, \cdots, c_n) = \int_a^b [f(x) - \sum_{j=1}^{n} c_j \phi_j(x)]^2 dx$$
  i wyznacz minimum funkcji $F(c_1, c_2, \cdots, c_n)$.
- Porównaj wyniki.

WIELOMIANY ORTOGONALNE

Ogólna teoria

Niech $\rho :\to \mathbf{R}^+$ będzie funkcją całkowalną. Założymy chwilowo, że jej *nośnik jest zbiorem nieskończonym* w przedziale $[a, b]$. Będziemy interesować się przestrzenią liniową rzeczywistą

$$L^2_\rho(a, b) = \{f | f : [a, b] \to \mathbf{R}, \int_a^b f(x)^2 \rho(x) dx < \infty\}.$$

W tej przestrzeni iloczyn skalarny jest określony wzorem

$$(f, g)_\rho = \int_a^b \rho(x) f(x) g(x) dx, \tag{2.10}$$

zaś normą jest

$$\|f\|_\rho = (\int_a^b \rho(x) f(x)^2 dx)^{\frac{1}{2}} = (f, f)_\rho^{\frac{1}{2}};$$

funkcja ρ nazywa się *wagą*.

Definicja. *Wielomiany ortogonalne* związane z iloczynem skalarnym $(\cdot, \cdot)_\rho$, to ciąg wielomianów

$$P_0, P_1, \cdots$$

takich, że

1. $P_k(x) = a_k x^k +$ wyrazy stopnia niższego od k, oraz $a_k > 0$ dla $k = 0, 1, \cdots$. Wynika stąd, że wielomian P_k jest stopnia *dokładnie* k,
2. $(P_k, P_l)_\rho = \delta_{k,l} \|P_k\|^2_\rho$,

Oczywiście wielomiany ortogonalne $P_0, P_1, \cdots, P_k$ stanowią *bazę* przestrzeni V_k wszystkich wielomianów stopnia $\leq k$.

Ponieważ wielomian $xP_k(x) \in V_{k+1}$ jest wielomianem stopnia $k+1$, więc istnieją współczynniki $\alpha_{k,j}, \quad j = 0, 1, \cdots, k+1$ takie, że

$$xP_k(x) = \sum_{j=0}^{k+1} \alpha_{k,j} P_j(x). \tag{2.11}$$

Nie trudno zauważyć, że ze względu na ortogonalność

$$\alpha_{k,j} = \frac{(xP_k, P_j)_\rho}{\|P_j\|_\rho^2}, \quad j = 0, 1, 2, \cdots, k+1. \tag{2.12}$$

Zauważmy jeszcze, że dla $\alpha_{k,k+1}$ mamy także inny wzór

$$\alpha_{k,k+1}\|P_{k+1}\|_\rho = (\|xP_k\|_\rho^2 - \sum_{j=0}^{k} \alpha_{k,j}^2 \|P_j\|_\rho^2)^{\frac{1}{2}} \tag{2.13}$$

Zadanie 2.10 Odpowiedz, dlaczego istnieją zawsze *rzeczywiste* współczynniki $\alpha_{k,l}$ $k = 0, 1, \cdots$ $j = 0, 1, \cdots, k+1$.

Wzór (2.11) możemy zapisać w postaci

$$xP_k(x) = \sum_{j=0}^{\infty} \alpha_{k,j} P_j(x) \tag{2.14}$$

określając dodatkowo $\alpha_{k,j} = 0$ dla $j > k+1$. Ze wzoru (2.14) wynika

$$(xP_k, P_l)_\rho = \int_a^b \rho(x) x P_k(x) P_l(x) dx = (xP_l, P_k)_\rho,$$

oraz

$$\alpha_{k,l}\|P_l\|^2 = \alpha_{l,k}\|P_l\|^2.$$

Ponieważ zaś dla $k > l+1$ $\alpha_{l,k} = 0$, to również $\alpha_{k,l} = 0$, dla $l < k-1$; oznacza to, wzory (2.11) i (2.14) mają na prawdę postać

$$xP_k(x) = \alpha_{k,k-1}P_{k-1}(x) + \alpha_{k,k}P_k(x) + \alpha_{k,k+1}P_{k+1}(x). \tag{2.15}$$

Udowodniliśmy więc następujące

Twierdzenie 2.3 *Wielomiany ortogonalne spełniają zawsze formułę trójczłonową postaci*

$$xP_k(x) = \alpha_{k,k-1}P_{k-1}(x) + \alpha_{k,k}P_k(x) + \alpha_{k,k+1}P_{k+1}(x),$$

gdzie

$$\alpha_{k,j} = \frac{(xP_k, P_j)_\rho}{\|P_j\|_\rho^2} \quad dla \quad j = k-1, k$$

$$\alpha_{k,k+1} = \frac{(\|xP_k\|_\rho^2 - \alpha_{k,k-1}^2\|P_{k-1}\|_\rho^2 - \alpha_{k,k}^2\|P_k\|_\rho^2)^{\frac{1}{2}}}{\|P_{k+1}\|_\rho}.$$

□

Zadanie 2.11 (Ważne!) Niech dany będzie *układ węzłów* w przedziale $[a, b]$:

$$a \le x_0 \le x_1 \le \cdots \le x_n \le b,$$

oraz odpowiadających im liczb dodatnich

$$\rho_0^2, \rho_1^2, \cdots, \rho_n^2,$$

tak zwanych *wag*. Określimy *iloczyn skalarny "dyskretny"*:

$$(f, g)_\rho = \sum_{j=0}^{n} \rho_j^2 f(x_j) g(x_j)$$

dla $f, g : [a, b] \to \mathbf{R}$.
Określ *wielomiany ortogonalne z wagą dyskretną* i zbadaj ich własności. Jak wygląda formuła trójczłonowa? Ile jest takich wielomianów?.

Uwaga. Formuła trójczłonowa może służyć do generowania ciągu wielomianów ortogonalnych, pod warunkiem, że na przykład, znamy *sposób unormowania* tych wielomianów (znamy ich normy $\|P_k\|_\rho$ $k = 0, 1, \cdots$). Tak jest, gdy interesują nas *wielomiany ortonormalne*, dla których $\|P_k\|_\rho = 1$ $k = 0, 1, \cdots$. Inny sposób *unormowania* ciągu wielomianów może polegać na zadaniu z góry wartości współczynnika przy x^k wielomianu P_k $k = 0, 1, \cdots$. Na przykład często mamy do czynienia z tak zwanymi *wielomianami monicznymi*, to jest wielomianami postaci:

$$P_k(x) = x^k + \text{wyrazy stopnia niższego niż} \quad k.$$

Zauważmy, że dla *wielomianów monicznych*

$$\alpha_{k,k+1} = 1,$$

a zatem formuła trójczłonowa jest postaci;

$$xP_k(x) = \alpha_{k,k-1}P_{k-1}(x) + \alpha_{k,k}P_k(x) + P_{k+1}(x),$$

gdyż
$$xP_k(x) = x^{k+1} + \quad \text{wyrazy stopnia niższego niż} \quad k+1.$$

Zadanie 2.12 Znajdź ogólny związek między współczynnikami a_k, gdzie $P_k(x) = a_k x^k + \cdots$, a współczynnikami formuły trójczłonowej $\alpha_{k,j}$.

PRZYKŁADY WIELOMIANÓW ORTOGONALNYCH

Wielomiany Czebyszewa 1-go rodzaju.

Weźmy pod uwagę funkcje
$$T_k(x) = \cos k\theta, \quad \text{gdzie} \ \ \theta = \arccos x \ \ k = 0, 1, \cdots, \quad |x| \le 1$$

Zadanie 2.13

1. Udowodnij, że
$$\int_{-1}^{1} \frac{T_k(x)T_l(x)}{\sqrt{1-x^2}} dx = \begin{cases} 0 & dla \quad k \ne l \\ \pi & dla \quad k = l = 0 \\ \frac{\pi}{2} & dla \quad k = l > 0 \end{cases}$$
   Zauważmy, że oznacza to, że funkcje T_k $k = 0, 1, \cdots$ są *ortogonalne z wagą* $\rho(x) = \frac{1}{\sqrt{1-x^2}}$ w przedziale $[-1, 1]$.

2. Wykorzystując znany wzór
$$\cos k\theta \cos l\theta = \frac{1}{2}[\cos(k-l)\theta + \cos(k+l)\theta],$$
   udowodnij, że funkcje $T_0, T_1, T_2, \cdots$ spełniają następującą *formułę trójczłonową*
$$T_{k+1}(x) + T_{k-1}(x) = 2xT_k(x).$$

3. Znajdź $T_0(x)$ i $T_1(x)$, oraz posługując się formułą trójczłonową udowodnij, że T_k jest wielomianem stopnia k postaci
$$2^{k-1}x^k + \quad \text{wyrazy stopnia niższego od } k.$$
   Jest to k-ty *wielomian Czebyszewa pierwszego rodzaju.*

4. Wyznacz pierwiastki wielomianu T_k. W jakim zbiorze związanym z wielomianami T_k leżą te pierwiastki? Wyznacz również punkty, w których T_k przyjmuje wartość $+1$ lub -1. Ile jest takich punktów w $[-1,1]$?

5. Udowodnij, że dla dowolnego $z \in \mathbf{C}$

$$T_k(z) = \frac{(z+\sqrt{z^2-1})^k + (z-\sqrt{z^2-1})^k}{2}, \quad k = 0, 1, \cdots.$$

Wskazówka. Skorzystaj z formuły trójczłonowej.

6. Udowodnij, że wśród wielomianów $w_k(x)$ stopnia k, takich, że

$$w_k(x) = x^k + \text{ wyrazy stopnia niższego od } k$$

najmniejszą normę $\|\cdot\|_{\infty,[-1,1]}$ ma wielomian

$$2^{1-k}T_k(x).$$

Wskazówka. Przypuść ze istnieje inny wielomian o tej własności, ale o mniejszej normie "sup" i rozważ różnice tych wielomianów. Jakiego jest stopnia ta różnica? Rozważ punkty w których wykresy tych wielomianów się przecinają. Ile jest takich punktów?

7. Niech $x_0 < -1$, i rozważmy zbiór wszystkich wielomianów w_k stopnia $\leq k$ spełniających warunek $w_k(x_0) = 1$. Udowodnij, że wśród wielomianów z tego zbioru najmniejszą normę $\|\cdot\|_{\infty,[-1,1]}$ ma wielomian

$$\frac{T_k(x)}{T_k(x_0)}.$$

Wyciągnij stąd następujący wniosek: niech $0 < a < b$; wśród wielomianów w_k stopnia $\leq k$ spełniających warunek $w_k(0) = 1$, najmniejszą normę $\|\cdot\|_{\infty,[a,b]}$ ma wielomian

$$\frac{T_k(\frac{b+a-2x}{b-a})}{T_k(\frac{b+a}{b-a})}.$$

Wskazówka. Przeprowadź dowód "ad absurdum". Skorzystaj z tego, że T_k w $k+1$ różnych punktach przedziału $[-1,1]$ przyjmuje naprzmian wartości +1 i -1. Jeśli istniałby wielomian stopnia $\leq k$ i mniejszej normie, to policz w ilu punktach wykresy tych wielomianów musiałyby się przecinać? Co stąd wynika?

8. Niech

$$R_k(x) = \frac{T_k(\frac{b+a-2x}{b-a})}{T_k(\frac{b+a}{b-a})}.$$

Udowodnij, że

$$\|R_k\|_{\infty,[a,b]} = \frac{1}{|T_k(\frac{b+a}{b-a})|} \leq 2(\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1})^k.$$

Wskazówka. Skorzystaj z wzoru

$$T_k(z) = \frac{(z+\sqrt{z^2-1})^k + (z-\sqrt{z^2-1})^k}{2}, \quad k = 0, 1, \cdots.$$

Wielomiany Legendre'a

Są to wielomiany ortogonalne w przedziale $[-1, 1]$ z wagą $\rho(x) = 1$. Dla wielomianów Legendre'a

$$P_0, P_1, \cdots$$

mamy następujące związki:

$$P_0(x) = 1,$$

$$P_1(x) = x.$$

Formuła trójczłonowa jest postaci

$$\frac{2k+1}{k+1} x P_k(x) = \frac{k}{k+1} P_{k-1}(x) + P_{k+1}(x), \quad \|P_k\|_\rho^2 = \frac{2}{2k+1}.$$

Wielomiany ortogonalne Hermite'a

Są to wielomiany $H_0, H_1, \cdots$, ortogonalne w przedziale $(-\infty, \infty)$ z wagą $\rho(x) = e^{-x^2}$. Zachodzą dla nich związki

$$H_0(x) = 1,$$

$$H_1(x) = 2x.$$

Formuła trójczłonowa jest postaci

$$2xH_k(x) = 2kH_{k-1}(x) + H_{k+1}(x), \quad \|H_k\|_\rho^2 = \sqrt{\pi} 2^k k!.$$

WŁASNOŚCI EKSTREMALNE WIELOMIANÓW ORTOGONALNYCH

W zadaniu dotyczącym wielomianów Czebyszewa poznaliśmy już *dwie własności ekstremalne* tych wielomianów. Sformułowane są one w punktach 6 i 7 tego zadania. Te własności odnoszą się do normy "sup" na odpowiednim przedziale. Okazuje się, że inne wielomiany ortogonalne mają również podobne *własności ekstremalne*, jednak związane z normą odpowiedniej przestrzeni typu $L^2_\rho(a,b)$. Fakt, że pewne wielomiany ortogonalne mają minimalne normy w określonych klasach wielomianów decyduje o roli jakie odgrywają one w zagadnieniach obliczeniowych. Twierdzenia podane poniżej dowodzimy w przypadku funkcji wagowych "ciągłych", określonych na przedziale $[a,b]$. Są one również prawdziwe dla *funkcji wagowych dyskretnych*, o których mowa w **Zadaniu do Twierdzenia 2.3**. Przeprowadzenie dowodów poniższych twierdzeń w przypadku *dyskretnym* zostawiamy czytelnikowi jako ćwiczenie.

Wielomiany jądrowe

Niech $P_0, P_1, \cdots$ będzie ciągiem wielomianów ortogonalnych z wagą ρ w przedziale $[a,b]$. Wielomian *dwóch zmiennych* x i y stopnia k ze względu na obie zmienne x i y

$$K_k(x,y) = K_k(y,x) = \sum_{j=0}^{k} \frac{P_j(x)P_j(y)}{\|P_j\|^2_\rho} \tag{2.16}$$

nazywa się *wielomianem jądrowym stopnia k.*

Twierdzenie 2.4 *Niech w_n będzie dowolnym wielomianem stopnia $\leq n$. Wtedy*

$$w_n(x) = \int_a^b \rho(y)K_n(x,y)w_n(y)dy. \tag{2.17}$$

Dowód. Mamy $w_n(x) = \sum_{j=0}^{n} c_j P_j(x)$ i, ponieważ $P_0, P_1, \cdots, P_n$ jest bazą ortogonalną przestrzeni wielomianów stopnia $\leq n$,

$$c_j = \frac{(w_n, P_j)_\rho}{\|P_j\|^2_\rho}, \quad j = 0, 1, \cdots, n.$$

Zatem

$$w_n(x) = \sum_{j=0}^{n} c_j P_j(x) = \sum_{j=0}^{n} (w_n, \frac{P_j P_j(x)}{\|p_j\|_\rho^2})_\rho =$$

$$= \int_a^b \rho(y) w_n(y) \sum_{j=0}^{n} \frac{P_j(y) P_j(x)}{\|P_j\|_\rho^2} dy = \int_a^b \rho(y) K_n(x,y) w_n(y) dy. \square$$

Wniosek 1. *Niech Q będzie dowolnym wielomianem stopnia $< n$, zaś niech $K_n(x,y)$ będzie wielomianem jądrowym. Wtedy*

$$\int_a^b \rho(y)(y-x) K_n(x,y) Q(y) dy = 0. \tag{2.18}$$

Dowód. Niech z będzie ustalone i niech $w_n(x) = (x-z)Q(x)$; $w_n(x)$ jest wielomianem stopnia $\leq n$, zatem

$$w_n(x) = \int_a^b \rho(y) K_n(x,y) w_n(y) dy = \int_a^b \rho(y)(y-z) Q(y) dy = (x-z)Q(x).$$

Połóżmy teraz $z = x$; otrzymamy

$$0 = \int_a^b \rho(y)(y-x) K_n(x,y) Q(y) dy. \quad \square$$

Weźmy pod uwagę wzór

$$\int_a^b \rho(y)(y-x) K_n(x,y) Q(y) dy = 0. \tag{2.19}$$

Zauważmy, że jeśli $\lambda < a \leq y \leq b$, to dla ustalonego λ funkcja zmiennej y

$$\omega(y) = (x-\lambda)\rho(y)$$

jest *nie ujemna* dla $y \in [a,b]$, a więc może ona odgrywać rolę *nowej wagi* dla *nowego iloczynu skalarnego*

$$(f,g)_\omega = \int_a^b \omega(y) f(y) g(y) dy = \int_a^b \rho(y)(y-\lambda) f(y) g(y) dy. \tag{2.20}$$

Załóżmy, że $\lambda < a \leq y \leq b$. Wtedy

$$(Q_l, K_n(\lambda, \cdot))_\omega = \int_a^b \rho(y)(y-\lambda)K_n(\lambda, y)Q_l(y)dy = 0$$

dla każdego wielomianu Q_l stopnia $l \leq n$, a więc także dla $Q_l(x) = K_l(\lambda, x)$. Stąd

Wniosek 2. *Wielomiany jądrowe*

$$K_0(\lambda, \cdot), K_1(\lambda, \cdot), K_2(\lambda, \cdot) \cdots$$

stanowią układ wielomianów ortogonalnych z nową wagą $\omega(x) = (x-\lambda)\rho(x)$ w przedziale $[a, b]$.

(*) Rozważmy teraz następujące zadanie na minimum normy: poszukujemy wielomianu w_n stopnia $\leq n$, który dla ustalonej liczby x_0, oraz dla ustalonej liczby α, spełnia warunek

$$w_n(x_0) = \alpha,$$

i który ma najmniejszą normę $\|\cdot\|_\rho$.

Twierdzenie 2.5 *Rozwiązaniem zadania* **(*)** *na minimum normy $\|\cdot\|_\rho$ jest wielomian*

$$w_{opt}(x) = \frac{K_n(x, x_0)}{K_n(x_0, x_0)}\alpha.$$

Dowód. Dowolny wielomian w_n stopnia $\leq n$ spełniający warunek $w_n(x_0) = \alpha$ przedstawimy w postaci rozwinięcia względem bazy $P_0, P_1, \cdots, P_n$ wielomianów ortogonalnych z wagą ρ na przedziale $[a, b]$

$$w_n(x) = \sum_{j=0}^{n} c_j P_j(x).$$

Jeśli $w_n(x_0) = \alpha$, to

$$\alpha = \sum_{j=0}^{n} c_j P_j(x_0),$$

i stąd

$$\alpha^2 = (\sum_{j=0}^{n} c_j P_j(x_0))^2 = (\sum_{j=0}^{n} c_j \|P_j\|_\rho \frac{P_j(x_0)}{\|P_j\|_\rho})^2.$$

Z nierówności Schwarz'a otrzymamy

$$\alpha^2 = [\sum_{j=0}^{n} c_j \|P_j\|_\rho (\frac{P_j(x_0)}{\|P_j\|_\rho})]^2 \leq \sum_{j=0}^{n} c_j^2 \|P_j\|_\rho^2 \sum_{j=0}^{n} (\frac{P_j(x_0)}{\|P_j\|_\rho})^2 = \|w_n\|_\rho^2 K_n(x_0, x_0),$$

lub inaczej

$$\frac{\alpha^2}{K_n(x_0, x_0)} \leq \|w_n\|_\rho^2. \tag{2.21}$$

Obliczmy teraz normę wielomianu $K_n(x_0, x)$.

$$\|K_n(x_0, \cdot)\|_\rho^2 = \int_a^b \rho(x) K_n(x_0, x)^2 dx =$$

$$= \int_a^b \sum_{j=0}^{n} \frac{P_j(x_0) P_j(x)}{\|P_j\|_\rho^2} \sum_{l=0}^{n} \frac{P_l(x_0) P_l(x)}{\|P_l\|_\rho^2} \rho(x) dx =$$

$$= \sum_{j=0}^{n} \sum_{l=0}^{n} \frac{P_j(x_0) P_l(x_0)}{\|P_j\|_\rho^2 \|P_l\|_\rho^2} \int_a^b \rho(x) P_j(x) P_l(x) dx =$$

$$= \sum_{j=0}^{n} \sum_{l=0}^{n} \frac{P_j(x_0) P_l(x_0)}{\|P_j\|_\rho^2 \|P_l\|_\rho^2} \delta_{j,l} \|P_j\|_\rho^2 = \sum_{j=0}^{n} \frac{P_j(x_0)^2}{\|P_j\|_\rho^2} = K_n(x_0, x_0).$$

Stąd

$$\|Q_{opt}\|_\rho^2 = \|\alpha \frac{K_n(x_0, \cdot)}{K_n(x_0, x_0)}\|_\rho^2 = \frac{\alpha^2}{K_n(x_0, x_0)^2} K_n(x_0, x_0) = \frac{\alpha^2}{K_n(x_0, x_0)}.$$

Wobec nierówności (2.21) mamy

$$\|Q_{opt}\|_\rho^2 \leq \|w_n\|_\rho^2$$

gdzie w_n jest dowolnym wielomianem stopnia $\leq n$ spełniającym warunek $w_n(x_0) = \alpha$. $\square$

Komentarz. Załóżmy teraz, że $x_0 < a < b$. Wtedy funkcja

$$\omega(x) = (x - x_0)\rho(x), \quad x \in [a, b]$$

przyjmuje tylko wartości nieujemne, gdy $x \in [a, b]$, a więc jest prawidłową funkcją - wagą. Udowodniliśmy, (patrz wniosek z Twierdzenia 2.4), że wielomiany jądrowe $K_k(x, x_0)$ $k = 0, 1, \cdots$ są *ortogonalne z wagą* ω *na przedziale* $[a, b]$. Z drugiej strony, Twierdzenie 2.5 mówi o tym, że wielomian jądrowy $K_n(x, x_0)$ po odpowiednim *unormowaniu*:

$$\frac{K_n(x, x_0)}{K_n(x_0, x_0)}\alpha$$

realizuje *minimum normy* $\| \cdot \|_\rho$. Zauważmy, że na odwrót, dowolne wielomiany ortogonalne z pewną wagą ω na przedziale $[a, b]$ mogą być uważane za *wielomiany jądrowe* pochodzące od wielomianów ortogonalnych z wagą $\rho(x) = \frac{\omega(x)}{x - x_0}$ na przedziale $[a, b]$; zatem po odpowiednim unormowaniu będą one realizować *minimum normy* $\| \cdot \|_\rho$ w zadaniu $(*)$. Wykorzystamy ten fakt w dalszej części tego rozdziału.

ZASTOSOWANIA WIELOMIANÓW ORTOGONALNYCH

Wielomiany ortogonalne stosuje się w bardzo wielu różnych dziedzinach matematyki obliczeniowej. Zajmiemy się tutaj tylko dwoma przykładami takiego zastosowania.

Optymalne węzły interpolacji wielomianowej Lagrange'a

Powróćmy na chwilę do interpolacji Lagrange'a przy pomocy *jednego wielomianu* na przedziale $[a, b]$. Założymy, że funkcja interpolowana

$$f : [a, b] \to \mathbf{R}$$

ma $n + 1$ pochodnych ciągłych w przedziale $[a, b]$, w którym mamy $n + 1$ różnych węzłów

$$a \leq x_0 < x_1 < x_2 \cdots < x_n \leq b.$$

Wiemy, że w tym przypadku **błąd interpolacji** wyraża sę wzorem

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi(x))}{(n+1)!}\omega(x),$$

gdzie P_n jest wielomianem interpolacyjnym, zaś $\xi(x)$ jest pewnym punktem przedziału otwartego $(\min\{x, x_0\}, \max\{x, x_n\})$. Zadajmy sobie pytanie, *czy można tak dobrać węzły interpolacji żeby błąd był możliwie najmniejszy.* Weźmy pod uwagę wielomian stopnia $n + 1$

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n);$$

Zauważmy, że jest to tak zwany *wielomian moniczny.* Wiemy (patrz **Zadanie** p. 6 - Wielomiany Czebyszewa), że na przedziale $[-1, 1]$ wielomian moniczny $2^{-n}T_{n+1}$ ma minimalną normę "sup". Przeksztłcenie liniowe $\frac{b+a-2x}{b-a}$ przeprowadza odcinek $[a, b]$ na odcinek $[-1, 1]$. Nie trudno znaleźć pierwiastki przekształconego wielomianu, znając pierwiastki $t_0 < t_1 < t_2 < \cdots < t_n$ wielomianu T_{n+1}.

Zadanie 2.14 Znajdź pierwiastki przekształconego wielomianu Czebyszewa.

Oznaczmy liczby znalezione w **Zadaniu 2.13** przez

$$y_0 < y_1 < y_2 < \cdots < y_n.$$

Jeśli przyjmiemy *jako nowe węzły interpolacji* liczby y_j, $j = 0, 1, 2, \cdots, n$ (wszystkie one leżą w przedziale $[a, b]$!), to uzyskamy wielomian interpolacyjny Lagrange'a dla którego wyraz $\omega(x)$, występujący w wyrażeniu na błąd będzie miał minimalną normę "sup" na przedziale $[a, b]$. Okazuje się, że efektem tego optymalnego doboru węzłów interpolacji jest *znaczne polepszenie własności aproksymacyjnych* wielomianu interpolacyjnego Lagrange'a. Można udowodnić, [5] że dla przedziału $[-1, 1]$, jeśli węzłami są pierwiastki t_j wielomianu Czebyszewa T_{n+1}, czyli liczby

$$t_j = \cos\frac{2j+1}{2(n+1)}\pi, \quad j = 0, 1, 2, \cdots, n,$$

to mamy następujące oszacowanie dla *wielomianów bazowych Lagrange'a* $l_j(x)$

$$\sum_{j=0}^{n+1} \|l_j\|_{\infty,[-1,1]} \leq \frac{2}{\pi}\ln(n) + 4$$

Posługując się **Twierdzeniem Jacksona** podaliśmy oszacowanie błędu dla wielomianu interpolacyjnego Lagrange'a w zależności od *stopnia gładkości* funkcji interpolowanej f. Załóżmy teraz, że $f \in \mathbf{C^1([-1,1])}$. Z naszych oszacowań uzyskanych dla wielomianu interpolacyjnego Lagrange'a P_n wynika, że

$$\|f - P_n\|_{\infty,[-1,1]} \leq (1 + \sum_{j=0}^{n} \|l_j\|_{\infty,[-1,1]})\|f - Q_n\|_{\infty,[-1,1]},$$

gdzie Q_n jest *wielomianem najlepszej aproksymacji w sensie normy "sup"* dla funkcji f. W rozważanym przypadku mamy

$$\|f - Q_n\|_{\infty,[-1,1]} \leq \frac{6}{n}\|f'\|_{\infty,[-1,1]}.$$

[5] Patrz: S.Paszkowski "Zastosowania numeryczne wielomianów i szeregów Czebyszewa" PWN 1975

Stąd

$$\|f - P_n\|_{\infty,[-1,1]} \leq (5 + \frac{2}{\pi} ln(n)) \frac{6}{n} \|f'\|_{\infty,[-1,1]}.$$

Ponieważ $\frac{ln(n)}{n} \to 0$, gdy $n \to \infty$, widzimy, że *jeśli używamy optymalnych węzłów, to, przy założeniu, że* $f \in \mathbf{C^1([-1,1])}$, *wielomian interpolacyjny Lagrange'a zbiega w normie "sup" do funkcji* f, *którą intertpoluje.*

Zadanie 2.15 Dla dowolnego, ograniczonego przedziału $[a, b]$ znajdź oszacowania odpowiadające opisanemu wyżej przypadkowi przedziału $[-1, 1]$.

Metody wielomianowe rozwiązywania numerycznego układów równań algebraicznych liniowych

Zajmiemy się teraz pewną klasą metod numerycznych iteracyjnych rozwiązywania układów równań algebraicznych liniowych. Są to tak zwane *metody wielomianowe.* Zajmiemy się układem równań algebraicznych liniowych postaci

$$Ax = d, \tag{2.22}$$

gdzie macierz A jest *symetryczna i dodatnio określona, wymiaru* $n \times n$. Weźmy pod uwagę następujący *proces iteracyjny Richardsona*

$$x_0 \quad \text{dowolny wektor "startowy"},$$

$$x_{k+1} = x_k + \frac{r_k}{q_k}. \tag{2.23}$$

Wektor $r_k = d - Ax_k$, jest tak zwanym *reziduum*, zaś q_k, $k = 0, 1, \cdots$ jest liczbą zwaną *współczynnikiem relaksacji.* W ten sposób określiliśmy *całą klasę metod* zależną od wyboru ciągu *współczynników relaksacji* $\{q_j\}_{j=0,1,\ldots}$. Współczynniki relaksacji będziemy wybierać tak, aby spełnione było określone *kryterium optymalności* procesu (2.23) zapewniające *szybką zbieżność* procesu Richardsona. Interpretacja tego procesu jest prosta: następny wektor *przybliżający rozwiązanie* x *równania* (2.22) wybieramy w ten sposób, że do poprzedniego przybliżenia dodajemy *poprawkę* proporcjonalną do reziduum na poprzednim kroku. Współczynnikiem proporcjalności jest odwrotność współczynnika relaksacji.

Znajdziemy najpierw zależność między kolejnymi reziduami

$$r_{k+1} = d - Ax_{k+1} = d - A(x_k + \frac{r_k}{q_k}) = (I - \frac{A}{q_k})r_k.$$

Stąd wnosimy, że dla każdego $k = 0, 1, 2, \cdots$

$$r_k = (I - \frac{A}{q_{k-1}})(I - \frac{A}{q_{k-2}}) \cdots (I - \frac{A}{q_0})r_0 \tag{2.24}$$

gdzie $r_0 = d - Ax_0$. Oznaczmy

(2.25). $$R_k(x) = (1 - \frac{x}{q_{k-1}})(1 - \frac{x}{q_{k-2}}) \cdots (1 - \frac{x}{q_0})$$

Wielomian stopnia k określony wzorem (2.25) nazywa się *k-tym wielomianem rezidualnym.* Zauważmy odrazu, że

$$R_k(0) = 1,$$

$$R_k(q_j) = 0 \quad j = 0, 1, \cdots k - 1.$$

Ogólnie: *każdy wielomian W_k stopnia k taki, że $W_k(0) = 1$ będziemy nazywać k-tym wielomianem rezidualnym.* Każdy taki wielomian *musi* być postaci (2.25). Wynika stąd, że dla naszego procesu Richardsona

$$r_k = R_k(A)r_0,$$

gdzie R_k jest k-tym wielomianem rezidualnym.

O macierzy A założyliśmy, że jest symetryczna i dodatnio określona. Niech więc jej *widmo*[6] $\sigma(A) \subset [a, b]$, gdzie $0 < a < b$. Oszacujemy z góry normę euklidesową k-tego reziduum

$$\|r_k\|^2 = (r_k, r_k) = \|R_k(A)r_0\|^2 \leq \|R_k(A)\|^2 \|r_0\|^2,$$

ale ponieważ macierz A jest symetryczna

$$\|R_k(A)\| = \max_{\lambda_j \in \sigma(A)} |R_k(\lambda_j)| \leq \sup_{x \in [a,b]} |R_k(x)| = \|R_k\|_{\infty,[a,b]}.$$

[6] zbiór wszystkich wartości własnych

Wiemy, że normę $\|R_k\|_{\infty,[a,b]}$ minimalizuje *przekształcony wielomian Czebyszewa 1-go rodzaju*

$$\frac{T_k(\frac{b+a-2x}{b-a})}{T_k(\frac{b+a}{b-a})},$$

którego normę szacujemy z góry[7] przez liczbę

$$2(\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1})^k.$$

Stąd mamy *optymalne oszacowanie* k-tego reziduum:

$$\|r_k\| \leq 2(\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1})^k \ \|r_0\|_{\infty,[a,b]}. \tag{2.26}$$

Zauważmy, że

$$2(\frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1})^k \rightarrow 0, \quad \text{gdy } k \rightarrow \infty.,$$

a więc **proces iteracyjny (2.23) jest geometrycznie zbieżny**. Jego szybkość zbieżności określa liczba

$$q = \frac{\sqrt{\frac{b}{a}}-1}{\sqrt{\frac{b}{a}}+1} < 1.$$

Zauważmy, że moglibyśmy przyjąć *dla celów oszacowania*, że

$$a = \min_{\lambda\in\sigma(A)} \lambda = \lambda_{min},$$

$$b = \max_{\lambda\in\sigma(A)} \lambda = \lambda_{max}.$$

Ale dla *współczynnika uwarunkowania* $\kappa(A)$ macierzy A mamy

$$\kappa(A) = \frac{\lambda_{max}}{\lambda_{min}}.$$

[7]Patrz, Wielomiany Czebyszewa, **Zadanie** p.7 i 8.

Stąd ostatecznie

$$\|r_k\| \leq 2(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1})^k\|r_0\|. \tag{2.27}$$

Należy tu podkreślić, że wzór (2.27) nie może być, poza bardzo szczególnymi przypadkami, traktowany jako oszacowanie szybkości zbieżności procesu iteracyjnego Czebyszewa, o którym będzie mowa w następnym paragrafie. Poza bardzo specjalnymi przypadkami, *nie znamy liczb* λ_{min} *ani* λ_{max}, zaś konkretny algorytm Czebyszewa może być określony, gdy znane są liczby a i b. Wzór (2.27) będzie jednak nam przydatny w daszej części tego rozdziału. Ponieważ liczby $q_0, q_1, \cdots, q_{k-1}$ są *pierwiastkami przekształconego wielomianu Czebyszewa* $T_k(\frac{a+b-2x}{b-a})$, więc potrafimy je łatwo znaleźć. Mając te liczby możemy zbudować

Dwupoziomową metodę Czebyszewa

Przy pomocy tego algorytmu wykonujemy N kroków iteracyjnych dla N zadanego z góry

$$x_{k+1} = x_k + \frac{r_k}{q_k}, \quad k = 0, 1, \cdots N-1,$$

gdzie $T_N(\frac{a+b-2q_j}{b-a}) = 0$, $j = 0, 1, 2, \cdots, N-1$, zaś x_0 jest dowolnym wektorem startowym - na przykład $x_0 = 0$. Łatwo sprawdzamy, że współczynniki relaksacji są postaci

$$q_j = \frac{a+b}{2} - s_j\frac{b-a}{2}, \quad \text{gdzie} \quad s_j = cos(\frac{\pi(2j+1)}{2N}), \tag{2.28}$$

dla $j = 0, 1, \cdots, N-1$. W tej chwili nasz algorytm jest określony z *z dokładnością do kolejności współczynników relaksacji.* Gdybyśmy mogli wykonywać obliczenia używając *"prawdziwej" arytmetyki* sprawa kolejności nie odgrywałaby żadnej roli. Jednak arytmetyka *"komputerowa"* różni się od *"prawdziwej"* i użycie współczynników relaksacji w nie właściwej kolejności może spowodować silne zaburzenie procesu, wprowadzając duże błedy. Właściwy dobór kolejności, to taki, przy którym, kolejne iloczyny czynników $(I - \frac{A}{q_j})$ we wzorze na reziduum *stopniowo się równoważą.* To znaczy, liczby q_j występują w takiej kolejności, że po dużym czynniku następuje mały i dzięki temu nie następuje ani gwałtowny wzrost ani gwałtowny spadek wielkości częściowych iloczynów. [8] Podamy tu, za wspomnianą pracą, sposób znajdowania optymalnej kolejności numerów we wzorze (2.28), dla $N = 2^p$, $p = 0, 1, 2, \cdots$. Dla $p = 0$, mamy $j_1 = 0$ i oczywiście nie ma tu wątpliwości co do kolejności. Jeśli znamy już kolejność numerów dla $N = 2^{p-1}$:

$$\{j_1, j_2, \cdots, j_{2^{p-1}}\},$$

to dla $N = 2^p$ będzie:

$$\{j_1, 2^p - 1 - j_1, j_2, 2^p - 1 - j_2, j_3, 2^p - 1 - j_3, \cdots, j_{2^{p-1}}, 2^p - 1 - j_{2^{p-1}}\}.$$

Przykład.

p	**N**	**ciąg numerów**
0	**1**	**0**
1	**2**	**0, 1**
2	**4**	**0, 3, 1, 2**
3	**8**	**0, 7, 3, 4, 1, 6, 2, 5**
4	**16**	**0, 15, 7, 8, 3, 12, 4, 11, 1, 14, 6, 9, 2, 13, 5, 10**

Metoda Czebyszewa może być używana w dwóch wersjach:

1. Ustalamy N dostatecznie duże dla osiągnięcia żądanej dokładności i wykonujemy N kroków opisanym algorytmem, *pamiętając o właściwej kolejności współczynników relaksacji.*

2. **Wersja cykliczna.**

   (a) Wybieramy jakieś N i x_0. Wykonujemy N kroków metody *zachowując zawsze właściwą kolejność współczynników relaksacji.*

   (b) Jako x_0 przyjmujemy wyliczone x_N i wykonujemy znów N kroków iteracyjnych

   (c) Powtarzamy punkt (b), aż do uzyskania żądanej dokładności.

[8] Ścisłe uzasadnienie - patrz V.I.Lebedev i S.A.Finogenov "O probleme vybora iteracionnych parametrov...." Žurnał vyč. matem. i mat. fiziki T.11 Nr 2 1971

Wadą metody Czebyszewa jest to, że aby ją stosować z optymalną możliwą efektywnością, musimy znać możliwie dokładne *dolne i górne oszacowanie widma macierzy* A, a i b. Można pokazać, że metoda Czebyszewa będzie funkcjonowała również gdy podamy zbyt wysoką wartość dla a, jednak zbieżność będzie wolniejsza niż to wynikałoby z wyprowadzonych wyżej oszacowań. Istnieje także inna wersja metody Czebyszewa, tak zwana *trzypoziomowa metoda Czebyszewa.*

Metody gradientów sprzężonych

Są to metody wywodzące się również od *procesu iteracyjnego Richardsona*

$$x_{k+1} = x_k + \frac{r_k}{q_k}, \quad k = 0, 1, 2, \cdots$$

gdzie punkt startowy x_0 jest dowolny, zaś $r_k = d - Ax_k$. O układzie

$$Ax = d$$

zakładamy, jak poprzednio, że macierz A wymiaru $m \times m$ jest *symetryczna i dodatnio określona.* Dla takiej macierzy mamy następujący *rozkład spektralny*:

$$A = Q^T \Lambda Q$$

gdzie $Q^T Q = QQ^T = I$, oraz Λ jest macierzą diagonalną mającą na głównej przekątnej *wartości własne* macierzy A:

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & \lambda_n \end{bmatrix}$$

Założymy, bez zmniejszania ogólności, że

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

Współczynniki relaksacji q_k będziemy teraz dobierać tak, aby uzyskać, nie optymalne oszacowanie reziduum r_k, jak to było w przypadku metody Czebyszewa, ale aby *zminimalizować pewną normę reziduum r_k dla każdego* $k = 1, 2, \cdots$. Normę, o której mowa, zwiążemy z pewną wybraną przez nas *macierzą wagową wymiaru* $n \times n$ B. O tej macierzy założymy, że *jest ona symetryczna i dodatnio określona* i że

$$B = Q^T D Q,$$

gdzie

$$D = \begin{bmatrix} d_1 & 0 & 0 & 0 & \cdots & 0 & 0 \\ 0 & d_2 & 0 & 0 & \cdots & 0 & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & 0 & \cdots & 0 & d_n \end{bmatrix}$$

oraz $d_j > 0$ dla $j = 1, 2, \cdots, n$. Oznacza to, że macierze A i B mają takie same wektory własne, i że *komutują*, to znaczy, że $AB = BA$. Ze względu na to, że macierz B jest *symetryczna i dodatnio określona*, to można przyjąc jako *nową normę* wektora $x \in \mathbf{R^n}$

$$\|x\|_B = (Bx, x)^{\frac{1}{2}}.$$

Rezidua r_k procesu Richardsona będziemy minimalizować w sensie takiej właśnie normy. Przykładami macierzy B o żądanych własnościach są A^p, $p \geq 0$ i A^{-1}.

Proces iteracyjny określimy tak, aby dla każdego ustalonego n *rezyduum* $r_n = d - Ax_n$ po n krokach iteracji miało najmniejszą możliwą normę $\| \cdot \|_B$:

$$\|r_n\|_B = min.$$

Pamiętamy, że dla procesu Richardsona

$$r_k = R_k(A) r_0,$$

gdzie $R_k(x)$ jest *wielomianem rezidualnym*, to jest takim wielomianem stopnia k, że $R_k(0) = 1$. Pierwiastki tego wielomianu są współczynnikami relaksacji q_j naszego procesu. Ze względu na to, że $Q^T Q = I$ mamy

$$\|r_n\|_B^2 = r_n^T B r_n = r_0^T R_n(A) B R_n(A) r_0 = \tag{2.29}$$

$$= r_0^T Q^T R_n(\Lambda) Q Q^T D Q Q^T R_n(\Lambda) Q r_0 = r_0^T Q^T R_n(\Lambda)^2 D Q r_0 =$$

$$= s_0^T R_n(\Lambda)^2 D s_0 = \sum_{j=1}^{m} k_j^2 d_j R_n(\lambda_j)^2.$$

gdzie $s_0 = [k_1, k_2, \cdots, k_m]^T = Qr_0$ i $\|s_0\| = \|Q_n r_0\| = \|r_0\|$. Oznaczmy teraz

$$\rho = \left\{ \begin{array}{ccccc} \lambda_1, & \lambda_2, & \lambda_3, & \cdots, & \lambda_m \\ k_1^2 d_1, & k_2^2 d_2, & k_3^2 d_3, & \cdots, & k_m^2 d_m \end{array} \right\}. \tag{2.30}$$

Jest to *dyskretna funkcja - waga* określona na widmie $\sigma(A)$ macierzy A. Używając tej funkcji - wagi, możemy napisać

$$\|r_n\|_B^2 = \sum_{j=1}^{n} k_j^2 d_j R_n(\lambda_j)^2 = \|R_n\|_\rho^2. \tag{2.31}$$

Zatem nasze zagadnienie zostało sprowadzone do *zadania wyznaczenia wielomianu R_n stopnia n spełniającego warunek $R_n(0) = 1$, który ma najmniejszą normę $\|\cdot\|_\rho$ związaną z dyskretną funkcją wagową ρ, określoną wzorem* (2.30). Znamy rozwiązanie tego zadania: podaje je **Twierdzenie 2.5**. Optymalnym wielomianem jest *wielomian jądrowy*

$$\frac{K_n(0, x)}{K_n(0, 0)},$$

gdzie

$$K_n(x, y) = \sum_{j=0}^{n} \frac{P_j(x) P_j(y)}{\|P_j\|_\rho^2},$$

zaś wielomiany $P_0, P_1, \cdots, P_m$ są ortogonalne w sensie iloczynu skalarnego dyskretnego z wagą ρ

$$(f, g)_\rho = \sum_{j=0}^{m} f(\lambda_j) g(\lambda_j) k_j^2 d_j.$$

Przypomnijmy (patrz **Wniosek 2. z Twierdzenia 2.4**), że wielomiany jądrowe $K_n(0, x)$, a więc także *wielomiany optymalne*

$$\frac{K_n(0, x)}{K_n(0, 0)},$$

są *ortogonalne w sensie iloczynu skalarnego dyskretnego* określonego przez funkcję wagową

$$\omega = \left\{ \begin{array}{ccccc} \lambda_1, & \lambda_2, & \lambda_3, & \cdots, & \lambda_m \\ k_1^2 d_1 \lambda_1, & k_2^2 d_2 \lambda_2, & k_3^2 d_3 \lambda_3, & \cdots, & k_m^2 d_m \lambda_m \end{array} \right\}. \tag{2.32}$$

Ta obserwacja pozwoli nam zbudować algorytm iteracyjny inaczej niż w przypadku *dwupoziomowej metody Czebyszewa*, gdzie wykorzystywaliśmy znajomość pierwiastków *wielomianów rezidualnych* R_n (były to pierwiastki "przesuniętych" wielomianów Czebyszewa). Teraz nie znamy z góry pierwiastków wielomianów rezidualnych R_n, a wyznaczanie ich numeryczne, byłoby bardzo pracochłonne i wobec tego mijałoby się z celem. Nasz algorytm oprzemy na *formule trójczłonowej* dla wielomianów optymalnych. Wypiszmy formułę trójczłonową dla optymalnych wielomianów rezidualnych

$$xR_n(x) = \alpha_{n,n-1}R_{n-1}(x) + \alpha_{n,n}R_n(x) + \alpha_{n,n+1}R_{n+1}.$$

Ponieważ $R_n(0) = 1$, to podstawiając $x = 0$ otrzymamy

$$\alpha_{n,n-1} + \alpha_{n,n} + \alpha_{n,n+1} = 0.$$

Kładąc teraz

$$\alpha_{n,n+1} = -(\alpha_{n,n-1} + \alpha_{n,n}),$$

otrzymamy

$$xR_n(x) = \alpha_{n,n-1}(R_{n-1}(x) - R_{n+1}(x)) + \alpha_{n,n}(R_n(x) - R_{n+1}(x)).$$

Stąd wynika następujący związek dla reziduów

$$Ar_n = \alpha_{n,n-1}(r_{n-1} - r_{n+1}) + \alpha_{n,n}(r_n - r_{n+1}),$$

gdyż, jak pamiętamy, $r_n = R_n(A)r_0$. Mamy jednak

$$r_{n-1} - r_{n+1} = d - Ax_{n-1} - d + Ax_{n+1} = A(x_{n+1} - x_{n-1}),$$

oraz

$$r_n - r_{n+1} = d - Ax_n - d + Ax_{n+1} = A(x_{n+1} - x_n)$$

Ponieważ macierz A jest *nieosobliwa*, po podstawieniu do (2.33) możemy "skrócić przez A", i wtedy dostaniemy następujący związek między x_{n-1}, x_n oraz x_{n+1}

$$x_{n+1} = \frac{1}{\alpha_{n,n-1} + \alpha_{n,n}}[r_n + \alpha_{n,n-1}x_{n-1} + \alpha_{n,n}x_n]. \tag{2.34}$$

gdzie $r_n = d - Ax_n$. Wypiszmy teraz jeszcze wzory dla współczynników we wzorze (2.34):

$$\alpha_{n,n-1} = \frac{(xR_n, R_{n-1})_\omega}{\|R_{n-1}\|_\omega^2},$$

$$\alpha_{n,n} = \frac{(xR_n, R_n)_\omega}{\|R_n\|_\omega^2}.$$

Przechodząc do macierzowej postaci tych wzorów dostaniemy

$$\alpha_{n,n-1} = \frac{\sum_{j=1}^m k_j^2 d_j \lambda_j^2 R_n(\lambda_j) R_{n-1}(\lambda_j)}{\sum_{j=1}^m k_j^2 d_j \lambda_j R_{n-1}(\lambda_j)^2} =$$

$$= \frac{(BA^2 r_n, r_{n-1})}{(BAr_{n-1}, r_{n-1})}, \tag{2.35}$$

oraz podobnie

$$\alpha_{n,n} = \frac{(BA^2 r_n, r_n)}{(BAr_n, r_n)}. \tag{2.36}$$

Jeśli sprecyzujemy jaka jest postać *macierzy wagowej* B, to wzory (2.34), (2.35) i (2.36) będą określać $n - ty$ krok

Metody Gradientów Sprzężonych.

Najczęściej używa się:

- $B = I = Q^T I Q$ - otrzymamy wtedy tak zwaną Metodę Minimalnych Reziduów, w skrócie CGMR - Conjugate Gradients Minimal Residuals.
- $B = A^{-1} = Q^T \Lambda^{-1} Q$ - otrzymamy wtedy tak zwaną Metodę Minimalnych Błędów, w skrócie CGME - Conjugate Gradients Minimal Errors.

Przyjrzyjmy się wzorom w obu przypadkach

1. **CGMR**

   Wtedy $B = I$

$$\alpha_{n,n-1} = \frac{(BA^2 r_n, r_{n-1})}{(BAr_{n-1}, r_{n-1})} = \frac{(A^2 r_n, r_{n-1})}{(Ar_{n-1}, r_{n-1})} = \frac{r_{n-1}^T A^2 r_n}{r_{n-1}^T A r_{n-1}},$$

$$\alpha_{n,n} = \frac{(BA^2 r_n, r_n)}{(BAr_n, r_n)} = \frac{r_n^T A^2 r_n}{r_n^T A r_n}.$$

$$x_{n+1} = \frac{1}{\alpha_{n,n-1} + \alpha_{n,n}}[r_n + \alpha_{n,n-1}x_{n-1} + \alpha_{n,n}x_n].$$

Algorytm ten minimalizuje na każdym kroku normę euklidesową reziduum

$$\|r_n\|^2 = r_n^T r_n.$$

2. **CGME**

   Wtedy $B = A^{-1}$

$$\alpha_{n,n-1} = \frac{(BA^2 r_n, r_{n-1})}{(BAr_{n-1}, r_{n-1})} = \frac{(Ar_n, r_{n-1})}{(r_{n-1}, r_{n-1})} = \frac{r_{n-1}^T A r_n}{r_{n-1}^T r_{n-1}},$$

$$\alpha_{n,n} = \frac{(BA^2 r_n, r_n)}{(BAr_n, r_n)} = \frac{r_n^T A r_n}{r_n^T r_n}.$$

$$x_{n+1} = \frac{1}{\alpha_{n,n-1} + \alpha_{n,n}}[r_n + \alpha_{n,n-1}x_{n-1} + \alpha_{n,n}x_n].$$

   Algorytm ten minimalizuje na każdym kroku normę $\| \cdot \|_{A^{-1}}$ reziduum r_n. Mamy

$$\|r_n\|_{A^{-1}}^2 = r_n^T A^{-1} r_n.$$

   Poniewaz $r_n = d - Ax_n = Ax - Ax_n = A(x - x_n) = Ae_n$, więc

$$\|r_n\|_{A^{-1}}^2 = r_n^T A^{-1} r_n = e_n^T A A^{-1} A e_n = e_n^T A e_n = \|e_n\|_A^2.$$

   Ponieważ $e_n = x - x_n$ gdzie x jest dokładnym rozwiązaniem układu $Ax = d$, można więc interpretować to wyrażanie jako *normę z wagą* A błedu przybliżenia na n-tym kroku e_n - stąd nazwa.

Porównajmy własności **Dwupoziomowej Metody Czebyszewa**, i opisanych wyżej algorytmów **Metody Gradientów Sprzężonych** - w skrócie **CG**. Oba typy metod, w opisanych wersjach, mogą być stosowane do układów równań

$$Ax = d,$$

gdzie macierz A jest *symetryczna i dodatnio określona.* Pierwsza rzecz która się rzuca w oczy to to, że we wzorach określających metody **CG**, x_{n+1} zależy od dwóch poprzednich przybliżeń x_n i x_{n-1}, podczas gdy dla metody Czebyszewa x_{n+1} zależało tylko od x_n. O metodach **CG** mówimy, że są one **trzypoziomowe**. Metoda Czebyszewa mogła być stosowana wtedy, gdy znany był przedział $[a, b]$ taki że $0 < a < b$ i $\sigma(A) \subset [a, b]$. Metody **CG** takiej informacji nie potrzebują. Ponadto, o ile metoda Czebyszewa jedynie *minimalizowała oszacowanie z góry* dla reziduum, to metody **CG** minimalizują poprostu normę tego reziduum. Dla dwupoziomowej metody Czebyszewa trzeba było **wybierać w specjalny sposób** kolejność wprowadzania współczynników relaksacji. W metodach trzypoziomowych taki problem wogóle nie występuje.

Zastanówmy się jeszcze nad sprawą **startu** algorytmów trzypoziomowych **CGMR** i **CGME**. Aby proces wystartował trzeba podać dwa wektory x_0 i x_1. Przyjmując x_0 dowolnie, x_1 dobieramy korzystając ze wzoru

$$x_1 = x_0 + \frac{r_0}{q_0},$$

gdzie q_0 jest tak dobrane, aby $\|r_1\|_B^2 = Min$.

Zadanie 2.16 Znajdź x_1 dla **CGMR** i dla **CGME**.

Zadanie 2.17 Opierając sie na formule trójczłonowej dla przekształconych wielomianów Czebyszewa skonstruuj *wersję trzypoziomową metody Czebyszewa.*

Na koniec zastanówmy się nad oceną szybkości zbieżności dla metod **CG**. Metody te minimalizują na każdym kroku normę reziduum

$$\sqrt{(r_n^T B r_n)} \leq \|R_n(A)\| \sqrt{\|B\|} \|r_0\|.$$

Normę euklidesową macierzy $R_n(A)$ szacujemy tak samo jak w przypadku metody Czebyszewa $\|R_n(A)\| \leq \|R_n\|_{\infty,[a,b]}$. Ale w tych metodach reziduum r_n jest minimalizowane ze względu na *wszystkie wielomiany rezidualne*. Stąd wynika, że aktualne jest oszacowanie dla normy euklidesowej

$$\|r_n\| \leq 2\sqrt{\|B\|} (\frac{\sqrt{\kappa(A)} - 1}{\sqrt{\kappa(A)} + 1})^n,$$

gdzie $\kappa(A)$ jest współczynnikiem uwarunkowania dla normy euklidesowej macierzy A.

Preconditing w metodach wielomianowych

Oszacowanie normy reziduum dla metod gradientowych, ktore jest też *granicznym, optymalnym* oszacowaniem dla metody Czebszewa

$$\|r_n\| \leq 2\sqrt{\|B\|}(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1})^n,$$

wskazuje na to jak bardzo ważną rolę dla zbieżności tych metod odgrwa *współczynnik uwarunkowania macierzy układu A*. Dlatego prawie nigdy nie stosuje się takiego algorytmu bez włączenia do niego *preconditingu*, to jest takiego wstępnego przekształcenia układu równań

$$Ax = d$$

na równoważny układ

$$\tilde{A}y = \tilde{d},$$

dla którego współczynnik $\kappa(\tilde{A})$ jest znacznie mniejszy niż $\kappa(A)$. Szczęśliwie się składa, że takie przekształcenie daje się stosunkowo łatwo włączyć odrazu do każdego procesu podobnego do procesu iteracyjnego metody gradientów sprzężonych. Najczęściej stosowane metody *preconditingu* polegają na znalezieniu macierzy M *blizkiej* macierzy A, na przykład w tym sensie, że $\kappa(M^{-1}A)$ *jest blizkie* 1, dla której rozwiązanie układu

$$Mz = b$$

jest *łatwe*. Podamy jeden z takich sposobów.[9] Ponieważ $A = A^T$ jest dodatnio określona, to macierzy *preconditingu* M będziemy również szukać wśród macierzy symetrycznych i dodatnio określonych. Każda macierz symetryczna i dodatnio określona *ma pierwiastek*, to znaczy istnieje taka macierz symetryczna i dodatnio określona, której kwadrat jest równy tej macierzy. Tak więc $M = CC$ i $A = GG$. Przyjmiemy $\tilde{A} = C^{-1}AC^{-1}$ i $\tilde{d} = C^{-1}d$. Uzasadnienie takiego rzekształcenia jest takie: $\tilde{A} = C^{-1}GGC^{-1}$, a ponieważ macierz M

[9] Patrz książka: G.H.Goloub & C.F.van Loan "Matrix Computation".

była blizka macierzy A, to macierz C powinna być blizka macierzy G, a więc współczynnik uwarunkowania $cond(C^{-1}GGC^{-1})$ powinien być nie wielki.

Algorytm zbudujemy tak, że wyliczanie pierwiastka $C = \sqrt{M}$ nie będzie wogóle potrzebne. Pokażemy jak to zrobić na przykładzie algorytmu **CGME**. Zastosujmy ten algorytm do układu $\tilde{A}y = \tilde{d}$.

Jeśli mamy już wyznaczone y_j dla $j = 0, 1, \cdots, n$, to

$$y_{n+1} = \frac{1}{\tilde{\alpha}_{n,n-1} + \tilde{\alpha}_{n,n}}[s_n + \tilde{\alpha}_{n,n-1}y_{n-1} + \tilde{\alpha}_{n,n}y_n],$$

gdzie $s_n = \tilde{d} - \tilde{A}y_n = C^{-1}(d - AC^{-1}y_n)$ jest reziduum na n-tym kroku tego procesu iteracyjnego, zaś

$$\tilde{\alpha}_{n,n-1} = \frac{s_{n-1}^T\tilde{A}s_n}{s_{n-1}^T s_{n-1}},$$

$$\tilde{\alpha}_{n,n} = \frac{s_n^T\tilde{A}s_n}{s_n^T s_n}.$$

Przyjrzyjmy się wzorom na współczynniki $\tilde{\alpha}_{n,j}$, $j = n-1, n$. Wygodnie będzie oznaczyć teraz

$$x_k = C^{-1}y_k,$$

dla $k = 0, 1, 2, \cdots$; wtedy $s_k = C^{-1}r_k$, gdzie $r_k = d - Ax_k$. Na przykład dla $\tilde{\alpha}_{n,n-1}$

$$\tilde{\alpha}_{n,n-1} = \frac{s_{n-1}^T\tilde{A}s_n}{s_{n-1}^T s_{n-1}} = \frac{r_{n-1}^T C^{-1}C^{-1}AC^{-1}C^{1}r_n}{r_{n-1}^T C^{-1}C^{-1}r_{n-1}}$$

Niech z_j będzie rozwiązaniem układu *preconditionera*

$$Mz_j = r_j.$$

Mamy więc

(2.37)
$$\tilde{\alpha}_{n,n-1} = \frac{z_{n-1}^T Az_n}{z_{n-1}^T r_{n-1}},$$

i podobnie

(2.38)
$$\tilde{\alpha}_{n,n} = \frac{z_n^T Az_n}{z_n^T r_n},$$

gdzie, zgodnie z naszymi oznaczeniami $r_k = d - Ax_k$.

W ten sposób, n-ty krok algorytmu *CGME z preconditingiem* ma następującą postać:

- Przypuśćmy, że już mamy $x_0, x_1, \cdots, x_n$ oraz z_n wyliczone jako rozwiązanie *łatwego układu preconditionera*

$$Mz_n = r_n, \quad r_n = d - Ax_n.$$

- Wyliczamy $\tilde{\alpha}_{n,n-1}$ i $\tilde{\alpha}_{n,n}$ przy pomocy wzorów (2.37) i (2.38), oraz

$$x_{n+1} = \frac{1}{\tilde{\alpha}_{n,n-1} + \tilde{\alpha}_{n,n}}[z_n + \tilde{\alpha}_{n,n-1}x_{n-1} + \tilde{\alpha}_{n,n}x_n], \tag{2.39}$$

  gdyż $x_{n+1} = C^{-1}y_{n+1}$ i, jak łatwo zauważyć

$$s_n = C^{-1}r_n = C^{-1}Mz_n = C^{-1}CCz_n = Cz_n.$$

  Na koniec wyliczamy $r_{n+1} = d - Ax_{n+1}$, oraz z_{n+1}, rozwiązując *układ preconditionera* $Mz_{n+1} = r_{n+1}$.

Aby proces mógł wystartować potrzebne są dwa punkty x_0 i x_1. Punkt x_0 wybieramy dowolnie oraz kładziemy

$$x_1 = x_0 + \frac{z_0}{q_0},$$

gdzie

$$q_0 = \frac{z_0^T A z_0}{z_0^T r_0}.$$

Zadanie 2.18

- wyjaśnij dla czego właśnie tak należy wybrać x_1,
- zbuduj wzory dla algorytmu **CGMR** z preconditingiem.

Tak więc **algorytm CGME z preconditingiem** różni się tylko tym od oryginalnego algorytmu **CGME**, że na każdym kroku wyliczamy dodatkowo wektor z_k rozwiązując *łatwy układ preconditionera* $Mz_k = r_k$. W algorytmie nie występuje nigdzie macierz $C = \sqrt{M}$.

Inna wersja metod typu CG [10]

Przedstawimy tu, na przykładzie metody **CGME**, inną, równoważną z punktu widzenia arytmetyki "dokładnej", wersję metody **CGME**. Wersja ta pochodzi (prawdopodobnie) od G. Golub'a. Doświadczenia pokazują, że przedstawiona poniżej wersja algorytmu (będziemy ją oznaczać skrótem **CGGG**), radzi sobie lepiej w praktyce obliczeniowej. Wydaje się, że algorytmy oparte bezpośrednio na formule 3-członowej, gdy współczynniki muszą być wyliczane w trakcie biegu algorytmu, napotykają na podobne trudności numeryczne jak, na przykład, algorytm Gramma-Schmidta. Zauważmy odrazu, że nie dotyczy to żadnej z wersji metody Czebyszewa (zastanów się dlaczego).

Wersja **CGGG** nie jest oparta na algorytmie ortogonalizacyjnym, lecz na znajdowaniu minimum funkcjonału, i to stopniowo, poprzez rozwiązywanie kolejno jednowymiarowych zadań na minimum. Często się zdarza, że podobne algorytmy, które stopniowo modyfikują dane wejściowe, są bardziej odporne na destrukcyjne działanie błędów zaokrąglenia. Istnieje podobny algorytm iteracyjny przeznaczony dla zadań o macierzach dowolnych, odwracalnych (patrz algorytm Y.Saada GMRES).

Przedstawiona poprzednio teoria metod **CG** nie staje się w ten sposób bezużyteczna, gdyż dostarcza nam wiele istotnych informacji: na przykład o szybkości zbieżności takich metod.

Będziemy zajmować się, jak poprzednio, układem równań liniowych algebraicznych wymiaru $n \times n$

$$Ax = d, \tag{2.40}$$

gdzie macierz A jest rzeczywista, symetryczna i dodatnio określona. Dla naszego równania (2.40) określimy funkcjonał

$$f(x) = \frac{x^T Ax}{2} - x^T d. \tag{2.41}$$

Niech $e = x^* - x$, gdzie x^* jest rozwiązaniem zadania (2.40), będzie "wektorem błędu". Nie trudno zauważyć, że

$$f(x) = \frac{e^T Ae - x^{*T} Ax^*}{2} = \frac{\|e\|_A{}^2 - \|x^*\|_A^2}{2},$$

[10] Patrz G.H.Golub & C.F.van Loan 'Matrix Computations'

zatem funkcjonał f i norma $\|e\|$ (która jest funkcjonałem od zmiennej x!) osiągają zawsze ekstremum w tym samym punkcie.

Zadanie 2.19 Znajdź minimum bezwarunkowe funkcjonału f. W jakim punkcie jest ono osiągane?

Lemmat *Załóżmy, że*

- $d_0, d_1, \cdots, d_{n-1}$ *jest układem ortogonalnym w* $\mathbf{R^n}$ *w sensie iloczynu skalarnego* $(\cdot,\cdot)_A$*, to znaczy* $d_k^T A d_l = \delta_{k,l}\|d_k\|_A^2$*,*
- $x_0, x_1, \cdots, x_{n-1}$ *jest ciągiem wektorów z* $\mathbf{R^n}$*, określonych rekurencyjnie:*

$$x_0 \quad \text{- dowolny,}$$

$$x_{k+1} = x_k + \alpha_k d_k, \quad k = 0, 1, \cdots, n-1, \tag{2.42}$$

gdzie

$$\alpha_k = \frac{d_k^T r_k}{d_k^T A d_k}, \quad r_k = d - Ax_k. \tag{2.43}$$

Wtedy:

- $$x_n = x^* \quad \text{- rozwiązanie zadania,}$$
- $$\forall k, \quad f(x_k) = \min_{z\in V_k} f(x_0 + z),$$

  gdzie $V_k = span\{d_0, d_1, \cdots, d_{k-1}\}$.

Dowód. Ponieważ $d_0, d_1, \cdots, d_{n-1}$ jest bazą ortogonalną w przestrzeni $\mathbf{R^n}$, to rozwiązanie x^*, równania (2.40), oraz x_0 można rozwinąć

$$x^* = \sum_{j=0}^{n-1} c_j d_j, \quad \text{gdzie} \quad c_j = \frac{d_j^T d}{d_j^T A d_j},$$

$$x_0 = \sum_{j=0}^{n-1} \gamma_j d_j, \quad \text{gdzie} \quad \gamma_j = \frac{d_j^T A x_0}{d_j^T A d_j}.$$

Ze wzorów rekurencyjnych (2.42)(2.43) wnioskujemy, że $\forall k, 0 < k \leq n$

$$x_k - x_0 = \alpha_0 d_0 + \alpha_1 d_1 + \cdots + \alpha_{k-1} d_{k-1},$$

gdzie

$$\alpha_j = \frac{d_j^T r_j}{d_j^T A d_j} = \frac{d_j^T (d - A x_j)}{d_j^T A d_j} = c_j - \frac{d_j^T A x_j}{d_j^T A d_j}.$$

Zauważmy, że ze względu na A-ortogonalność bazy

$$d_j^T A x_j = d_j^T A (\sum_{s=0}^{j-1} \alpha_s d_s) = d_j^T A x_0,$$

a więc

(2.44) $$\alpha_j = c_j - \gamma_j, \quad j = 0, 1, \cdots, n-1.$$

Stąd

$$x_n - x_0 = \sum_{j=0}^{n-1} \alpha_j d_j = \sum_{j=0}^{n-1} (c_j - \gamma_j) = x^* - x_0,$$

a więc $x_n = x^*$.

Ponieważ macierz A jest dodatnio określona, aby udowodnić, że

$$f(x_k) = \min_{z \in V_k} f(x_0 + z),$$

wystarczy pokzać, że

$$\forall h \in V_k = span\{d_0, d_1, \cdots, d_{k-1}\}, \quad f'(x_k) h = 0,$$

lub równoważnie, że

$$\forall j = 0, 1, \cdots, k-1, \quad d_j^T (A x_k - d) = 0.$$

Mamy

$$d_j^T (A x_k - d) =$$

$$= d_j^T Ax_0 + \sum_{s=0}^{k-1} \alpha_s d_j^T Ad_s - d_j^T d = d_j^T Ax_0 + \alpha_j d_j^T Ad_j - d_j^T d =$$

$$= (\gamma_j + \alpha_j - c_j) d_j^T Ad_j,$$

i ze względu na to, że $\alpha_j = c_j - \gamma_j$

$$d_j^T (Ax_k - d) = 0.$$

□

Teraz zdefiniujemy algorytm CGGG.

- Wybieramy dowolnie x_0, oraz przyjmujemy $d_0 = r_0 = d - Ax_0$.
- Jeśli już mamy $x_0, x_1, \cdots, x_k$ i $d_0, d_1 \cdots, d_k$, to określamy

$$x_{k+1} = x_k + \alpha_k d_k,$$

$$r_{k+1} = r_k - \alpha_k Ad_k,$$

$$d_{k+1} = r_{k+1} + \beta_k d_k,$$

gdzie

$$\alpha_k = \frac{d_k^T r_k}{d_k^T Ad_k} = \frac{r_k^T r_k}{d_k^T Ad_k}, \quad \beta_k = -\frac{d_k^T Ar_{k+1}}{d_k^T Ad_k} = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}.$$

Zadanie 2.20 Udowodnij, że zaproponowany wybór współczynników α_k i β_k pociąga spełnienie warunków

- niech $\phi(\alpha) = f(x_k + \alpha d_k)$, wtedy $\phi(\alpha_k) = \min_{\alpha \in \mathbf{R}} \phi(\alpha)$,
- $d_{k+1}^T Ad_k = 0$.

Zadanie 2.21 Udowodnij, że współczynniki α_k i β_k można wyrazić w sposób wygodniejszy dla obliczeń

$$\alpha_k = \frac{r_k^T r_k}{d_k^T Ad_k}, \quad \beta_k = \frac{r_{k+1}^T r_{k+1}}{r_k^T r_k}.$$

Zadanie 2.22 Udowodnij (przez indukcję), że dla algorytmu **CGGG**

$$\forall k,\ \ 1 \le k \le n,\ \ V_k = span\{d_0, d_1, \cdots, d_{k-1}\} =$$

$$= span\{r_0, r_1, \cdots, r_{k-1}\} = span\{r_0, Ar_0, \cdots, A^{k-1} r_0\}.$$

Twierdzenie 2.5 *Jeśli* $r_{k-1} \neq 0$, *to dla algorytmu* **CGGG**

1. $V_k = span\{d_0, d_1, \cdots, d_{k-1}\} = span\{r_0, r_1, \cdots, r_{k-1}\} =$
   $= span\{r_0, Ar_0, \cdots, A^{k-1} r_0\}$,
2. *dla* $j < l \le k-1, \quad d_j^T r_l = 0$,
3. $d_l^T A d_j = \delta_{l,j} \|d_l\|_A^2, \qquad 0 \le l, j \le k-1$,
4. *dla* $e_k = x^* - x_k, \quad \|e_k\|_A^2 = \min_{z \in V_k} \|x^* - (x_0 + z)\|_A^2$.

Dowód. Dowód punktu *1.* - patrz **Zadanie 2.22**.
Zastosujemy indukcję jednocześnie do punktów *2.* i *3.* Mamy

$$d_0^T r_1 = r_0^T (r_0 - \alpha_0 r_0) = r_0^T r_0 - \frac{r_0^T r_0}{d_0^T A d_0} d_0^T A d_0 = 0$$

oraz

$$d_0^T A d_1 = d_0^T A (r_1 + \beta_0 d_0) = d_0^T A (r_1 - \frac{r_1^T A d_0}{d_0^T A d_0} d_0) = 0.$$

Krok indukcyjny.
Z założenia indukcyjnego

$$d_j^T r_{l-1} = 0 \ \text{ i } \ d_j^T A d_{l-1}$$

dla $j < l-1$. Zajmiemy się najpierw wyrażeniem $d_j^T r_l$. Niech $j < l-1$; wtedy

$$d_j^T r_l = d_j^T (r_{l-1} - \alpha_{l-1} A d_{l-1}) = d_j^T r_{l-1} - \alpha_{l-1} d_j^T A d_{l-1} = 0,$$

ponieważ $d_j^T r_{l-1} = 0$ i $d_j A d_{l-1} = 0$ z założenia indukcyjnego. Teraz niech $j = l-1$. Mamy

$$d_{l-1}^T r_l = d_{l-1}^T (r_{l-1} - \alpha_{l-1} A d_{l-1}) = d_{l-1}^T r_{l-1} - \frac{d_{l-1}^T r_{l-1}}{d_{l-1}^T A d_{l-1}} d_{l-1}^T A d_{l-1} = 0.$$

Podobnie,

$$d_j^T A d_l = d_j^T A(r_l - \beta_{l-1} d_{l-1}) = d_j^T A r_l - \beta_{l-1} d_j^T A d_{l-1}.$$

Załóżmy najpierw, że $j \leq l-2$. Wtedy z założenia indukcyjnego $d_j^T A d_{l-1} = 0$. Natomiast

$$Ad_j \in span\{Ad_0, Ad_1, \cdots, Ad_{l-2}\} \subset span\{r_0, Ar_0, \cdots, A^{l-1} r_0\} = V_{l-1},$$

i wtedy $d_j^T A r_l = 0$, ponieważ

$$Ad_j = \sum_{s=0}^{l-1} c_s d_s,$$

i udowodniliśmy już, że $d_s^T r_l = 0$ dla $s = 0, 1, \cdots, l-1$. Pozostaje do rozpatrzenia przypadek, gdy $j = l-1$; ale wtedy $d_{l-1}^T A d_l = 0$, z definicji ciągu $d_0, d_1 \cdots$. Wreszcie, warunek *4*.

$$\|e_k\|_A^2 = \min_{z \in V_k} \|x^* - (x_0 + z)\|_A^2$$

wynika stąd, że funkcjonał f osiąga minimum na V_k w tym samym punkcie x_k. Wynika to bezpośrednio z **Lemmatu**, gdyż układ $d_0, d_1, \cdots, d_{k-1}$, który konstruujemy jest A-ortogonalny. □

Wniosek 3. *Algorytmy* **CGGG** *i* **CGME** *są równoważne, gdyż oba w wyniku wykonania k - kroków dają wektor realizujący warunek*

$$\|e_k\| A^2 = \min_{z \in span\{r_0, r_1, \cdots, r_k\}} \|x^* - (x_0 + z)\|_A^2.$$

Aby się o tym przekonać, wystarczy zauważyć, że algorytm **CGME** spełnia na kroku k zależność

$$x_k = x_0 + \frac{r_0}{q_0} + \frac{r_1}{q_1} + \cdots + \frac{r_{k-1}}{q_{k-1}} \in V_k,$$

gdzie $r_k = d - Ax_k$ i $q_j, \quad j = 0, 1, \cdots, k-1$ są pierwiastkami wielomianu rezidualnego $R_k(x)$ dla tego algorytmu. □

Zadanie 2.23 Dobierając odpowiednio macierz wagową B, skonstruuj odpowiednik metody **CGMR** podobny do **CGGG**.

Zadanie 2.24 Wzorując się na sposobie preconditingu opisanym dla metod **CGME** i **CGMR** skonstruuj podobny preconditing dla **CGGG**.

Rozdział 3
ROZWIĄZYWANIE UKŁADÓW RÓWNAŃ LINIOWYCH ALGEBRAICZNYCH

Metody iteracyjne "tradycyjne"

Będziemy zajmować się układami równań algebraicznych liniowych

$$Ax = d, \tag{3.1}$$

gdzie A jest macierzą kwadratową wymiaru $m \times m$ **nieosobliwą**. Rozwiązywanie układów równań liniowych algebraicznych jest jednym z najważniejszych zadan z którymi zajmują się metody numeryczne. Takie bowiem zadania występuja jako części składowe bardzo wielu innych zagadnień numerycznych liniowych i nieliniowych. We współczesnej numeryce mamy często do czynienia z układami o ogromnych rozmiarach, rzedu setek tysięcy równań. Takie zadanie jest *praktycznie niskończenie wymiarowe*. Bardzo wielkie układy dość często odznaczają się regularną budową; są to czesto układy o *macierzach pasmowych* to jest mających niezerowe elementy zgrupowane jedynie na pewnej liczbie diagonal położonych wokół głównej diagonali. Taką szczególną budowę, ze zrozumiałych względów technicznych, staramy się zwykle zachować podczas procesu obliczeń. Dlatego do układów tego typu chętnie stosuje się rozmaite metody *iteracyjne*, których cechą jest to, że podczas działania nie zmieniają macierzy układu. W poprzednim rozdziale poznaliśmy już takie metody: była to **metoda Czebyszewa** oraz dwie wersje **metody gradientów sprzężonych**. Jeśli stosujemy metody iteracyjne, jest ważne, aby dla osiągnięcia wystarczającej dla naszych celów dokładności, wystarczyło wykonać **znacznie mniej iteracji niż wynosi wymiar zadania**. Stąd dbałość o szybkość zbieżności metod iteracyjnych. Ten aspekt sprawy na ogół eliminuje z konkurencji *zwykłe metody bezpośrednie* typu eliminacji Gauß'a. **Metody bezpośrednie** stosujemy na ogół do zadań o nie wielkich rozmiarach. W tym wykładzie będziemy zajmować się jedynie metodami iteracyjnymi.

Przypomnienie.
Normy. W przestrzeni liniowej macierzy kwadratowych można zdefiniować różne normy. Takie normy można podzielić na dwie klasy:

1. **Normy operatorowe** - indukowane przez odpowiednie normy w przestrzeni wektorowej $\mathbf{R^m}$ (lub $\mathbf{C^m}$). Traktujemy wtedy macierz jako *operator* działający na tej przestrzeni wektorowej o wartościach w tej samej przestrzeni. Zgodnie z ogólną definicją **normy operatora**

$$\|A\| = \sup_{\|x\|=1} \|Ax\|.$$

   Po prawej stronie tego wzoru występuje norma $\|\cdot\|$ w przestrzeni wektorowej. Zatem *postać* normy macierzy będzie zależeć od tego jaką normę przyjmiemy w przestrzeni wektorowej.

2. Macierz kwadratową wymiaru $m \times m$ można także traktować jako *wektor* wymiaru m^2. Można więc używać również normy wektora z tej przestrzeni jako normy macierzy. Przykładem takiej normy jest **norma Frobieniusa**

$$\|A\|_F = (\sum_{i,j=1}^{m} a_{i,j}^2)^{\frac{1}{2}}.$$

   Oczywiście normy tego typu mają całkiem inne własności niż *normy operatorowe*.

Najczęściej używane **normy operatorowe** macierzy to:

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

   Odpowiada ona normie wektorowej $\|x\|_\infty = \max_{1\le i\le m} |x_i|$.

2. $$\|A\|_1 = \max_{1\le j\le m} \sum_{i=1}^{m} |a_{i,j}|.$$

   Odpowiada ona normie wektorowej $\|x\|_1 = \sum_{j=1}^{m} |x_j|$.

3.
$$\|A\| = \max_{1\le j\le m} \sqrt{s_j}.$$

Odpowiada ona *euklidesowej normie wektorowej*

$$\|x\| = (\sum_{j=1}^{m} |x_j|^2)^{\frac{1}{2}}.$$

Liczby s_j, $s_j \geq 0$, $j = 1, 2, \cdots, m$ są *wartościami szczególnymi* macierzy A, to jest wartościami własnymi macierzy $A^T A$.

Zadanie 3.1 Udowodnij, że wzory podane powyżej określają normy operatorowe macierzy indukowane przez podane normy wektorowe. Które z tych norm są łatwe do obliczenia?

Uwarunkowanie. Jeśli dane układu równań $Ax = d$ zaburzymy przy pomocy *niewielkich zaburzeń* macierzy A, ΔA i wektora d, Δd, to rozwiązanie x zaburzy się i będzie postaci $x + \Delta x$. *Względne zaburzenie rozwiązania* $\frac{\|\Delta x\|}{\|x\|}$ liczone w ustalonej normie wektorowej można oszacować w zależności od *współczynnika uwarunkowania macierzy* A, $cond(A) = \|A\|\|A^{-1}\|$ liczonego w *odpowiedniej normie macierzy.* Zachodzi oszacowanie

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{cond(A)(\frac{\|\Delta d\|}{\|d\|} + \frac{\|\Delta A\|}{\|A\|})}{1 - cond(A)\frac{\|\Delta A\|}{\|A\|}}. \tag{3.2}$$

Zadanie 3.2 Udowodnij, że zachodzi nierówność (3.2).

Wzór (3.2) pokazuje, jak ważną rolę odgrywa współczynnik uwarunkowania macierzy przy numerycznym rozwiązywaniu układu (3.1).

Metody bezpośrednie. Wspomnimy tu tylko najważniejsze algorytmy.

Eliminacja Gauß'a.
Algorytm składa się z dwóch kroków

- Sprowadzenie układu do postaci trójkątnej

$$\begin{bmatrix} x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \\ x & x & x & x & x \end{bmatrix} x = d \Longrightarrow \begin{bmatrix} x & x & x & x & x \\ \cdot & x & x & x & x \\ \cdot & \cdot & x & x & x \\ \cdot & \cdot & \cdot & x & x \\ \cdot & \cdot & \cdot & \cdot & x \end{bmatrix} x = \tilde{d}.$$

  Odmiany:

  - bez wyboru głównego elementu,
  - z częściowym wyborem głównego elementu,
  - z pełnym wyborem głównego elementu.

- Rozwiązanie układu o macierzy trójkątnej.

Metoda Householdera. Jest to *rozkład typu* $A = QR$, gdzie Q - macierz ortogonalna, R - macierz trójkątna górna. Macierz Q jest iloczynem $m - 1$ *macierzy Householdera* zbdowanych przy pomocy macierzy postaci $H = I - 2uu^T$, gdzie $u^T u = 1$; są to macierze ortogonalne i symetryczne

$$Q = H_{m-1} H_{m-2} \cdots H_1.$$

Macierz Householdera H_j *eliminuje* j-tą kolumę macierzy A to znaczy doprowadza ją do takiej postaci, że poniżej elementu o nomerze j występują tylko zera.

Zadanie 3.3 Przypomnij jak wygląda algorytm Householdera, jak wyznacza się macierze H_j, jakie są cechy tej metody.

Metoda Cholesky'ego - Banachiewicza. Polega na rozkładzie macierzy symetrycznej i dodatnio określonej A na iloczyn $A = LL^T$ gdzie L jest macierzą trójkątną dolną. Następnie rozwiązujemy dwa układy trójkątne. Wersja tej metody w zastosowaniu do układu o macierzy *trójdiagonalnej* nosi popularną nazwę *metody progonki.*

Zadanie 3.4 Przypomnij dowód istnienia rozkładu $A = LL^T$ dla dowolnej macierzy symetrycznej i dodatnio określonej A, oraz algorytm rozkładu Cholesky'ego - Banachiewicza.

Klasyczne metody iteracyjne

Ogólny dwupoziomowy schemat iteracyjny.
Nasz układ $Ax = d$ przekształcamy w dowolny sposób do postaci

$$x = Cx + b, \tag{3.3}$$

tak, aby układy były równoważne.

Przykłady typowych przekształceń do postaci (3.3).

1. $x = x + \kappa r$, gdzie κ jest skalarem $\kappa \neq 0$, zaś r jest *reziduum* $r = d - Ax$. Mamy wtedy

   $$x = (I - \kappa A)x + \kappa d. \tag{3.4}$$

2. Ogólniej, $x = x + Br$, gdzie B jest macierzą odwracalną. Otrzymamy

   $$x = (I - BA)x + Bd. \tag{3.5}$$

3. Zawsze możemy napisać $A = L + D + U$, gdzie

   - L - trójkątna dolna bez diagonali *(Left)*,
   - D - diadonala,
   - U - trójkątna górna bezdiagonali *(Upper)*.

   Jeśli D^{-1} istnieje, to mamy następujące często używane formy typu (3.3):

   - Postac Jordana

     $$x = -D^{-1}(L + U)x + D^{-1}d, \tag{3.6}$$

   - Gauß- Seidel

     $$x = -(D + L)^{-1}U + (D + L)^{-1}d. \tag{3.7}$$

   - Podrelaksacja - Nadrelaksacja. Niech $\omega \neq 0$,

     $$Dx = (1 - \omega)Dx - \omega[(L + U)x - d]. \tag{3.8}$$

Mając równanie postaci (3.3)

$$x = Cx + b,$$

możemy, startując od dowolnego wektora $x_0 \in \mathbf{R^m}$, wygenerować ciąg

$$x_0, x_1, x_2, \cdots,$$

przy pomocy procesu iteracyjnego

$$x_{n+1} = Cx_n + b. \tag{3.9}$$

Zauważmy, że proces (3.9) w trakcie działania nie zmienia macierzy układu. Ponadto, łatwo zauważyć, że jeśli proces (3.9) zbiega, to zbiega do rozwiązania x równania $Ax = d$.

Każdej z wymienionych wyżej form układu równań odpowiada pewna metoda iteracyjna.

1. Metoda Iteracyjna Richardsona

   $$x_{n+1} = x_n + \kappa r_n,$$

   gdzie r_n jest *reziduum na n-tym kroku*: $r_n = d - Ax_n$. Prces ten w ogólniejszej postaci poznaliśmy już przy omawianiu metody Czebyszewa i metod gradientów sprzężonych. Teraz *współczynnik relaksacji* κ jest stały.

2. Metoda Jacobiego

   $$Dx_{n+1} = -(L + U)x_n + d,$$

   wymaga rozwiązania na każdym kroku iteracji układu równań z macierzą diagonalną D.

3. Metoda Gauß'a -Seidel'a

   $$(D + L)x_{n+1} = -Ux_n + d$$

   wymaga rozwiązania na każdym kroku iteracji układu równań z macierzą trójkątną dolną $D + L$.

4. Metoda nad (pod) - relaksacji

$$(D + \omega L)x_{n+1} = [D(1 - \omega) - \omega U]x_n + \omega d$$

jest uogólnieniem metody Gauß'a - Seidel'a, (metodę Gauß'a - Seidel'a otrzymujemy dla $\omega = 1$).

Warunki zbieżności procesu iteracyjnego (3.9)
Niech $x \in \mathbf{R^m}$ będzie rozwiązaniem równania (3.3). Odejmując stronami równania

$$x = Cx + b$$

i

$$x_{n+1} = Cx_n + b$$

otrzymamy

$$e_{n+1} = Ce_n,$$

gdzie oznaczyliśmy $e_k = x - x_k$ - *błąd na k-tym kroku iteracji.* Otrzymujemy stąd

$$e_k = C^k e_0. \tag{3.10}$$

Stąd $\|e_n\| = \|C^n e_0\| \leq \|C\|^k \|e_0\|$, dla dowolnej normy $\|\cdot\|$. Widzimy więc, że $\|e_n\| \to 0$, gdy $n \to \infty$, jeśli $\|C\| < 1$.
Zatem, warunkiem dostatecznym zbieżności ciągu (3.9) jest $\|C\| <$ 1. Warunek konieczny i dostateczny zbieżności procesu iteracyjnego (3.9) podaje następujące

Twierdzenie 3.1. *Ciąg $\{x_k\}_{k=1,2,\ldots}$ określony procesem iteracyjnym $x_{n+1} = Cx_n + b$ jest zbieżny do rozwiązania x układu $Ax = d$ wtedy i tylko wtedy, gdy wszystkie wartości własne macierzy C mają moduły < 1.*

Dowód. Dowód przeprowadzimy w przypadku, gdy $C = C^T$. Mamy wtedy, po zastosowaniu Twierdzenia Jordana o rozkładzie spektralnym

$$C = Q\Lambda Q^T \quad \text{gdzie} \quad Q^TQ = QQ^T = I,$$

zaś Λ jest macierzą diagonalną, na jej diagonali leżą wartości własne A.

$$\Lambda = \begin{bmatrix} \lambda_1 & \cdot & \cdot & \cdots & \cdot \\ \cdot & \lambda_2 & \cdot & \cdots & \cdot \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdot & \cdot & \cdot & \cdots & \lambda_m \end{bmatrix}$$

Stąd $C^k = Q\Lambda^k Q^T$ i jasne jest, że $C^k \to 0$, gdy $k \to \infty$, wtedy i tylko wtedy, gdy dla każdego $j = 1, 2, \cdots, m$, $|\lambda_j| < 1$. $\Box$.

Zadanie 3.5 Posługując się rozkładem spektralnym Jordana na *klatki jordanowskie* dowolnej macierzy kwadratowej C, przeprowadź dowód twierdzenia 3.1 bez założenia o tym, że $C^T = C$.

Powyższe twierdzenia pozwalają znaleźć warunki zbieżności dla niektórych z opisanych procesów.

Zadanie 3.6 Udowodnij, że warunkiem dostatecznym zbieżności Metody Jacobiego dla układu $Ax = d$ jest istnienie takiej liczby ρ, $0 \le \rho < 1$ że dla każdego $i = 1, 2, \cdots, m$

$$\sum_{j=1}^{m} |a_{i,j}| \le \rho |a_{i,i}|,$$

gdzie $A = (a_{i,j})_{i,j=1,2,\cdots,m}$
Uwaga Nie potrzebne tu jest założenie o symetrii macierzy A.

Przeprowadzimy dyskusję zbieżności procesu iteracyjnego Richardsona. Założymy teraz, że macierz A jest symetryczna i dodatnio określona. Zatem dla każdej wartości własnej λ_j macierzy A zachodzi warunek

$$0 < \lambda_m \le \lambda_j \le \lambda_M,$$

gdzie λ_m - minimalna, zaś λ_M - maksymalna wartość własna macierzy A. Warunkiem koniecznym i dostatecznym zbieżności procesu Richardsona

$$x_{n+1} = x_n + \kappa r_n$$

jest, aby widmo macierzy $C = I - \kappa A$ było zawarte w przedziale otwartym $(-1, 1)$. Każda wartóść własna λ_C macierzy C jest związana zależnością

$$\lambda_C = 1 - \kappa \lambda_A,$$

z pewną wartością własną λ_A macierzy A. Stąd mamy warunek konieczny i dostateczny zbieżności

$$0 < \kappa\lambda_M < 2,$$

lub

$$\kappa < \frac{2}{\lambda_M} = \frac{2}{\|A\|},$$

gdzie $\|\cdot\|$ jest normą euklidesową macierzy. Iteracja jest najszybciej zbieżna gdy maksymalna co do modułu wartość własna macierzy C osiąga wartość minimalną. Nie trudno stwierdzić, rozważając trójkąty utworzone przez prostą o równaniu $y = 1 - \kappa x$, oś x oraz odcinki równoległe do osi y wyprowadzone z punktów $x = \lambda_m$ i λ_M w kierunku tej prostej, że warunek ten jest spełniony, gdy współczynnik relaksacji κ przyjmuje wartość

$$\kappa_{opt} = \frac{2}{\lambda_m + \lambda_M}.$$

Odpowiada to sytuacji, w której wspomniana prosta przecina oś x w środku odcinka $[\lambda_m, \lambda_M]$. Moduł maksymalnej wartości własnej macierzy C dla $\kappa = \kappa_{opt}$ (jest to norma euklidesowa tej macierzy) jest równy współczynnikowi zbieżności metody Richardsona w przypadku optymalnym. Łatwo obliczamy ten współczynnik:

$$\frac{\frac{\lambda_M}{\lambda_m} - 1}{\frac{\lambda_M}{\lambda_m} + 1} = \frac{cond(A) - 1}{cond(A) + 1}.$$

Warto porównać ten współczynnik ze współczynnikiem zbieżności Metod Gradientów Sprzężonych, które można uważać również za metody Richardsona, jednak *ze zmiennym współczynnikiem relaksacji.* Dla Metod Gradientów Sprzężonych wyprowadziliśmy:

$$\frac{\sqrt{cond(A)} - 1}{\sqrt{cond(A)} + 1}.$$

Ponieważ $cond(A) > 1$ to współczynnik zbieżności dla metod Gradientów sprzężonych jest mniejszy, a więc Metody Gradientów Sprzężonych zbiegają szybciej niż rozważana tu metoda.

Procesy iteracyjne dwupoziomowe w postaci kanonicznej

Dla układu $Ax = d$ będziemy rozważać procesy iteracyjne dwupoziomowe w *postaci kanonicznej*

$$B\frac{x_{n+1} - x_n}{\tau} + Ax_n = d, \tag{3.11}$$

gdzie $\tau > 0$ jest stałą, zaś B jest pewną macierzą nieosobliwą. Zauważmy odrazu, że jeśli proces iteracyjny (3.11) jest zbieżny do pewnego wektora x, to granica x ciągu $x_0, x_1, \cdots$ wygenerowanego przez proces (3.11) jest rozwiązaniem układu równań $Ax = d$. Procesowi iteracyjnemu (3.11) można nadać postać (3.9)

$$x_{n+1} = x_n - \tau B^{-1}(d - Ax_n).$$

Macierzy B we wzorze (3.11) można nadać następującą interpretację. Wzory te można zapisać w równoważnej postaci

$$\frac{x_{n+1} - x_n}{\tau} + B^{-1}Ax_n = B^{-1}d.$$

a powyższy proces iteracyjny rozwiązuje układ równań postaci

$$B^{-1}Ax = B^{-1}d$$

równoważny układowi oryginalnemu $Ax = d$. Ten nowy układ, może mieć lepsze własności numeryczne, jeśli odpowiednio dobierzemy macierz B. Przez właściwy dobór B możemy, na przykład, *obniżyć współczynnik uwarunkowania macierzy układu*:

$$cond(B^{-1}A) \ll cond(A).$$

Operacja przejścia od układu $Ax = d$ do równoważnego układu $B^{-1}Ax = B^{-1}d$ o mniejszym współczynniku uwarunkowania, to znany już nam z poprzedniego rozdziału *preconditing*. Zatem, możemy uważać, że proces (3.11) *zawiera w sobie operację preconditingu*. Aby uzyskać pożądany efekt powinno się jako B dobierać macierz bliską A, ale taką, żeby układ $Bz = g$ był *łatwy do rozwiązania*, posługiwanie się procesem (3.11) wymaga bowiem na każdym kroku rozwiązania układu równań z macierzą B.

Wygodnie nam będzie teraz operować *relacją nierówności między macierzami*. Niech A i B będą macierzami kwadratowymi wymiaru $m \times m$. Relację tę rozumiemy w następujący sposób

$$A \geq (>) B$$

wtedy i tylko wtedy, gdy dla każdego niezerowego wektora $x \in \mathbf{R^m}$

$$((A - B)x, x) \geq (>) 0.$$

Jeśli $A = A^T > 0$, to z tą macierzą możemy związać *nowy iloczyn skalarny i nową normę w* $\mathbf{R^m}$ $(x, y)_A = (Ax, y)$ i $\|x\|_A = \sqrt{(Ax, x)}$. Normami tego rodzaju posługiwaliśmy się już poprzednio.
Komentarz. Jak dobrze wiadomo, wszystkie rodzaje norm są *równoważne* w przestrzeni $\mathbf{R^m}$; oznacza to, że dla dwóch dowolnych norm $\|\cdot\|_1$ i $\|\cdot\|_2$ w $\mathbf{R^m}$ istnieją stałe dodatnie α i β, takie, że dla $x \in \mathbf{R^m}$

$$\alpha \|x\|_1 \leq \|x\|_2 \leq \beta \|x\|_1.$$

Potrzeba rozróżniania norm nie jest tak wyraźna jeśli interesujemy się *tylko jedną przestrzenią*. Potrzeba ta jednak staje się znacznie bardziej ewidentna, gdy mamy do czynienia nie z jednym zadaniem w ustalonej przestrzeni, ale *z ciągiem zadań w ciągu przestrzeni skończonego wymiaru.* Z taką sytuacją spotykamy się dość często, na przykład rozważając układy równań liniowych otrzymane z aproksymacji równań różniczkowych. Stałe α i β, określające równoważność norm $\|\cdot\|_1$ i $\|\cdot\|_2$ *mogą zalezyć od* n.

Potrzebna będzie nierówność, której prosty dowód proponujemy jako

Zadanie 3.7 Niech $B > 0$ (nie zakładamy symetrii macierzy B) i niech $x \in \mathbf{R^m}$, $x \neq 0$. Udowodnij posługując się rozkładem spektralnym macierzy symetrycznej, że

$$(Bx, x) = (\frac{B + B^T}{2} x, x)$$

zaś wyrażenie (Bx,x) można oszacować z dołu i z góry w następujący sposób

$$0 < \lambda_{min} \|x\|^2 \leq (Bx, x) \leq \lambda_{max} \|x\|^2,$$

gdzie $0 < \lambda_{min} \leq \lambda_{max}$ to najmniejsza i największa wartość własna macierzy

$$\frac{B + B^T}{2}.$$

Twierdzenie 3.2 *Rozważamy układ równań*

$$Ax = d,$$

oraz proces iteracyjny dla tego układu

$$B\frac{x_{n+1} - x_n}{\tau} + Ax_n = d \quad \tau > 0, \quad x_0 \text{ - dowolne.}$$

Jeśli $A = A^T > 0$, oraz jeśli $B - \frac{\tau}{2}A > 0$, to proces iteracyjny jest zbieżny do rozwiązania x rozważanego układu równań w normie $\|\cdot\|_A$. Inaczej mówiąc $\|x_n - x\|_A \to 0$ gdy $n \to \infty$.

Dowód. Jeśli x jest rozwiązaniem, to

$$B\frac{x - x}{\tau} + Ax = d,$$

i oznaczając $e_n = x - x_n$ (błąd na n-tym kroku iteracji) otrzymamy

$$B\frac{e_{n+1} - e_n}{\tau} + Ae_n = 0. \tag{3.12}$$

Jest to *równanie błędu*. Zauważmy, że

$$e_n = \frac{e_{n+1} + e_n}{2} - \frac{\tau}{2}\frac{e_{k+1} - e_k}{\tau}.$$

Wstawiając to wyrażenie do (3.12), dostaniemy

$$(B - \frac{\tau}{2}A)\frac{e_{n+1} - e_n}{\tau} + A\frac{e_{n+1} + e_n}{2} = 0.$$

To ostatnie równanie pomnożymy skalarnie przez $2(e_{n+1} - e_n)$; otrzymamy

$$2\tau((B - \frac{\tau}{2}A)\frac{e_{n+1} - e_n}{\tau}, \frac{e_{n+1} - e_n}{\tau}) + \|e_{n+1}\|_A^2 - \|e_n\|_A^2 = 0,$$

skąd ze względu na warunek

$$((B-\frac{\tau}{2}A)\frac{e_{n+1}-e_n}{\tau},\frac{e_{n+1}-e_n}{\tau})\geq 0$$

mamy

$$0\leq \|e_{n+1}\|_A^2\leq \|e_n\|_A^2.$$

Wynika stąd, że ciąg liczbowy $\{\|e_n\|_A^2\}$ jest zbieżny, jako ciąg malejący i ograniczony z dołu przez 0. Pozostaje więc pokazać, że zbiega on do zera. Ze zbieżności tego ciągu wynika, że $\|e_{n-1}\|^2-\|e_n\|^2$ zbiega do zera, a zatem $((B-\frac{\tau}{2}A)\frac{e_{n+1}-e_n}{\tau},\frac{e_{n+1}-e_n}{\tau})$ także zbiega do zera. Korzystając teraz z nierówności udowodnionej w **Zadaniu 3.7** wnosimy, że istnieje stała dodatnia $\lambda>0$, dla której

$$\lambda\|e_{n+1}-e_n\|^2\leq ((B-\frac{\tau}{2}A)\frac{e_{n+1}-e_n}{\tau},\frac{e_{n+1}-e_n}{\tau})\to 0,$$

gdy $n\to\infty$, awięc $\|e_{n+1}-e_n\|\to 0$. Z równania (3.12)

$$Ae_n=-B\frac{e_{n+1}-e_n}{\tau}.$$

Ponieważ $A=A^T>0$ to istnieje macierz "pierwiastek z A", także symetryczna i dodatnio określona $A^{\frac{1}{2}}$. Istnieje więc także $A^{-\frac{1}{2}}$. Mnożąc ostatnią równość przez tę macierz dostaniemy

$$A^{\frac{1}{2}}e_n=-A^{-\frac{1}{2}}B\frac{e_{n+1}-e_n}{\tau}.$$

Stąd

$$(A^{-\frac{1}{2}}e_n,A^{-\frac{1}{2}}e_n)=(Ae_n,e_n)=\|e_n\|_A^2=$$

$$=\|A^{-\frac{1}{2}}e_n\|^2\leq \|A^{-\frac{1}{2}}\|^2\|B\|^2\frac{\|e_{n+1}-e_n\|^2}{\tau^2}\to 0,\quad \text{gdy}\quad n\to\infty.$$

$\square$.

Twierdzenie to można wykorzystać przy dowodzie zbieżności procesu iteracyjnego Gauß'a - Seidela. Załóżmy znów, że $A=A^T>0$. Mamy

$$(L+D)x_{n+1}=-Ux_n+d.$$

Zapiszemy ten proces w *postaci kanonicznej* (dodajemy stronami $-(D+L)x_n$)

$$(L+D)(x_{n+1}-x_n)+Ax_n=d.$$

Zatem $B = L + D$ i $\tau = 1$, więc

$$B - \frac{\tau}{2}A = L + D - \frac{L}{2} - \frac{D}{2} - \frac{L^T}{2} = \frac{L}{2} + \frac{D}{2} - \frac{L^T}{2}.$$

Trzeba sprawdzić, czy $B - \frac{\tau}{2}A > 0$.

$$((\frac{L - L^T}{2} + \frac{D}{2})x, x) = \frac{1}{2}((Lx, x) - (L^T x, x) + (Dx, x)).$$

Ale $(Lx, x) = (L^T x, x)$, zatem $B - \frac{\tau}{2}A = \frac{1}{2}(Dx, x) > 0$, ponieważ macierz dodatnio określona ma diagonalę dodatnią (**dlaczego?**).
Zatem proces Gauß'a - Seidela jest zbieżny zawsze, gdy macierz układu A jest symetryczna i dodatnio określona.

Zadanie 3.8 Zbadaj zbieżność procesu iteracyjnego nad - pod relaksacji dla układu $Ax = d$ z macierzą A symetryczną i dodatnio określoną:

$$(D + \omega L)x_{n+1} = [(1 - \omega)D - U\omega]x_n + \omega d, \quad x_0 \text{ - dowolne,}$$

gdzie współczynnik $\omega > 0$.

Udowodnij, że proces jest zbieżny, gdy $0 \leq \omega < 2$. Dla $0 < \omega < 1$ proces nazywa się *podrelaksacją* zaś dla $1 < \omega < 2$ *nadrelaksacją*. Dla $\omega = 1$ to poprostu proces Gauß'a - Seidela. Wiadomo, że przy dodatkowych założeniach o macierzy układu *istnieje optymalna wartość parametru* ω, przy której proces zbiega znacznie szybciej niż przy innych jego wartościach. Metoda z taką wartością parametru jest nadrelaksacją i nosi nazwę (SOR).

Rozdział 4
KWADRATURY NUMERYCZNE

Będziemy zajmować się teraz aproksymacją całek. Zauważmy odrazu, że jedynie bardzo nieliczne funkcje potrafimy scałkować poprzez wykorzystanie wzorów. Dlatego bardzo ważnym zadaniem obliczeniowym jest numeryczne, przybliżone obliczanie całek.

Niech $\rho : [a, b] \to \mathbf{R}^+$ będzie funkcją całkowalną, - *funkcją wagą*. Będziemy zajmować się aproksymacją funkcjonału

$$I(f) = \int_a^b \rho(x) f(x) dx, \tag{4.1}$$

gdzie *funkcja - waga* ρ jest ustalona, zaś argumentem funkcjonału jest funkcja ciągła $f : [a, b] \to \mathbf{R}$. I jest funkcjonałem *ograniczonym*, a więc ciągłym, określonym na przestrzeni Banacha $C([a, b])$ wyposażonej znaną nam dobrze normę $\| \cdot \|_{\infty,[a,b]}$.

Zadanie 4.1 Oblicz normę funkcjanału I.

Funkcjonał (4.1) będziemy starali się aproksymować innym funkcjonałem *nad przestrzenią* $C([a, b])$ - *kwadraturą numeryczną*. Tutaj będziemy rozpatrywać jedynie *kwadratury* postaci

$$Q(f) = \sum_{j=0}^{m} A_j f(x_j). \tag{4.2}$$

Liczby $a \le x_0 < x_1 < \cdots, < x_m \le b$ noszą nazwę *węzłów* kwadratury (4.2), zaś A_j $j = 0, 1, 2, \cdots, m$ to *współczynniki* tej kwadratury.

Definicja. *Kwadratura numeryczna* (4.2) *jest rzędu* p*, jeśli dla każdego wielomianu* P *stopnia* $< p$ *zachodzi* $Q(P) = I(P)$*, zaś istnieje wielomian* P_0 *stopnia* p*, taki że* $Q(P_0) \ne I(P_0)$.

Obliczmy normę funkcjonału Q. Z definicji

$$\|Q\| = \sup_{\|f\|_{\infty,[a,b]}=1} |Q(f)|.$$

Oszacujemy z góry, dla f spełniającego warunek $\|f\|_{\infty,[a,b]} = 1$:

$$|Q(f)| \leq \sum_{j=0}^{m} |A_j||f(x_j)| \leq \sum_{j=0}^{m} |A_j|.$$

Stąd, ze względu na to, że wyrażenie $\sum_{j=0}^{m} |A_j|$ nie zależy od f,

$$\|Q\| \leq \sum_{j=0}^{m} |A_j|.$$

Wystarczy teraz pokazać, że wartość $\sum_{j=0}^{m} |A_j|$ jest osiągana dla pewnej funkcji o normie równej 1. Nie trudno taką funkcję ciągłą znaleźć:

$$f_0(x) = \begin{cases} sgn(A_j) & \text{dla } \ x = x_j, \quad j = 0, 1, \cdots, m \\ \text{liniowa ciągła} & \text{dla innych wartości} \quad x \in [a, b] \end{cases}$$

Ostatecznie mamy

$$\|Q\| = \sum_{j=0}^{m} |A_j|.$$

Przypuśćmy, że dany jest *ciąg układów węzłów* w przedziale $[a, b]$,

$$a \leq x_0^m < x_1^m < x_2^m < \cdots < x_m^m \leq b,$$

oraz związany z nim ciąg kwadratur numerycznych $\{Q^m\}_{m=1,2,\cdots}$

$$Q^m(f) = \sum_{j=0}^{m} A_j^m f(x_j^m). \tag{4.3}$$

Zajmiemy się najpierw sprawą *zbieżności* ciągu kwadratur numerycznych (4.3) do funkcjonału I. Dokładniej, odpowiemy na pytanie

Przy jakich założeniach

$$Q^m(f) \to I(f) \quad gdy \quad m \to \infty$$

dla dowolnej funkcji ciągłej $f : [a, b] \to \mathbf{R}$.

Będzie nam potrzebne następujące

Twierdzenie 4.1 (Helly) *Niech będą dane*

- *Funkcjonał liniowy i ograniczony* $F : C([a, b]) \to \mathbf{R}$*,*
- *Ciąg funkcjonałów liniowych* $\{F_n\}_{n=0,1,\cdots}$ $F_n : C([a, b]) \to \mathbf{R}$ *dla których istnieje stała* $K > 0$*, taka że dla każdego* n*,* $\|F_n\| \leq K$ *(ciąg wspólnie ograniczonych funkcjonałów),*
- *Zbiór* G*, gęsty*[11] *w przestrzeni* $C([a, b])$*.*

Jeśli dla każdego $g \in G$*, zachodzi* $|F_n(g) - F(g)| \to 0$ *gdy* $n \to \infty$*, to dla każdego* $f \in C([a, b])$

$$|F_n(f) - F(f)| \to 0, \quad gdy \;\; n \to \infty.$$

Dowód. Bez zmniejszenia ogólności możemy założyć, że $\|F\| \leq K$. Niech $g \in G$ będzie dowolnym elementem zbioru gęstego G. Mamy dla dowolnego $f \in C([a, b])$

$$\leq |F(f) - F_n(f)| = |F(f) - F(g) + F(g) - F_n(g) + F_n(g) - F_n(f)| \leq$$

$$\leq |F(f) - F(g)| + |F(g) - F_n(g)| + |F_n(g) - F_n(f)|.$$

Z przyjętych założeń

$$|F(f) - F(g)| \leq K\|f - g\|_{\infty,[a,b]},$$

$$|F_n(g) - F_n(f)| \leq K\|f - g\|_{\infty,[a,b]}.$$

Przyjmijmy, ze względu na gęstość zbioru G, że element g został tak dobrany do f, że $\|f - g\|_{\infty,[a,b]} \leq \frac{\epsilon}{2K}$, gdzie ϵ jest dowolną liczbą dodatnią. Ze względu na założenie o zbieżności na zbiorze G, możemy znaleźć takie n_0, że dla $n > n_0$ $\|F_n(g) - F(g)\|_{\infty,[a,b]} \leq \frac{\epsilon}{2}$. Ostatecznie widzimy, że dla dowolnego $f \in C([a, b])$ i dla dowolnego ϵ dodatniego istnieje takie n_0, że dla każdego $n > n_0$

$$\|F(f) - F_n(f)\|_{\infty,[a,b]} \leq \epsilon.\Box$$

[11] Zbiór gęsty w przestrzeni metrycznej X, to taki zbiór, którego domknięcie jest równe X.

Możemy interpretować teraz jako F nasz funkcjonał $I(f) = \int_a^b \rho(x)f(x)dx$, jako F_n - kwadratury numeryczne $Q_n(f) = \sum_{j=0}^{n} A_j^n f(x_j^n)$, jako zbiór gęsty G - zbiór wszystkich wielomianów jednej zmiennej.[12]

Możemy teraz sformułować

Twierdzenie 4.2 *Niech dla całki* (4.1) $I(f) = \int_a^b \rho(x)f(x)dx$ *będzie dany ciąg kwadratur numerycznych* (4.3) $Q_n(f) = \sum_{j=0}^{n} A_j^n f(x_j)$ *przyczym zakładamy, że*

1. $A_j^n > 0 \quad j = 0, 1, 2, \cdots, n \quad n = 1, 2, \cdots,$
2. *kwadratura numeryczna* $Q_n, \quad n = 0, 1, 2 \cdots,$ *jest rzędu conajmniej 1,*
3. $Q_n(w) \to I(w), \quad n \to \infty$ *dla dowolnego wielomianu* w*.*

Wtedy dla każdego $f \in C([a, b])$ $Q_n(f) \to I(f)$ *gdy* $n \to \infty$. *(Mówimy krótko, że* **kwadratura** Q_n **jest zbieżna** *dla każdej funkcji z* $C([a, b])$*).*

Wniosek. *Założenie 1. oraz 2. Twierdzenia 4.2 jest spełnione, jeśli kwadratura numeryczna* Q_n *jest rzędu conajmniej* n*.*

Wobec tego, jeśli kwadratura numeryczna Q_n *jest rzędu przynajmniej* n*, oraz ma współczynniki dodatnie, to jest zbieżna dla każdej funkcji ciągłej z przestrzeni* $C([a, b])$*.*

Dowód Wniosku. Niech kwadratura numeryczna Q_n będzie rzędu n. Jest oczywiste, że spełnione jest założenie 1. Twierdzenia 4.2. Ponadto, z definicji rzedu kwadratury, $Q_n(w_{n-1}) = I(w_{n-1})$ dla dowolnego wielomianu w_{n-1} stopnia co najwyżej $n - 1$. Stąd wynika, że $Q_n(w) \to I(w)$ dla $n \to \infty$, gdyż dla n większych od stopnia wielomianu w, $Q_n(w) = I(w)$. $\Box$

Dowód Twierdzenia 4.2. Jeśli kwadratura numeryczna Q_n jest przynajmniej rzędu 1, to przyjmjąc $f(x) = 1$

$$Q_n(1) = I(1) \ \text{ dla każdego } \ n.$$

[12]Zgodnie z **Twierdzeniem Weierstrassa** zbiór wszystkich wielomianów jest gęsty w przestrzeni $C([a, b])$ ze względu na normę $\| \cdot \|_{\infty,[a,b]}$.

To znaczy, że dla każdego n

$$\|Q_n\| = \sum_{j=0}^{n} |A_j^n| = Q_n(1) = I(1) = \int_a^b \rho(x)dx = K,$$

gdyż współczynniki A_j^n są dodatnie. Wobec przyjętego założenia o zbieżności kwadratur dla wszystkich wielomianów, widzimy, że spełnione są wszystkie założenia **Twierdzenia Helly**. Stąd zbieżność kwadratur

$$Q_n(f) \to I(f) \ \text{ gdy } \ n \to \infty$$

dla każdego $f \in C([a, b])$. $\square$

Przykładem kwadratur numerycznych rozważanego typu są *kwadratury interpolacyjne*, to znaczy powstające w ten sposób, że zamiast całkować funkcję f, całkujemy jej wielomian interpolacyjny. Najprostsze są *kwadratury* **Newtona- Cotes'a**.

Na przedziałę $[a, b]$ założymy siatkę jednakowo odległych węzłów:

$$a = x_0^n < x_1^n < x_2 \cdots < x_n^n = b \tag{4.4}$$

gdzie $x_j = a + jh$, $h = \frac{b-a}{n}$. Niech P_n będzie *wielomianem interpolacyjnym Lagrange'a* funkcji $f \in C([a, b])$, opartym na węzłach (4.4). Wyraźmy wielomian P_n przy pomocy bazy Lagrange'a

$$P_n(x) = \sum_{j=0}^{n} l_j(x) f(x_j),$$

gdzie

$$l_j(x) = \frac{(x - x_0^n)(x - x_1^n) \cdots (x - x_{j-1}^n)(x - x_{j+1}^n) \cdots (x - x_n^n)}{(x_j^n - x_0^n)(x_j^n - x_1^n) \cdots (x_j^n - x_{j-1}^n)(x_j^n - x_{j+1}^n) \cdots (x_j^n - x_n^n)}.$$

Wykorzystajmy jeszcze fakt, że węzły są równoodległe i wprowadźmy nową zmienną niezależną s określoną przez związek

$$s(x) = \frac{x - a}{h}.$$

Zatem $s(x_j^n) = \frac{a+jh-a}{h} = j$, dla $j = 0, 1, \cdots, n$, oraz $\frac{dx}{ds} = h = \frac{b-a}{n}$. Stąd

$$Q_n(f) = \int_a^b \rho(x) P_n(x) dx = \sum_{j=0}^{n} \int_a^b \rho(x) l_j(x) dx \cdot f(x_j^n) = \sum_{j=0}^{n} A_j^n f(x_j^n).$$

Wprowadzona nowa zmienna pozwoli nam wyrazić współczynniki kwadratury A_j^n w sposób niezależny od przedziału i węzłów, uzależniając je tylko od liczby węzłów. Zapiszemy wzory dla tych współczynników w przypadku najczęściej występującym w zastosowaniach, to jest dla $\rho(x) = 1$.

$$A_j^n = \int_a^b l_j(x) dx = \frac{b-a}{n} \int_0^n l_j(s) ds.$$

Łatwo policzyć że

$$l_j(s) = \frac{s(s-1)(s-2)\cdots(s-j+1)(s-j-1)\cdots(s-n)}{j(j-1)(j-2)\cdots 2\cdot 1\cdot(-1)(-2)\cdots(j-n)}.$$

Współczynnik A_j^n wyraża się przy pomocy całki z wielomianu, którą policzyć można dokładnie. Ponadto jeśli oznaczymy

$$B_j^n = \int_0^n l_j(s) ds,$$

to

$$Q_n(f) = \frac{b-a}{n} \sum_{j=0}^{n} B_j^n f(x_j^n)$$

to znaczy naszą kwadraturę numeryczną zapisaliśmy w taki sposób że wszelkie informacje o przedziale całkowania i węzłach są zawarte we współczynniku $\frac{b-a}{n}$ oraz w stablicowanych wartościach funkcji f - $f(x_j^n)$. Komplet współczynników B_j^n $j = 0, 1, \cdots, n$ zależy zaś jedynie od n - jest zatem *uniwersalny* i może być umieszczony w tablicach.

Warto zadać sobie istotne pytanie: czy współczynniki B_j^n są dodatnie, **gdyż od tego zależy zbieżność kwadratury Newtona - Cotesa.** Okazuje się że **jest tak jedynie dla** $n \leq 7$ **oraz dla** $n = 9$**.** Można pokazać, że kwadratury te *są rzędu n+1 dla n nieparzystych, zaś rzędu n+2 dla n parzystych.*

Jak więc jest z użytecznością tych formuł kwadraturowych? Oczywiście nie jest dobrym wyjsciem stosowanie formuł niezbieżnych. Można jednak znaleźć sposób użycia formuł Newtona - Cotesa w sposób taki, aby uzyskać kwadratury zbieżne. Trudności ze zbieżnością kwadratur interpolacyjnych wynikają z podobnych trudności związanych ze zbieżnością globalnych wielomianów interpolacyjnych. I sposób radzenia sobie z tym problemem jest podobny jak w przypadku interpolacji. Stosujemy **kwadratury złożone**. Dokonajmy podziału odcnka $[a, b]$ na przykład na n równych części o długości $h = \frac{b-a}{n}$, zaś na każdej z tych cześci stosujmy kwadraturę Newtona - Cotesa o współczynnikach dodatnich, opartą na ustalonej liczbie węzłów. Jeśli na każdym podprzedziale stosowaliśmy formułe Newtona - Cotesa opartą o k węzłow, to otrzymamy kwadraturę numeryczną dla przedziału $[a, b]$ opartą na $m = n \cdot k$ węzłach, i dodatnich współczynnikach. Będzie ona takiego rzędu jak zastosowana lokalnie kwadratura Newtona - Cotesa.

Zadanie 4.2 Udowodnij zbieżnośc kwadratury złożonej otrzymanej w sposób opisany powyżej. Zastosuj wzór na oszacowanie błędu interpolacji Lagrange'a w przypadku, gdy funkcją interpolowaną jest wielomian (wysokiego stopnia). Wykorzystaj **Twierdzenie 4.2**.

Kwadratury Gauß'a

Dla całki
$$I(f) = \int_a^b \rho(x) f(x) dx$$
będziemy poszukiwali teraz kwadratury numerycznej z ustaloną funkcją wagową ρ i z ustaloną liczbą $n + 1$ węzłów, opartej na globalnej interpolacji Lagrange'a, mającej **maksymalny możliwy rząd.**

Niech więc f będzie wielomianem stopnia $m \geq n$. Poszukujemy takiej kwadratury numerycznej
$$Q_n(f) = \sum_{j=0}^{n} A_j f(x_j),$$
aby
$$Q_n(f) = I(f).$$
dla dowolnego wielomianu f stopnia m, **dla możliwie dużego** m**.** Niech P_n będzie wielomianem interpolacyjnym Lagrange'a dla f o węzłach
$$a \leq x_0 < x_1 < x_2 < \cdots < x_n \leq b.$$

Wtedy

$$P_n(x) = \sum_{j=0}^{n} l_j(x) f(x_j),$$

gdzie $l_j(x)$ są funkcjami bazowymi Lagrange'a. Mamy więc $f(x_j) - P_n(x_j) = 0$ dla $j = 0, 1, 2, \cdots, n$. Oznacza to, że wielomian stopnia $\leq m$, $f - P_n$ musi **dzielić się przez wielomian stopnia n+1**

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n).$$

Stąd wynika, że

$$f(x) - P_n(x) = \omega(x) g(x),$$

gdzie g jest wielomianem stopnia $l \leq m - n - 1$. Chcielibyśmy, aby

$$\int_a^b \rho(x)[f(x) - P_n(x)]dx = \int_a^b \rho(x)\omega(x)g(x)dx = 0$$

dla *możliwie dużego* m. Warunek

$$\int_a^b \rho(x)\omega(x)g(x)dx = 0$$

będzie spełniony dla każdego wielomianu g stopnia $\leq n$, jeśli tylko ω jest **wielomianem ortogonalnym na przedziale** $[a, b]$ **z wagą** ρ**.** Ponieważ

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$$

oznacza to, że **węzły** $x_0, x_1, \cdots, x_n$ **są pierwiastkami** $n+1$**-go wielomianu ortogonalnego na przedziale [a,b], z wagą** ρ**.** Zatem, jeśli dobierzemy węzły właśnie tak, to

$$\int_a^b \rho(x) f(x)dx = \int_a^b \rho(x) P_n(x)dx,$$

dla $l = m - n - 1 < n + 1$, czyli dla $m < 2n + 2$. **Oznacza to, że nasza formuła kwadraturowa jest rzędu** $2n + 2$**.** W ten sposób udowodniliśmy

Twierdzenie 4.3. *Jeśli węzłami kwadratury numerycznej interpolacyjnej z wagą* ρ *w przedziale* $[a, b]$ *są* **pierwiastki wielomianu stopnia** $n+1$**, ortogonalnego na tym przedziale z tą właśnie wagą**, *to otrzymana kwadratura* **jest rzędu** $2n+2$. *Tak zbudowane kwadratury noszą nazwę* **kwadratur Gauß'a.** □

Widzimy więc, że kwadratury Gauß'a mają rząd znacznie wyższy niż inne kwadratury interpolacyjne oparte na tej samej liczbie węzłów. W zależności od przedziału i funkcji wagi, w grę wchodzą różne wielomiany ortogonalne i różne, związane z nimi kwadratury. Istnieją tablice węzłów i współczynników kwadratur Gauß'a. [13] Rząd kwadratury odgrywa istotną rolę z punktu widzenia jakości aproksymacji rozważanej całki. Ale, kwadratury Gauß'a mają także i inne pozytywne cechy.

Przyjrzyjmy się bliżej wzorom dla omawianych kwadratur numerycznych.

$$Q_n(f) = \int_a^b \rho(x) P_n(x) dx = \sum_{j=0}^{n} \int_a^b \rho(x) l_j(x) dx f(x_j) = \sum_{j=0}^{n} A_j f(x_j),$$

gdzie

$$A_j = \int_a^b \rho(x) l_j(x) dx \quad j = 0, 1, \cdots, n$$

i $l_j(x) \;\; j = 0, 1, \cdots, n$ są bazowymi wielomianami interpolacji Lagrange'a dla pierwiastków $n + 1$-go wielomianu ortogonalnego rozważanego ciągu wielomianów ortogonalnych, które przyjmujemy jako węzły naszej kwadratury.

Zauważmy odrazu, że $l_k(x)^2, \quad k = 0, 1, \cdots, n$ jest wielomianem stopnia $2n < 2n+1$, więc **jest całkowalny dokładnie** przy pomocy naszej formuły kwadraturowej. Zauważmy jeszcze, że

$$l_k(x_l) = \delta_{k,l} = l_k(x_l)^2 \quad k, l = 0, 1, \cdots, n.$$

Zatem

$$\int_a^b \rho(x) l_k(x)^2 dx = \sum_{j=0}^{n} A_j l_k(x)^2$$

i w powyższej sumie wystąpi tylko jeden składnik niezerowy - dla $j = k$. Ostatecznie, ponieważ $l_k(x)^2 \geq 0$ i $\rho(x) \geq 0$

$$0 < \int_a^b \rho(x) l_k(x)^2 dx = A_k,$$

a więc **współczynniki kwadratury Gauß'a są zawsze dodatnie.** Na mocy **Wniosku z Twierdzenia 4.2** stwierdzamy, że **kwadratury Gauß'a**

[13] Godne polecenia są tablice kwadratury Gauß'a - Legendra i kwadratury Gauß'a - Laguerre'a wydane przez National Bureau of Standards z datą 10.11.1954.

są zbieżne dla dowolnej funkcji ciągłej. Ponadto w kwadraturach złożonych mogą być użyte formuły Gauß'owskie o dowolnej liczbie węzłów.

Zadanie 4.3 Kwadratura Romberga, to kwadratura złożona zbudowana z kwadratur Newtona - Cotes'a opartych na dwóch węzłach (kwadratura trapezów). Napisz wzory dla kwadratury Romberga i oszacuj jej bład.

Zadanie 4.4 Wykorzystując tablice dla węzłów i współczynników kwadratur Gauß'a - Legendre'a zbuduj podprogram dla kwadratury złożonej opartej o wzory dla 16 węzłów. Jako argumenty ("parametry") podprogramu powinny wystąpić

- liczba podprzedziałów N,
- krańce przedziału całkowania
- nazwa funkcji podcałkowej f - funkcja f powinna być zadana innym podprogramem.

Przeprowadź testy numeryczne.

Rozdział 5
ROZWIĄZYWANIE NUMERYCZNE RÓWNAŃ NIELINIOWYCH

Interesujące nas zagadnienie postawimy w sposób dość ogólny. Niech $(X, \|\cdot\|_X)$ i $(Y, \|\cdot\|_Y)$ będą dwiema przestrzeniami liniowymi, unormowanymi (najlepiej przestrzeniami Banacha), i niech

$$F : X_0 \to Y, \quad X_0 \subset X$$

będzie zadaną funkcją. Poszukujemy takiego elementu $\alpha \in X_0$, że

$$F(\alpha) = 0, \tag{5.1}$$

gdzie 0 jest *elementem zerowym przestrzeni liniowej* Y. Oczywiście równanie (5.1) może nie mieć wogóle rozwiązania, może mieć tylko jedno rozwiązanie i może mieć ich wiele. Nas będzie najczęściej interesował przypadek *lokalnej jednoznaczności* rozwiązania równania (5.1).

Rozwiązanie α równania (5.1) jest *lokalnie jednoznaczne* jeśli istnieje takie otoczenie punktu α w X_0, że w tym otoczeniu α jest jedynym rozwiązaniem równania (5.1).

Wiele sposobów rozwiązywania numerycznego równań typu (5.1) polega na *lokalnej linearyzacji* rozwiązywanego zadania, oraz na *iteracyjnym* rozwiązywaniu zlinearyzowanych zadań.

Zadanie (5.1) może określać, na przykład, nieliniowe zagadnienie postawione dla równań różniczkowych lub całkowych. Wtedy zazwyczaj przestrzenie X i Y mają wymiar nieskończony.

Najczęściej wynikiem zastosowania metody numerycznej do zadania (5.1) jest wygenerowany ciąg elementów przestrzeni X

$$x_0, x_1, x_2, \cdots, \tag{5.2}$$

który zbiega do poszukiwanego rozwiązania α.

Przypuśćmy, że rozważamy metodę generującą ciąg (5.2) dla równania (5.1). Wyrażenie

$$e_k = \alpha - x_k$$

nazywamy błędem na k-tym kroku naszej metody.

Mówimy, że rozważana metoda jest rzędu γ, jeśli istnieje stała dodatnia C taka, że dla każdego k

$$\|e_{k+1}\|_X \leq C(\|e_k\|_X)^\gamma, \tag{5.3}$$

zaś powyższy warunek nie zachodzi dla żadnego $\gamma_1 > \gamma$.

Zauważmy, że szybkość zbieżności ciągu (5.2) (szybkość malenia normy błędu) zależy od rzędu metody. Zachowanie się procesu iteracyjnego zależy w sposób bardzo istotny od tego, jaki jest jego rząd. Mówi o tym

Twierdzenie 5.1 *Przypuśćmy, że proces iteracyjny rzędu γ mający aproksymować element α produkuje ciąg*

$$x_0, x_1, x_2, \cdots$$

Wtedy

1. *jeśli $\gamma < 1$, to proces może nie być zbieżny,*
2. *jeśli $\gamma = 1$, to błąd na k-tym kroku iteracji spełnia warunek*

   $$\|e_k\| \leq C^k\|e_0\|, \tag{5.4}$$

   a zatem jest **zbieżny geometrycznie** *dla dowolnego punktu startowego x_0, gdy stała C we wzorze (5.3) i (5.4) spełnia nierówność $0 \leq C < 1$,*
3. *jeśli $\gamma > 1$, to*

   $$\|e_k\| \leq \frac{1}{C_1}(C_1\|e_0\|)^{\gamma^k}, \tag{5.5}$$

   gdzie $C_1 = C^{\frac{1}{\gamma-1}}$, a więc proces zbiega z rzędem γ jesli tylko

   $$\|e_0\| < \frac{1}{C_1}. \tag{5.6}$$

   Oznacza to, że proces zbiega nie zależnie od wartości stałej C, ale wtedy, gdy punkt startowy x_0 został wybrany **dostatecznie blizko** *poszukiwanego punktu α.*

Dowód. Ponieważ $\|e_{k+1}\| \le C\|e_k\|^\gamma$ to, stąd otrzymujemy

$$\|e_k\| \le \begin{cases} C^{1+\gamma+\cdots+\gamma^{k-1}}\|e_0\|^{\gamma^k}, & k=1,2,\cdots, \quad \gamma \ne 1 \\ C^k\|e_0\|, & k=1,2,\cdots, \quad \gamma = 1 \end{cases}.$$

Niech $\gamma \ne 1$. Zauważmy, że

$$1+\gamma+\gamma^2+\cdots+\gamma^{k-1} = \frac{\gamma^k-1}{\gamma-1} = \frac{1}{\gamma-1}\gamma^k - \frac{1}{\gamma-1}$$

i stąd

$$C^{1+\gamma+\gamma^2+\cdots+\gamma^{k-1}} = \frac{C_1^{\gamma^k}}{C_1}$$

gdzie $C_1 = C^{\frac{1}{\gamma-1}}$. Zatem

$$\|e_k\| \le \frac{1}{C_1}(C_1\|e_0\|)^{\gamma^k}. \tag{5.7}$$

1. Niech $\gamma < 1$. Rozważmy proces iteracyjny spełniający warunek

   $$\|e_{k+1}\| = C\|e_k\|^{\gamma^k}.$$

   Wtedy

   $$\|e_k\| = \frac{1}{C_1}(C_1\|e_0\|)^{\gamma^k},$$

   i ponieważ $0 \le \gamma < 1$, to $\gamma^k \to 0$ gdy $k \to \infty$. Stąd $\|e_k\| \to \frac{1}{C_1}$, a więc rozważany proces nie jest zbieżny.

2. Niech $\gamma = 1$. Wtedy

   $$\|e_k\| \le C^k\|e_0\|, \tag{5.8}$$

   i proces zbiega **geometrycznie, dla każdego** x_0, pod warunkiem, że $0 \le C < 1$.

3. Niech $\gamma > 1$. Wtedy

   $$\|e_k\| \le \frac{1}{C_1}(C_1\|e_0\|)^{\gamma^k},$$

   i $\gamma^k \to \infty$, gdy $k \to \infty$. Zatem proces zbiega, gdy

   $$\|e_o\| < \frac{1}{C_1}. \tag{5.8}$$

   □

Zajmiemy się najpierw najprostszym przypadkiem *jednego równania skalarnego.* Wtedy $X = Y = \mathbf{R}$ i w obu przestrzeniach normą jest moduł. Rozważmy teraz równanie

$$f(x) = 0, \tag{5.9}$$

gdzie $f : \mathbf{R} \to \mathbf{R}$. Jeśli f jest funkcją ciągłą, to można niekiedy poszukiwać przybliżonego rozwiązania równania (5.9) *metodą bisekcji.* Niech $a < b$ i $f(a) < 0$, zaś $f(b) > 0$. Ponieważ f jest funkcją ciągłą, to przedział $[a, b]$ zawiera napewno przynajmniej jedno rozwiązanie α równania (5.9). Połóżmy $x_0 = \frac{a+b}{2}$. Możliwe są trzy przypadki

1. $f(x_0) = 0$,
2. $f(x_0) > 0$,
3. $f(x_0) < 0$.

Przyjmijmy $a_0 = a$ i $b_0 = b$.

Jeśli zachodzi 1., to $\alpha = x_0$, proces jest zakończony.

Jeśli zachodzi 2., to kładziemy $a_1 = a_0$ i $b_1 = x_0$.

Jeśli zachodzi 3., to kładziemy $a_1 = x_0$ i $b_1 = b$.

Teraz wyliczamy $x_1 = \frac{a_0+b_0}{2}$. Postępując w ten sposób, albo w pewnym momencie znajdziemy jakieś rozwiązanie α, albo wytworzymy ciąg

$$x_0, x_1, \cdots, x_n, \cdots$$

o tej własności, że $x_k = \frac{a_{k-1}+b_{k-1}}{2}$, zaś $|b_k - a_k| = \frac{|b_0-a_0|}{2^k}$ i każdy z przedziałów $[a_k, b_k]$ zawiera pierwiastek równania (5.9).

Inna metoda polega na przedstawieniu równania (5.1) w *równoważnej postaci*

$$x = \Phi(x). \tag{5.10}$$

Jest to *zadanie znalezienia punktu stałego (fixpunktu)* funkcji $\Phi : X \to X$. Równanie (5.1) można sprowadzać do postaci (5.10) różnymi sposobami, między innymi takimi które zostały omówione w Rozdziale 3 przy okazji rozważania metod iteracyjnych dla układów równań algebraicznych liniowych. Jeśli funkcja F określająca równanie (5.1) działa z przestrzeni X w przestrzeń Y

i operator liniowy $G : Y \to X$ jest odwracalny na Y, to można na przykład przyjąć $\Phi(x) = x + G(F(x))$. Funkcja Φ określa wtedy rodzaj *nieliniowego procesu iteracyjnego Richardsona*

$$x_{k+1} = \Phi(x_k), \quad x_0 \;\; - \text{ zadane,} \tag{5.11}$$

gdyż $F(x)$ jest *reziduum* równania (5.1) w punkcie x_k. O zbieżności procesu iteracyjnego

$$x_{k+1} = \Phi(x_k)$$

mówi dobrze znane **Twierdzenie Banacha o punkcie stałym.** W przypadku, gdy $\Phi : X \to X$, gdzie X jest przestrzenią Banacha, to twierdzenie można tak sformułować

Twierdzenie Banacha *Jeśli* $\Phi : X \to X$ *przyczym istnieje stała* $0 \le L < 1$ *taka, że dla dowolnych* $x, y \in X$

$$\|\Phi(x) - \Phi(y)\| \le L\|x - y\|,$$

to istnieje **jedyny w** X *punkt stały* α *funkcji* Φ

$$\alpha = \Phi(\alpha).$$

Ponadto, dla dowolnego x_0*, ciąg* $x_0, x_1, x_2, \cdots$*, gdzie*

$$x_{k+1} = \Phi(x_k), \quad k = 0, 1, 2, \cdots \tag{5.12}$$

zbiega do α*:*

$$\|x_k - \alpha\| \to 0 \;\; \textit{gdy} \;\; k \to \infty.$$

Zadanie 5.1 Przypomnij dowód **Twierdzenia Banacha.** Zwróc uwagę na to, że dowodzi się tu

- istnienie punktu stałago,
- zbieżność ciągu (5.12).

Zastanów się jaką rolę odgrywa założenie o **zupełności** prestrzeni X. (Przestrzeń Banacha jest zupełna!).

Proces iteracyjny określony wzorem (5.12) nazywa się *iteracją prostą*. Zbadajmy jego rząd.

Mamy:

$$\|e_{k+1}\| = \|\alpha - x_{k+1}\| = \|\Phi(\alpha) - \Phi(x_k)\| \leq L\|e_k\|,$$

Oznacza to, ze iteracja prosta jest rzedu 1, a więc przy przyjętych założeniach jest ona zbieżna geometrycznie dla dowolnego punktu startowego x_0.

Aby skorzystać z Twierdzenia Banacha, należy równanie (5.1) przekształcić do postaci (5.10). Czasem równanie jest już w tej postaci. Na przykład tak jest dla równania

$$x - \frac{\sin x}{2} = 0.$$

Jeśli $f : X \to X$, to dla rozwiązania numerycznego równania

$$f(x) = 0$$

można probować zastosować *itreację Richardsona* z liczbowym współczynnikiem relaksacji κ

$$x_{k+1} = x_k - \kappa f(x_k) \quad k = 0, 1, 2, \cdots$$

Zadanie 5.2 Zakładając, że $f \in C^1$ znajdź warunek dostateczny jaki powinien spełniać *współczynnik relaksacji* κ, aby iteracja Richardsona była zbieżna.

Sensowne wydaje sie, jeśli to możliwe, łączenie dwóch procesów iteracyjnych

- najpierw stosujemy proces rzędu 1 aby zbliżyć się do rozwiązania równania
- następnie, gdy już zbliżyliśmy się dostatecznie dobrze stosujemy proces rzędu wyższego niż 1, który zbiega szybciej.

Dla tego też warto zainteresować sie procesami wyższego rzędu. Takim procesem jest **Metoda Newtona**. Najpierw określimy tę metodę dla równania skalarnego

$$f(x) = 0,$$

gdy $f \in C^1$:

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \quad x_0 \text{ punkt startowy.} \tag{5.13}$$

Zadanie 5.3 Zinterpretuj geometrycznie wzór (5.13). Udowodnij, że punkt x_{k+1} leży na przecięciu osi x ze styczną do wykresu funkcji f wychodzącą z punktu $(x_k, f(x_k))$.

Zauważmy, że wzór (5.13) można interpretować w następujący sposób: Rozwjamy f przy pomocy wzoru Taylora, biorąc tylko dwa wyrazy:

$$f(x) = f(x_k) + f'(x_k)(x - x_k) + r_k$$

Odrzucamy resztę r_k i rozwiązujemy **równanie liniowe**

$$f(x_k) + f'(x_k)(x - x_k) = 0, \tag{5.14}$$

którego rozwiązanie to właśnie x_{k+1} ze wzoru (5.13). Jest to zatem *linearyzacja* równania oryginalnego, dokonywana na każdym kroku iteracji. Jeśli interesuje nas układ równań, lub ogólniej *równanie w przestrzeni Banacha*

$$F(x) = 0, \tag{5.1}$$

to metoda Newtona jest określona poprzez równania liniowe

$$F'(x_k)(x_{k+1} - x_k) + F(x_k) = 0, \tag{5.15}$$

przy założeniu, że *pochodna Frécheta* funkcji F, F' **istnieje i jest odwracalna** w obszarze który nas interesuje. [14] Rozważmy bardzo prosty przykład układu dwóch równań

$$G_1(x_1, x_2) = 0,$$

[14]Pochodną Frécheta $F'(x)$ w punkcie x funkcji $F : X \to Y$ działającej w przestrzeniach Banacha X, Y określa się jako **część liniową względem** $h \in X$ **przyrostu** $F(x+h) - F(x) = F'(x)h + r$, gdzie $r = o(\|h\|)$. Pochodna Frécheta, jeśli istnieje, jest **operatorem liniowym:** $F'(x) : X \to Y$.

$$G_2(x_1, x_2) = 0. \tag{5.16}$$

Zakładamy, że obie funkcje mają ciągłe pochodne cząstkowe. Naszą przestrzenią Banacha jest teraz $X = \mathbf{R^2}$ i

$$F(x) = \begin{bmatrix} G_1(x_1, x_2) \\ G_2(x_1, x_2) \end{bmatrix},$$

zaś $x = (x_1, x_2)$. Wtedy $F'(x)$ jest **macierzą jacobianu** (a więc jest to **operator liniowy działający w** X).

$$F'(x) = \begin{bmatrix} \frac{\partial G_1}{\partial x_1}(x_1, x_2), & \frac{\partial G_1}{\partial x_2}(x_1, x_2) \\ \frac{\partial G_2}{\partial x_1}(x_1, x_2), & \frac{\partial G_2}{\partial x_2}(x_1, x_2) \end{bmatrix}.$$

Otrzymujemy w ten sposób układ dwóch równań algebraicznych liniowych do rozwiązania na każdym kroku iteracji

$$\frac{\partial G_1}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1} - x_1^k) + \frac{\partial G_1}{\partial x_2}(x_1^k, x_2^k)(x_2^{k+1} - x_2^k) + G_1(x_1^k, x_2^k) = 0,$$
$$\frac{\partial G_2}{\partial x_1}(x_1^k, x_2^k)(x_1^{k+1} - x_1^k) + \frac{\partial G_2}{\partial x_2}(x_1^k, x_2^k)(x_2^{k+1} - x_2^k) + G_2(x_1^k, x_2^k) = 0,$$

z którego wyznaczamy $x_{k+1} = (x_1^{k+1}, x_2^{k+1})$. Warunkiem wykonalności jest odwracalność macierzy jakobianu.

Zbadamy teraz rząd iteracji Newtona, w przypadku równania skalarnego

$$f(x) = 0.$$

Twierdzenie 5.2 *Jeśli* $f \in C^2$, $f(\alpha) = 0$, $f'(\alpha) \neq 0$, *to iteracja Newtona* (5.13) *jest rzędu 2.*

Dowód Ponieważ $f(\alpha) = 0$ i $f'(\alpha) \neq 0$, mamy

$$\alpha - x_{k+1} = \alpha - x_k - \frac{f(\alpha) - f(x_k)}{f'(x_k)}.$$

Rozwijając przy pomocy wzoru Taylora w punkcie α otrzymamy

$$f(x_k) - f(\alpha) = f'(\alpha)(x_k - \alpha) + \frac{f''(d)}{2}(x_k - \alpha)^2,$$

$$f'(x_k) = f'(\alpha) + f''(d')(x_k - \alpha),$$

gdzie d i d' leżą w przedziale $(\min\{\alpha, x_k\}, \max\{\alpha, x_k\})$. Stąd

$$\alpha - x_{k+1} = \alpha - x_k + \frac{-f'(\alpha)(\alpha - x_k) + \frac{f''(d)}{2}(\alpha - x_k)^2}{f'(\alpha)[1 - \frac{f''(d')}{f'(\alpha)}(\alpha - x_k)]} =$$

$$= \alpha - x_k +$$

$$+ \frac{-f'(\alpha)[1 - \frac{f''(d')}{f'(\alpha)}(\alpha - x_k)](\alpha - x_k) + [\frac{f''(d)}{2} - f'(\alpha)\frac{f''(d')}{f'(\alpha)}](\alpha - x_k)^2}{f'(\alpha)[1 - \frac{f''(d')}{f'(\alpha)}(\alpha - x_k)]} =$$

$$= \frac{[\frac{f''(d)}{2} - f'(\alpha)\frac{f''(d')}{f'(\alpha)}](\alpha - x_k)^2}{f'(\alpha)[1 - \frac{f''(d')}{f'(\alpha)}(\alpha - x_k)]} = O(\alpha - x_k)^2$$

co oznacza, że iteracja jest rzędu 2. □

Twierdzenie 5.1 i Twierdzenie 5.2 pozwalają stwierdzić, że przy przyjętych założeniach o funkcji f metoda Newtona jest zbieżna kwadratowo, jeśli tylko punkt startowy x_0 został wybrany dostatecznie blizko rozwiązania α. Podobne twierdzenia można udowodnić dla równań w dowolnych przestrzenach Banacha [15]

Zadanie 5.4 Niech $X = C([a, b])$ (norma "sup"), oraz niech

$$F(x)(t) = x(t) + \int_a^b f(x(s))ds, \quad t \in [a, b], \quad f \in C^2([a, b]), \quad x \in X.$$

Wypisz wzory procesu iteracyjnego Newtona dla równania

$$F(x) = 0.$$

Kiedy ten proces będzie rzędu 2? Zastanów się co nam daje zastosowanie procesu Newtona do rozważanego zadania.

[15] Patrz na przykład N.S. Bahvalov "Čislennye Metody", tom I, Nauka, Moskva 1973 str. 411-416

Zajmiemy się teraz znajdowaniem pierwiastków wielomianów jednej zmiennej.

Zadanie 5.5 Niech P_n będzie wielomianem stopnia n. Do numerycznego rozwiązania równania

$$P_n(x) = 0$$

zastosuj metodę Newtona, wykorzystując *schemat Hornera*, dwukrotnie na każdym kroku iteracji.

Jeśli wielomian P_n, stopnia n, ma *współczynniki rzeczywiste* i poszukujemy zer zespolonych tego wielomianu, to warto zastosować wygodniejszy **algorytm Bairstowa**. Taki wielomian może mieć pierwiastki rzeczywiste oraz pary sprzężone pierwiastków zespolonych. Poszukiwanie zer zespolonych przy bezpośrednim użyciu metody Newtona (patrz zadanie!) musi wykorzystywać *arytmetykę zespoloną*. Metoda Bairstowa działa wyłącznie w dziedzinie rzeczywistej. Będziemy poszukiwali **dzielników kwadratowych** wielomianu P_n, postaci

$$x^2 + px + q, \tag{5.16}$$

gdzie p i q są liczbami rzezywistymi. Dzieląc P_n przez $x^2 + px + q$ otrzymamy

$$P_n(x) = Q_{n-2}(x)(x^2 + px + q) + Rx + S, \tag{5.17}$$

gdzie Q_{n-2} jest ilorazem, zaś $Rx + S$ jest **resztą** stopnia nie wyższego niz 1.

Zadanie 5.6 Napisz algorytm, podobny do schemau Hornera, który wyznacza iloraz Q_{n-2} oraz resztę $Rx + S$ z dzielenia wielomianu $P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0$ przez wielomian kwadratowy $x^2 + px + q$.

Zauważmy, że współczynniki ilorazu Q_{n-2}, oraz reszty R i S są funkcjami zmiennych p i q, zaś oczywiście, współczynniki $a_0, a_1 \cdots, a_n$ wielomianu P_n od p i q nie zależą.

Znalezienie dzielnika kwadratowego, jest zatem równoważne rozwiązaniu układu dwóch równań

$$R(p, q) = 0,$$

$$S(p, q) = 0.$$

Zastosujemy do tego **metodę Newtona.** Wartości funkcji R i S dla zadanych p i q znajdujemy z uogólnionego schematu Hornera dzielenia P_n przez $x^2 + px + q$ (Zadanie!). Potrzebne są nam jeszcze pochodne cząstkowe

$$\frac{\partial R}{\partial p}, \frac{\partial R}{\partial q}, \frac{\partial S}{\partial p}, \frac{\partial S}{\partial q}.$$

Aby skonstruować algorytm wyznaczający te pochodne, zróżniczkujmy tożsamość (5.17) względem p i q.

$$0 = \frac{\partial Q_{n-2}}{\partial p}(x^2 + px + q) + xQ_{n-2} + \frac{\partial R}{\partial p} + \frac{\partial S}{\partial p}.$$

Na ten ostatni wzór możemy spojrzeć jak na **dzielenie** wielomianu $-xQ_{n-2}$, stopnia $n-1$ przez $x^2 + px + q$:

$$-xQ_{n-2} = \frac{\partial Q_{n-2}}{\partial p}(x^2 + px + q) + \frac{\partial R}{\partial p} + \frac{\partial S}{\partial p}.$$

Podobnie, Różniczkując (5.17) względem q otrzymamy wzór na dzielenie wielomianu $-Q_{n-2}$ stopnia $n-2$ przez $x^2 + px + q$:

$$-Q_{n-2} = \frac{\partial Q_{n-2}}{\partial q}(x^2 + px + q) + \frac{\partial R}{\partial q} + \frac{\partial S}{\partial q}.$$

Wielomian Q_{n-2} otrzymujemy z pierwszego dzielenia P_n przez czynnik kwadratowy. Musimy zatem na każdym kroku iteracji wykonać trzy dzielenia:

- P_n przez $x^2 + px + q$,
- $-xQ_{n-2}$ przez $x^2 + px + q$,
- $-Q_{n-2}$ przez $x^2 + px + q$.

Kolejne reszty to

$$R \;\; i \;\; S,$$
$$\frac{\partial R}{\partial p} \;\; i \;\; \frac{\partial S}{\partial p},$$
$$\frac{\partial R}{\partial q} \;\; i \;\; \frac{\partial S}{\partial q}.$$

Otrzymane reszty określają wszystkie współczynniki algorytmu Newtona.

Deflacja. Deflacja, to operacja usuwania z wielomianu czynników odpowiadających już wynaczonym pierwiastkom. Deflacja jest potrzebna po to, by nie wyznaczać ponownie już wyznaczonych pierwiastków.

Deflacja czynnika liniowego $x - \alpha$, to poprostu dzielenie

$$P_n(x) = Q_{n-1}(x)(x - \alpha) + R,$$

gdzie $R = P_n(\alpha)$. Dzielenie to wykonujemy przy pomocy schematu Hornera

	a_n	a_{n-1}	a_{n-2}	$\cdots$	a_1	a_0
α	b_{n-1}	b_{n-2}	b_{n-3}	$\cdots$	b_0	**R**

$$P_n(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_0,$$
$$Q_{n-1}(x) = b_{n-1}x^{n-1} + b_{n-2}x^{n-2} + \cdots + b_0.$$

Jeśli $R = P_n(\alpha) = 0$, to schemat Hornera można wykonywać "w dwie strony":

- z lewej do prawej:

$$b_{n-1} = a_n,$$
$$b_{n-2} = a_{n-1} + \alpha b_{n-1},$$
$$b_{n-3} = a_{n-2} + \alpha b_{n-2},$$
$$\cdots$$
$$b_0 = a_1 + \alpha b_1,$$
$$R = a_0 + \alpha b_0,$$

- z prawej do lewej:

$$\begin{aligned} c_0 &= \frac{R - a_0}{\alpha}, \\ c_1 &= \frac{c_0 - a_1}{\alpha}, \\ c_2 &= \frac{c_1 - a_2}{\alpha}, \\ &\cdots \\ c_{n-2} &= \frac{c_{n-3} - a_{n-2}}{\alpha}, \\ c_{n-1} &= \frac{c_{n-2} - a_{n-1}}{\alpha}. \end{aligned}$$

W algorytmie "z prawej do lewej" $c_0, c_1, \cdots, c_{n-1}$ oznaczają współczynniki wielomianu Q_{n-1}. Jeśli działania są wykonywane *w arytmetyce dokładnej*, to oczywiście $c_j = b_j$ dla $j = 0, 1, 2, \cdots, n-1$. Nie jest tak, gdy działania wykonuje się w "arytmetyce komputerowej". Jeśli wszystkie etapy obliczania pierwiastków i deflacji wykonuje się w arytmetyce "fl", to jeśli dobierze się k tak, aby

$$\frac{|c_k - b_k|}{|a_k| + |c_k|} = \min_{|a_j|+|c_j|>0} \frac{|c_j - b_j|}{|a_j| + |c_j|},$$

to wielomian o współczynnikach

$$c_{n-1}, c_{n-2} \cdots, c_{k+1}, b_k, \cdots, b_0$$

daje numerycznie poprawną deflację czynnika liniowego $x - \alpha$, pod warunkiem, że pierwiastek α został obliczony algorytmem numerycznie poprawnym.

Metoda Bairstowa i deflacja. Dobre rezultaty daje metoda Bairstowa, jeśli pierwiastki wielomianu wyznaczamy zgodnie z rosnącymi modułami. Jeśli wyznaczony czynnik $x^2 + px + q$ odpowiada dwóm pierwiastkom o bardzo różnych modułach (są one zatem rzeczywiste), to na ogól dobrze jest wyznaczony tylko ten, o większym module. Trzeba zatem dokonać deflacji tego "lepszego" pierwiastka. Gdy pierwiastki mają moduły porównywalne (na przykład, gdy stanowią parę sprzężoną), można odrazu dokonać deflacji czynnika kwadratowego.

Rozdział 6
NUMERYKA W RÓWNANIACH RÓŻNICZKOWYCH

RÓWNANIA RÓŻNICZKOWE ZWYCZAJNE - TROCHĘ TEORII.

Równanie różniczkowe zwyczajne, to równanie następującej postaci

$$\frac{du(t)}{dt} = f(t, u(t)), \tag{6.1}$$

gdzie $t \in \mathbf{R}$, $u(t) \in \mathbf{R^m}$, zaś funkcja f jest ciągła ze względu na wszystkie argumenty. **Rozwiązaniem jest jest funkcja u.** Aby funkcja u mogła być rozwiązaniem równania (6.1), musi ona być klasy C^1. Taka funkcja klasy C^1, która spełnia równanie (6.1) nazywa się *rozwiązaniem klasycznym.* Często rozważa się również *rozwiązania uogólnione* od których nie wymaga się takiej regularności. My będziemy tutaj zajmować się jedynie *rozwiązaniami klasycznymi.* Równanie postaci (6.1), to równanie **rzędu 1.** Jest to naprawdę *układ m równań różniczkowych zwyczajnych.* My przeważnie nie będziemy rozróżniać jednego równania od układu, traktując (6.1) jako jedno równanie ze względu na funkcję wektorową u, mającą wartości w przestrzeni $\mathbf{R^m}$.

Równanie różniczkowe zwyczajne **rzędu n** jest postaci

$$\frac{d^n u(t)}{dt^n} = f(t, u(t), \frac{du(t)}{dt}, \cdots, \frac{d^{n-1}u(t)}{dt^{n-1}}) \tag{6.2}$$

Zauważmy, że definiując n nowych fukcji

$$u_j = \frac{d^{(j-1)}u}{dt^{j-1}}, \quad j = 1, 2, \cdots, n$$

funkcję wektorowe wymiaru nm, $v = [u_1, u_2, \cdots, u_n]^T$, oraz

$$F(t, v) = [u_2, u_3, \cdots, u_n, f(t, u_1, u_2, \cdots, u_n)]^T,$$

równanie (6.2) możemy zastąpić równoważnym równaniem rzedu 1

$$\frac{dv(t)}{dt} = F(t, v(t)). \tag{6.3}$$

Wynika stąd, że wystarczy zajmować się równaniami rzędu 1.

Równania postaci (6.1) **mogą mieć wiele rozwiązań.** Przykładem jest bardzo proste równanie skalarne

$$\frac{du(t)}{dt} = 0 \tag{6.4.}$$

Jego rozwiązaniem jest $u(t) = C$, gdzie C jest dowolną stała. Natomiast równanie

$$\frac{du(t)}{dt} = \begin{cases} 0 & \text{dla } t = 0 \\ 1 & \text{dla } t \neq 0 \end{cases}$$

nie ma wogóle rozwiązania klasycznego (to jest rozwiązania klasy C^1) na żadnym przedziale zawierającym we wnętrzu 0.

Zagadnienie Cauchy'ego (zagadnienie początkowe). Załóżmy, że funkcja f jest określona i ciągła na zbiorze $D \subset \mathbf{R} \times \mathbf{R}$

$$D = \{(t, u)| \ \ |t - t_0| \leq a, \ \ |u_j - u_{j_0}| \leq b, \ \ j = 1, 2, \cdots, m\},$$

gdzie $0 \leq a \leq \infty$, $0 \leq b \leq \infty$.

Zagadnienie Cauchy'ego polega na poszukiwaniu rozwiązania u równania różniczkowego

$$\frac{du(t)}{dt} = f(t, u(t)), \tag{6.5}$$

spełniającego **warunek pczątkowy** (warunek Cauchy'ego)

$$u(t_0) = u_0, \tag{6.6}$$

gdzie t_0 i u_0 są zadane.

Zauważmy, że nasze przykładowe równanie (6.4)

$$\frac{du(t)}{dt} = 0$$

uzupełnione warunkiem początkowym $u(0) = 0$ ma już **jednoznaczne rozwiązanie** $u(t) = 0$. Okazuje sią, że dla dużej klasy równań (6.1) można udowodnić *istnienie i jednoznaczność* zagadnienia Cauchy'ego (6.5), (6.6). Założymy, że funkcja f jest ciągła, i że w zbiorze D postaci (6.4) spełnia ona **warunek Lipschitza** ze względu na zmienna (wektorową) u.

Istnieje stała $L \geq 0$, taka że dla dowolnych (t, u_1) i (t, u_2) ze zbioru D

$$|f(t, u_1) - f(t, u_2)| \leq L|u_1 - u_2|. \tag{6.7}$$

Tutaj $|\cdot|$ oznacza dowolną ustaloną normę w przestrzeni $\mathbf{R^m}$.

Nasze równanie

$$\frac{du(t)}{dt} = f(t, u(t))$$

scałkujemy względem t w przedziale (t_0, t) gdzie $t \in [t_0, t_0 + a]$, $a > 0$, uwzględniając warunek początkowy

$$u(t) = u_0 + \int_{t_0}^{t} f(s, u(s))ds. \tag{6.8}$$

Jest to *równanie całkowe* Zauważmy, że każde rozwiązanie równania (6.5) z warunkiem (6.6) spełnia równanie całkowe (6.8).

Na równanie (6.8) popatrzmy teraz nieco inaczej. Niech X będzie przestrzenią wszystkich funkcji ciągłych o wartościach w $\mathbf{R^m}$, określonych na przedziale $[t_0, t_0 + a]$. Załóżmy, że $0 \leq a < \infty$. W prestrzeń X wyposażymy w normę

$$\|u\|_{\infty,[t_0,t_0+a]} = \sup_{t_0 \leq t \leq t_0+a} |u(t)|.$$

Wiemy, że $(X, \|\cdot\|_{\infty,[t_0,t_0+a]})$ jest przestrzenią Banacha. Dla uproszczenia rachunków założymy, że funkcja f jest określona ciągła i że spełnia warunek Lipschitza w zbiorze $D = [t_0, t_0 + a] \times \mathbf{R^m}$. Niech

$$\Phi(u)(t) = u_0 + \int_{t_0}^{t} f(s, u(s))ds$$

dla $u \in X$. Jest oczywiście

$$\Phi : X \to X.$$

Będzie nas interesowało rozwiązanie u równania (6.8) dla $t \in [t_0, t_0+\alpha]$, gdzie $0 \leq \alpha \leq a$. Dla dowolnych $u_1, \quad u_2 \in X$ mamy

$$\|\Phi(u_1) - \Phi(u_2)\|_{\infty,[t_0,t_0+a]} = \sup_{t_0 \leq t \leq t_0+a} |\int_{t_0}^{t} [f(s, u_1(s)) - f(s, u_2(s)]ds| \leq$$

$$\leq \sup_{t_0 \leq t \leq t_0+a} \int_{t_0}^{t} |[f(s, u_1(s)) - f(s, u_2(s))]| ds \leq$$

$$\leq \sup_{t_0 \leq t \leq t_0+a} L \int_{t_0}^{t} |u_1(s)) - u_2(s)| ds \leq L\alpha \|u_1 - u_2\|_{\infty,[t_0,t_0+a]}.$$

Mamy więc

$$\|\Phi(u_1) - \Phi(u_2)\|_{\infty,[t_0,t_0+a]} \leq L\alpha \|u_1 - u_2\|_{\infty,[t_0,t_0+a]}.$$

Stąd wynika, że Φ jest *przekształceniem zwężającym*, gdy

$$0 \leq \alpha < \frac{1}{L}.$$

Wiemy na podstawie Twierdzenia Banacha, że warunek ten pociąga istnienie jedynego punktu stałego $u = \Phi(u)$. Oznacza to, że równanie (6.8) **ma dokładnie jedno rozwiązanie** u dla $t \in [t_0, t_0 + \alpha]$, gdy $\alpha < \frac{1}{L}$. Ponieważ zaś u i f są funkcjami ciągłymi, to u jest różniczkowalna w sposób ciągły, a zatem jest rozwiązaniem zagadnienia Cauchy'ego (6.5)(6.6) dla $t \in [t_0, t_0 + \alpha]$. Wykazaliśmy w ten sposób

Twierdzenie Picard'a - Lindelöf'a. *Jeśli $f : D \to \mathbf{R}^{\mathbf{m}}$ jest ciągła, gdzie D jest postaci* (6.4)*, oraz jeśli funkcja f spełnia warunek Lipschitza* (6.7)*, to dla $t \in [t_0, t_0 + \alpha]$ gdzie α spełnia nierówność $0 < \alpha \leq a$ i jest dostatecznie małe,* **istnieje jednoznaczne rozwiązanie** *zagadnienia Cauchy'ego* (6.5) (6.6).

Komentarze.

1. Twierdzenie Picard'a - Lindelöf'a ma *charakter lokalny*. To znaczy, mówi ono o istnieniu i jedoznaczności rozwiązania u ale tylko w pewnym przedziale $[t_0, t_0 + \alpha]$, dla pewnego *małego* α, spełniającego nierówność $0 < \alpha \leq a$. Dowodzi się, że istnieje przedział maksymalnej długości, zawierający przedział $[t_0, t_0 + \alpha]$, na który można przedłużyć to rozwiązanie lokalne.

2. Jeśli założyć, że funkcja f, określająca równanie

$$\frac{du(t)}{dt} = f(t, u(t))$$

jest jedynie ciągła i ograniczona w D, to można jedynie udowodnić istnienie rozwiązania lokalnego. Przy tych założeniach rozwiązanie nie musi być jednoznaczne.

3. Jeśli funkcja f zależy dodatkowo od *parametru* λ, i spełnione są założenia Twierdzenia Picarda - Lindelöf'a, na przykład

$$\frac{du(t)}{dt} = f(t, u(t), \lambda), \quad \lambda \in \mathbf{R}^{\mathbf{d}},$$

to rozwiązanie u także jest funkcją parametru λ. Przy tem, jeśli f jest funkcją ciągłą zmiennej λ, (jest różniczkowalna p-krotnie względem λ i u), to także u jest funkcją ciągłą λ (jest różniczkowalna p-krotnie względem λ.) To samo dotyczy *warunku początkowego*. Rozwiązanie jest funkcją ciągłą warunku początkowego, zaś przy założeniu p-krotnej różniczkowalności f względem u, u jest p-krotnie różniczkowalna względem warunku począkowego.

4. Pochodna rozwiązania u względem parametru λ $v = \frac{\partial u}{\partial \lambda}$ spełnia równanie różniczkowe otrzymane przez formalne zróżniczkowanie równania oryginalnego względem tego parametru:

$$\frac{dv(t, \lambda)}{dt} = \frac{\partial}{\partial \lambda} f(t, u, \lambda) + \frac{\partial}{\partial u} f(t, u, \lambda) v(t, \lambda),$$

i spełnia warunek początkowy

$$v(t_0, \lambda) = \frac{\partial}{\partial \lambda} u_0.$$

Równanie o zmiennych rozdzielonych. Niektóre równania różniczkowe można rozwiązać efektywnie, albo też rozwiązanie wyrazić przez całki pewnych funkcji. Takimi równaniami są między inymi równania **skalarne o zmiennych rozdzielonych**

$$\frac{du(t)}{dt} = f(t) g(u(t)), \tag{6.9}$$

gdzie $u(t) \in \mathbf{R}$, $f(t) \in \mathbf{R}$, $g(u) \in \mathbf{R}$. Załóżmy, że $g(u) \neq 0$ w całej dziedzinie g. Wtedy łatwo udowodnić, że rozwiązanie u spełnia **równanie całkowe**

$$\int \frac{du}{g(u)} = \int f(t)dt. \tag{6.10}$$

Jeśli potrafimy efektywnie obliczyć całki, to otrzymamy równanie (nieliniowe) określające, na ogół w sposób *uwikłany* u jako funkcję zmiennej t, lub t jako funkcję zmiennej u.

Zadanie. Udowodnij, że rozwiązanie równania (6.9) istnieje i spełnia równanie całkowe (6.10).

Przykład. Równanie skalarne, liniowe, jednorodne.

$$\frac{du(t)}{dt} = a(t)u(t), \tag{6.11}$$

sprowadza się do

$$\int \frac{du}{u} = \int a(t)dt.$$

Całka po lewej stronie da się obliczyć, zatem, po przekształceniach

$$u(t) = e^{\int_{t_0}^{t} a(s)ds} C, \tag{6.12}$$

gdzie C jest dowolną stałą, którą wyznaczamy przy pomocy zadanego warunku początkowego. Jeśli $u(t_0) = u_0$, to

$$u(t) = e^{\int_{t_0}^{t} a(s)ds} u_0.$$

Zadanie. Udowodnij, że **równanie liniowe niejednorodne**

$$\frac{du(t)}{dt} = a(t)u(t) + f(t), \tag{6.13}$$

gdzie a i f są ciągłe, ma rozwiązanie postaci

(6.14). $$u(t) = e^{\int_{t_0}^{t} a(s)ds} C + w(t)$$

Wyrażenie wykładnicze jest *rozwiązaniem równania jednorodnego*, zaś w jest *jakimkolwiek* rozwiązaniem równania (6.13). Funkcję $w(t)$ znajdujemy tak zwaną *metodą uzmienniania stałej.* Polega ona na tym, że rozwiązania równania (6.13) poszukujemy w postaci

$$w(t) = e^{\int_{t_0}^{t} a(s)ds} C(t),$$

gdzie teraz $C(t)$ jest funkcją zmiennej t, którą należy wyznaczyć. Wyprowadź ostateczny wzór dla rozwiązania u.

Definicja. *Rozwiązaniem ogólnym równwnia różniczkowego*

$$\frac{du(t)}{dt} = f(t, u(t)), \quad u \in \mathbf{R^m}$$

nazywamy rozwiązanie u zależne od t i **dowolnej stałej** $C \in \mathbf{R^m}$.

Przykład. Funkcja (6.12) jest rozwiązaniem ogólnym równania (6.11) zaś funkcja (6.14), rozwiązaniem ogólnym równania (6.13).

Ważna uwaga. Jeśli przyjrzymy się wzorom (6.12) i (6.14), podającym odpowiednio rozwiązanie równania jednorodnego i niejedorodnego skalarnego,

$$\frac{du}{dt} = au,$$

$$\frac{du}{dt} = au + f,$$

to zauważymy, że *zbiór wszystkich rozwiązań* równania jednorodnego jest *jednowymiarową przestrzenią liniową.* Baza tej przestrzeni składa się z jednego elementu $\phi(t) = e^{\int_{t_0}^{t} a(s)ds}$. Zbiór wszystkich rozwiązań równania niejednorodnego jest *jednowymiarową rozmaitością liniową* zawierającą punkt $w(t)$.

Układy równań różniczkowych zwyczajnych liniowych. Niech $A(t)$ będzie macierzą wymiaru $m \times m$ zależną w sposób ciąły od $t \in [t_0, t_0+a]$, $a > 0$. Zajmiemy się najpierw *układem jednorodnym*

$$\frac{du(t)}{dt} = A(t)u(t). \tag{6.15}$$

Zadanie. Odpowiedz, czy równanie (6.15) można rozwiązać metodą *rozdzielenia zmiennych*, gdy $m > 1$.

Zadanie. Udowodnij, że równanie (6.15) z warunkiem początkowym $u(t_0) = u_0 \in \mathbf{R^m}$ ma jednoznaczne rozwiązanie.

Twierdzenie 6.1 *Zbiór wszystkich rozwiązań równania (6.15) jest m-wymiarową przestrzenią liniową.*

Dowód. Ze względu na jednoznaczność rozwiązania równania (6.15) z warunkiem początkowym $u(t_0) = u_0$ możemy każdenu wektorowi $u_0 \in \mathbf{R^m}$ przyporządkować w sposób wzajemnie jednoznaczny rozwiązanie $u(t)$ spełniające ten warunek początkowy. Otrzymujemt w ten sposób *izomorfizm* przestrzeni wszystkich rozwiązań i $\mathbf{R^m}$. $\Box$

Jeśli $\phi_1(t), \phi_2(t), \cdots, \phi_m(t)$ jest bazą przestrzeni rozwiązań równania (6.15), to macierz

$$X(t) = [\phi_1(t), \phi_2(t), \cdots, \phi_m(t)]$$

wymiaru $m \times m$, której kolumnami są funkcje wektorowe $\phi_1, \phi_2, \cdots, \phi_m$ nazywa się *macierzą fundamentalną układu* (6.15). Dowolne rozwiązanie $u(t)$ równania (6.15) da się wyrazić w postaci

$$u(t) = X(t)c,$$

gdzie $c \in \mathbf{R^m}$ jest pewnym wektorem stałym. Ten fakt można inaczej sformułować tak:
Zbiór wszystkich rozwiązań równania (6.15) *tworzy przestrzeń liniową wymiaru m* Nie trudno zauważyć, że

$$\frac{dX(t)}{dt} = A(t)X(t).$$

Dowodzi się, że jeśli $\det(X(t_0)) \neq 0$ dla pewnego t_0, to dla każdego $t \in [t_0, t_0 + a]$, $\det(X(t)) \neq 0$.

Zadanie. Udowodnij, że $X^{-1}(t)$ spełnia równanie

$$\frac{dX^{-1}(t)}{dt} = -X^{-1}(t)A(t).$$

Jak wybrać warunek początkowy?

Zadanie. Udowodnij, że dowolne rozwiązanie u liniowego układu niejednorodnego

$$\frac{du(t)}{dt} = A(t)u(t) + f(t), \quad f \in C([t_0, t_0 + a]), \tag{6.16}$$

wyraża się przez *macierz fundamentalną* $X(t)$

$$u(t) = X(t)c + w(t),$$

gdzie $c \in \mathbf{R^m}$, a $w(t)$ jest *jakimś* rozwiązaniem równania (6.16). Wyznacz wzory dla w i u. Dla wyznaczenia w użyj opisanej już *metody uzmienniania stałej.*

Powyższe zadanie można interpretować tak
Zbiór rozwiązań układu liniowego niejednorodnego tworzy rozmaitość liniową zawierającą punkt w.

Układy równań liniowych o stałych współczynnikach. Zajmiemy się teraz układami liniowymi postaci

$$\frac{du(t)}{dt} = Au(t), \tag{6.17}$$

gdzie A jest macierzą stałą wymiaru $m \times m$.

Macierz wykładnicza. Jeśli B jest dowolną macierzą stałą wymiaru $m \times m$, to z definicji

$$e^B = I + \frac{B}{1!} + \frac{B^2}{2!} + \frac{B^3}{3!} + \cdots \tag{6.18}$$

Zadanie. Udowodnij, że szereg (6.18) jest bezwzględnie zbieżny, to znaczy, że

$$\|I\| + \frac{\|B\|}{1!} + \frac{\|B\|^2}{2!} + \frac{\|B\|^3}{3!} + \cdots.$$

Stąd wynika poprawność definicji (6.18).

Nie trudno sprawdzić, że *macierzą fundamentalną* układu (6.17) jest

$$X(t) = e^{At}. \tag{6.19}$$

Zadanie. Odpowiedz, czy wzór $e^{A(t)t}$ przedstawia macierz fundamentalną układu (6.15). Uzasadnij odpowiedź.

Przyjrzyjmy się czym jest naprawdę macierz wykładnicza, bowiem operowanie szeregiem (6.18) jest raczej nie wygodne.

$$X(t) = e^{At} = I + \frac{At}{1!} + \frac{(At)^2}{2!} + \frac{(At)^3}{3!} + \cdots .$$

Zastosujmy Twierdzenie Jordana o rozkładzie spektralnym do macierzy A. Mamy

$$A = TJT^{-1}, \tag{6.20}$$

gdzie T jest macierzą nieosobliwą, zaś J jest macierzą klatek Jordana

$$J = \begin{bmatrix} J_1 & 0 & 0 & \cdots & 0 \\ 0 & J_2 & 0 & \cdots & 0 \\ 0 & 0 & J_3 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & J_p \end{bmatrix}.$$

Klatki są postaci $J_s = \lambda_s I + E_s$, gdzie λ_s jest wartością własną macierzy A;

$$E_s = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & 1 \\ 0 & 0 & 0 & \cdots & 0 \end{bmatrix}.$$

Niech rozważana klatka ma wymiar $d_s \times d_s$. Macierz E_s ma tę własność, ze podniesienie jej do l-tej potęgi powoduje przesunięcie jedynek o $l-1$ miejsc w prawo. Stąd wynika, że

$$E_s^{d_s} = 0.$$

Zastosujmy teraz rozkład (6.20) w równaniu (6.17); jeśli oznaczymy $v(t) = T^{-1}u(t)$, to otrzymamy

$$\frac{dv(t)}{dt} = Jv(t). \tag{6.21}$$

Łatwo sprawdzić, że macierz fundamentalna $Y(t)$ układu (6.21) jest postaci

$$Y(t) = e^{Jt} = \begin{bmatrix} e^{J_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{J_2 t} & 0 & \cdots & 0 \\ 0 & 0 & e^{J_3 t} & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 0 & \cdots & e^{J_p t} \end{bmatrix}.$$

Przyjrzyjmy sią klatkom $e^{J_s t}$:

$$e^{J_s t} = e^{(\lambda_s + E_s)t} = e^{\lambda_s t} e^{E_s t}.$$

Zauważmy, że $e^{\lambda_s t}$ jest skalarem. Przekształćmy jeszcze $e^{E_s t}$. Z definicji *macierzy wykładniczej*, ze względu na to, że $E_s^{d_s} = 0$

$$e^{E_s t} = I + \frac{E_s t}{1!} + \frac{{E_s}^2 t^2}{2!} + \cdots + \frac{{E_s}^{d_s - 1} t^{d_s - 1}}{d_{s-1}!},$$

gdzie d_s jest wymiarem klatki J_s. Oznacza to, że $e^{E_s t}$ jest poprostu *wielomianem* od macierzy $E_s t$. Oryginalne rozwiązanie u otrzymamy mnożąc

$$u(t) = Tv(t) = TY(t)c,$$

gdzie c jest dowolą stałą, $c \in \mathbf{R^m}$. Macierzą fundamentalną układu oryginalnego (6.17) jest więc

$$X(t) = e^{TJT^{-1}}t = Te^{Jt}T^{-1}.$$

We wzorze określającym $X(t)$ występują tylko funkcje wykładnicze $e^{\lambda_s t}$ oraz wielomiany pewnych macierzy.

Ponieważ *rozwiązanie ogólne* równania różniczkowego zależy od dowolnej stałej $C \in \mathbf{R^m}$. Zagadnienie Cauchy'eo ma już często jednoznaczne rozwiązanie. Dodanie warunku początkowego nie jest jedynie zabiegiem *ujednoznaczniającym.* Zagadnienia początkowe, są często *naturalnymi*zagadnieniami opisującymi pewne zjawiska fizyki, przyrody, techniki itp. Inny typ zagadnień stawianych dla równań różniczkowych, to *zagadnienia brzegowe.* Zagadnienie brzegowe powstaje przez dodanie do równania różniczkowego zwyczajnego *warunku brzegowego* to jest pewnego warunku wiążącego wartości rozwiązania na brzegach odcinka, na którym rozpatrujemy równanie. Przykładem zagadnienia brzegowego jest

$$\frac{d}{dt}u(t) = f(t, u(t)) \quad t \in (a, b),$$
$$g(u(a), u(b)) = 0,$$

gdzie g jest pewną funkcją, $g : \mathbf{R^m} \times \mathbf{R^m} \to \mathbf{R^m}$. Teoria zagadnienia brzegowego jest znacznie bardziej skomplikowana niż teoria zagadnienia Cauchy'ego.

NUMERYKA ZAGADNIENIA CAUCHY'EGO

Opiszemy tu jedynie niektóre **metody różnicowe**, to jest takie, które równanie różniczkwe zastępują pewnym **równaniem różnicowym**. Weźmy pod uwagę zagadnienie Cauchy'ego

$$\frac{d}{dt}u(t) = f(t, u(t)), \tag{6.22}$$

$$u(t_0) = u^0. \tag{6.23}$$

Załóżmy, że rozwiązanie u istnieje i jest jednoznaczne w przedziale $[t_0, t_0 + a]$ i że ma w tym przedziale tyle pochodnych ile będzie nam potrzeba. Na tym przedziale zbudujemy *siatkę punktów*, dla uproszczenia, *równoodległych*

$$t_0 < t_1 < t_2 \cdots < t_N,$$
$$t_j = t_0 + jh, \quad h = \frac{a}{N}, \quad h > 0.$$

Liczbę h będziemy nazywali *krokiem siatki* lub *krokiem całkowania.* W praktyce często potrzebne są siatki ze *zmiennym krokiem*, my jednak ograniczymy się tu do siatek o stałym kroku. Rozwiązanie u zagadnienia Cauchy'ego (6.22),(6.23) rozwiniemy przy pomocy wzoru Taylora dla $t = t_{k+1}$, w punkcie t_k

$$u(t_{k+1}) = u(t_k) + hu'(t_k) + \frac{h^2}{2}u''(t_k) + \frac{h^3}{6}u'''(t_k) + \cdots,$$

lub

$$u(t_{k+1}) = u(t_k) + hf(t_k u(t_k)) + O(h^2).$$

Odrzucając wyrazy zawierające h w potędze 2 i wyższej, otrzymamy *równanie różnicowe*

$$u_{k+1} = u_k + hf(t_k, u_k), \quad u_0 = u^0. \tag{6.24}$$

Rozwiązaniem tego równania jest ciąg $\{u_j\}, \quad j = 0, 1, \cdots$. Element ciągu u_k odpowiada wartości rozwiązania $u(t_k)$, zagadnienia Cauchy,ego (6.22),(6.23). Można postąpić inaczej; rozwinąć $u(t_k)$ w punkcie t_{k+1}

$$u(t_k) = u(t_{k+1}) + hu'(t_{k+1}) + \frac{h^2}{2}u''(t_{k+1}) + O(h^3).$$

Podobnie jak poprzednio, odrzucając wyrazy zawierające h w potędze 2 i wyższej otrzymamy *inne równanie różnicowe*

$$u_{k+1} = u_k + hf(t_{k+1}, u_{k+1}), \quad u_0 = u^0. \tag{6.25}$$

Wygodnie będzie dalej oznaczać

$$f_k = f(t_k, u_k).$$

Równania (6.24) i (6.25) noszą nazwę **Schematów Eulera** (schemat = metoda).

(6.24) $\qquad u_{k+1} = u_k + hf_k$ **Schemat otwarty Eulera**,

(6.25) $\qquad u_{k+1} = u_k + hf_{k+1}$ **Schemat zamknięty Eulera**.

Schematem otwartym Eulera łatwo się posługiwać. Znając warunek początkowy u_0, drogą podstawiania do wzoru (6.24) obliczymy u_k dla każdego interesującego nas k.

Zupełnie inaczej jest ze schematem zamkniętym (6.25). Aby wyliczyć u_{k+1} znając u_k, trzeba rozwiązać **równanie nieliniowe** (układ m-równań!)

$$u_{k+1} = u_k + hf(t_{k+1}, u_{k+1}).$$

Chwilowo **nie potrafimy nic powiedzieć** na temat zależności ciągów o elementach u_k, oraz $u(t_k)$, $k = 0, 1, 2, \cdots$. Chcielibyśmy, aby spełniony był

Warunek zbieżności schematu. *Przypuśćmy, że dla dowolnego ustalonego* $t \in [t_0, t_0 + a]$, *siatka punktów* $\{t_k\}$ *została tak dobrana, że* $t = t_k = t_0 + kh$.

Będziemy uważać rozważany schemat za **zbieżny**, *jeśli warunek*

$$u_k \to u(t), \quad \textit{gdy} \quad h \to 0 \quad (\textit{stąd} \;\; k = \frac{t - t_0}{h}, \quad k \to \infty). \tag{6.26}$$

zachodzi

- *dla dowolnego rozwiązania u dowolnego równania* (6.1) *z warunkiem początkowym* $u(t_0) = u^0$, *należącego do klasy równań spełniających założenia Twierdzenia Picard'a - Lindelöf'a,*
- *dla dowolnego rozwiązania* $\{u_k\}, \quad k = 0, 1, \cdots$ *rozważanego schematu, dla którego wartość* $u_0 = u_0(h)$ *spełnia warunek*

$$u_0 \to u^0, \quad \textit{gdy} \quad h \to 0.$$

Dalej będą nas interesowały jedynie te schematy, które są **zbieżne** w powyższym sensie. Zobaczymy też jak odróżniać schematy zbieżne od niezbieżnych.

Chwilowo powróćmy do schematu zamkniętego Eulera

$$u_{k+1} = u_k + f_{k+1}.$$

Aby obliczyć u_{k+1} trzeba **rozwiązać zadanie na punkt stały**

$$x = \Phi(x),$$

gdzie $x = u_{k+1}$, $\Phi(x) = u_k + hf(t_{k+1}, x)$. Sprobujmy zastosować Twierdzenie Banacha o punkcie stałym. Zbudujmy ciąg wektorów $\{x_l\}$ $l = 0, 1, 2, \cdots$, x_0-dowolny element,

$$x_{l+1} = \Phi(x_l).$$

Ciąg będzie zbiegał do punktu stałego $x = u_{k+1}$, jeśli Φ spełmia **warunek Lipschitza ze stałą** L_1**,** $0 \leq L_1 < 1$**.** Przypuśćmy, że funkcja f (prawa strona równania (6.1)) spełnia założenia Twierdzenie Picard'a - Lindelöf'a ze stałą Lipschitza L. Wtedy dla dowolnych x i y takich, że (t, x) i (t, y) należą do dziedziny funkcji f

$$|\Phi(x) - \Phi(y)| = h|[f(t_{k+1}, x) - f(t_{k+1}, y)]| \leq hL|x - y| = L_1|x - y|.$$

Widzimy, że $L_1 < 1$, gdy

$$0 < h < \frac{1}{L}. \tag{6.27}$$

Zatem iteracja

$$u_{k+1}^{l+1} = u_k + hf(t_{k+1}, u_{k+1}^l) \tag{6.28}$$

zbiega do u_{k+1} dla dowolnego punktu startowego u_{k+1}^0, jeśli

$$h < \frac{1}{L}.$$

Warunek (6.27) nie jest *bardzo ograniczający*, jeśli stała Lipschitza L funkcji f nie jest zbyt duża. W przypadku wielkich wartości L lepiej stosować *iterację Newtona*. Zauważmy, że koszt algorytmu wykorzystującego schemat zamknięty skupia się głównie w wyliczaniu wartości funkcji f. Zatem należy wyliczać wartości f jak najmniej razy. Liczba iteracji zależy od tego **jak dobrze dobrany został punkt startowy** u_{k+1}^0. Dobry start iteracji zapewnia przyjęcie jako u_{k+1}^0 wartości u_{k+1} uzyskanej z zastosowania schematu otwartego Eulera.
W ten sposób doszliśmy do tak zwanej **METODY PREDICTOR - CORRECTOR** opartej na schematach Eulera.

- **PREDICTOR**, to schemat otwarty podający punkt startowy dla iteracji - stosowany 1 raz na krok.
- **CORRECTOR**, to schemat zamknięty służący do iterowania. Iterujemy małą liczbę razy, gdyż punkt startowy jest **blizko rozwiązania.**

Metodę **PREDICTOR - CORRECTOR** w taki sam sposób można budować w oparciu o inne pary schematów [16], złożone ze schematu otwartego (PREDICTOR) i zamkniętego (CORRECTOR).

Narzuca sie pytanie: **po co stosować skomplikowane w użyciu schematy zamknięte, skoro dysponujemy bardzo wygodnymi schematami otwartymi?** Okazuje się, że pewne cechy stawiają metodę zamkniętą zdecydowanie wyżej od metody otwartej. Są zadania, których nie daje się

[16] Schematy takie poznamy w dalszej części tego wykładu.

wogóle policzyć metodą otwartą, a którym *daje radę* metoda zamknięta. To co odróżnia schemat Eulera otwarty od zamkniętego, to na pewno **nie jest rząd**.
Co to jest **rząd schematu?**

Niech $u(t)$ będzie rozwiązaniem zagadnienia Cauchy'ego (6.1), (6.2), *o którym zakładamy, że ma $p+1$ pochodnych ciągłych. Oznaczmy przez*

$$S(\{u_l\}, l = 0, 1, 2, \cdots) = 0 \tag{6.29}$$

nasz schemat różnicowy.
Mówimy, że schemat (6.29) **jest rzędu p**, *jeśli podstawiając do* (6.29) *ciąg*

$$\{u(t_j)\}, \quad j = 0, 1, 2, \cdots$$

zamiast ciągu $\{u_j\}$ $\;j = 0, 1, 2, \cdots$, otrzymamy

$$S(\{u(t_j)\}, j = 0, 1, 2, \cdots) = R,$$

gdzie **reszta R** *spełnia warunek*

$$R = O(h^{p+1}),$$

zaś istnieje takie zadanie Cauchy'ego spełniające powyższe warunki, dla którego $R \neq O(h^{p+2})$.

Biorąc pod uwagę sposób w jaki otrzymaliśmy oba schematy Eulera widzimy, że **oba są rzędu 1**.

Zanim przejdziemy, do wyjaśnienia na czym polega wyższość schematu zamkniętego nad otwartym, przyjrzyjmy się jeszcze innym schematom różnicowym. Niech $u \in C^3$. Mamy

$$u(t_{k+1}) = u(t_k) + hu'(t_k) + \frac{h^2}{2!}u''(t_k) + \frac{h^3}{3!}u'''(t_k) + \cdots, \tag{6.29}$$

$$u(t_{k+1}) = u(t_k) + hu'(t_{k+1}) - \frac{h^2}{2!}u''(t_{k+1}) + \frac{h^3}{3!}u'''(t_{k+1}) + \cdots. \tag{6.30}$$

Zauważmy jeszcze, że $u''(t_{k+1}) = u''(t_k) + hu'''(t_k) + O(h^2)$. Dodajmy stronami wzory (6.29) i (6.30) uwzględniając powyższą uwagę. Otrzymamy tak zwany **schemat trapezów**

$$u_{k+1} = u_k + \frac{h}{2}(f_k + f_{k+1}). \tag{6.31}$$

Jest to **schemat zamknięty, rzędu 2**.

Zadanie. Zapisz *iteracje Banacha i Newtona* dla schematu trapezów.

Wszystkie trzy schematy, które dotychchczas poznaliśmy są **jednokrokowe**, to znaczy, że mając do dyspozycji jedynie u_k, możemy wyliczyć u_{k+1}.

Zadanie. Schematy Taylora. Używając rozwinięcia Taylora dla rozwiązania $u(t)$ zagadnienia początkowego (6.1), (6.2), uwzględniając drugie i ewentualnie wyższe pochodne u zbuduj **schematy jednokrokowe rzędu wyższego niż 1.**
Wskazówka. Zauważ, że

$$u''(t) = \frac{\partial}{\partial t} f(t, u(t)) + \frac{\partial}{\partial u} f(t, u(t)) f(t, u(t)).$$

Podobnie dla wyższych pochodnych.

Odwołajmy się jeszcze raz do wzoru Taylora. Podobnie jak poprzednio

$$u(t_{k+1}) = u(t_k) + hu'(t_k) + \frac{h^2}{2} u(t_k) + \frac{h^3}{6} u'''(t_k) + dots,$$

$$u(t_{k-1}) = u(t_k) - hu'(t_k) + \frac{h^2}{2} u(t_k) - \frac{h^3}{6} u'''(t_k) + \cdots.$$

Odejmijmy stronami te równości. Otrzymamy schemat "Midpoint"

$$u_{k+2} = u_k + 2hf_{k+1}.$$

Schemat Midpoint, to schemat otwarty. Nie jest on schematem jednokrokowym, gdyż u_{k+2} możemy wyliczyć tylko jeśli dysponujemy dwoma wartościami u_k i u_{k+1}. Aby schemat *mógł wystartować* potrzebne są *dwa warunki początkowe u_o* i u_1. Mówimy, że taki schemat *nie jest samostartujący.* Jeśli dysponujemy warunkiem początkowym u^0, to aby uruchomić schemat Midpoint musimy dodatkowo *doliczyć* wartość u_1. Można to zrobić używając jakiejś metody jednokrokowej. Nie jest jednak obojętne jakiej metody użyjemy. Ze sposobu konstrukcji schematu Midpoint wynika, że jest on rzędu 2 (reszta odrzucona jest rzędu $O(h^3)$). Zatem dla zachowania rzędu powinniśmy zadbać o to, aby u_1 wyliczyć również schematem rzędu 2.

Okazuje się, że schemat Midpoint, mimo że ma rząd 2, zawodzi w pewnych przypadkach z którymi schemat otwarty Eulera (który jest rzędu 1) radzi sobie całkiem dobrze.

Zadanie. Napisz program rozwiązujący zagadnienie Cauchy'ego

$$\frac{d}{dt}u(t) = -\lambda u(t), \quad \lambda > 0.$$

$$u(0) = 1.$$

Użyj schematu otwartego Eulera i schematu Midpoint dla tego samego zadania. Porównaj zachowanie się schematów gdy wykonujesz dużą liczbę kroków przy jednakowej wartości kroku h i stałej $\lambda > 0$. Porównaj co się dzieje dla różnych wartości h i λ.

Schematy liniowe wielokrokowe. Przykładem takiego schematu jest schemat Midpoint. Schemat liniowy q - krokowy jest równaniem różnicowym, *na ogół nieliniowym*, postaci

$$\sum_{j=0}^{q} \alpha_j u_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}, \tag{6.30}$$

gdzie jak poprzednio $f_l = f(t_l, u_l)$.

Aby wystartować, taki schemat potrzebuje q warunków początkowych $u_0, u_1, \cdots, u_{q-1}$, które trzeba *doliczyć* schematem jednokrokowym odpowiednio wysokiego rzędu. Współczynniki $\alpha_j, \quad \beta_j, \quad j = 0, 1, \cdots, q$ można wyznaczyć tak, aby *rząd schematu był odpowiednio wysoki*, oraz żeby posiadał on jeszcze inne cechy, o których powiemy później. Zauważmy teraz, że schemat (6.30) jest

- otwarty, gdy $\beta_q = 0$,
- zamknięty, gdy $\beta_q \neq 0$.

Zadanie. Zbuduj schemat postaci (6.30) dlla q=1, który ma najwyższy możliwy rząd.

Powróćmy jeszcze do schematów jednokrokowych. Specjalną klasę takich schematów stanowią **schematy Runge - Kutty.** Schemat Runge - Kutty q - poziomowy jest postaci

$$u_{k+1} = u_k + h[c_1K_1 + c_2K_2 + \cdots + c_qK_q], \tag{6.31}$$

gdzie

$$K_j = f(t_k + ha_j, u_k + h\sum_{l=1}^{q} b_{j,l}K_l) \quad j = 1, 2, \cdots, q. \tag{6.32}$$

Współczynniki

$$\begin{matrix} c_1 & c_2 & c_3 & \cdots & c_q \\ a_1 & a_2 & a_3 & \cdots & a_q \\ b_{1,1} & b_{1,2} & b_{1,3} & \cdots & b_{1,q} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ b_{q,1} & b_{q,2} & b_{q,3} & \cdots & b_{q,q} \end{matrix} \tag{6.32}$$

wyznacza się tak, aby uzyskać możliwie wysoki rząd, oraz jeszcze inne cechy schematu. Taką cechą może być na przykład *jego otwartość*. Schemat będzie otwarty, jeśli zażądamy, aby

$$b_{j,l} = 0 \;\text{ dla }\; l \geq j.$$

Schemat zamknięty wymaga rozwiązania na każdym kroku układu qm równań dla wyznaczenia $K_1, K_2, \cdots, K_q$. Współczynniki (6.32) dla różnych schematów są znane od wielu dziesiątek lat.
Przytoczymy tu dwa przykłady schematów Runge - Kutty.
Schemat 4- poziomowy otwarty, rzędu 4.

$$u_{k+1} = u_k + \frac{h}{6}[K_1 + 2K_2 + 2K_3 + K_4],$$

$$K_1 = f(t_k, u_k),$$

$$K_2 = f(t_k + \frac{h}{2}, u_k + \frac{h}{2}K_1), \tag{6.33}$$

$$K_3 = f(t_k + \frac{h}{2}, u_k + \frac{h}{2}K_2),$$

$$K_4 = f(t_k + h, u_k + hK_3),$$

Jest to bardzo często używany schemat.
Schemat 2- poziomowy zamknięty, rzędu 4.

$$u_{k+1} = u_k + \frac{h}{2}[K_1 + K_2],$$

$$K_1 = f(t_k + (\frac{1}{2} + \frac{\sqrt{3}}{6})h, u_k + \frac{h}{4}K_1 + (\frac{1}{2} + \frac{\sqrt{3}}{6})hK_2), \tag{6.34}$$

$$K_2 = f(t_k + (\frac{1}{2} - \frac{\sqrt{3}}{6})h, u_k + (\frac{1}{2} - \frac{\sqrt{3}}{6})K_1 + \frac{h}{4}K_2).$$

Udowodniono, że można zbudować schematy otwarte Runge - Kutty, dla których liczba poziomów oraz rząd spełniają następujące zależności

liczba poziomów	**rząd**
1	1
2	2
3	3
4	4
5	4
6	5
7	6
8	6
9	7
$q \leq 10$	$q - 2$

Z tej tabelki widać, że schematy otwarte 4 poziomowe są optymalne w tym sensie, że osiągają maksymalny rząd przy minimalnej liczbie poziomów. Dla schematów zamkniętych q - poziomowych, można zawsze osiągnąć rząd $2q$.

Koszt schematu determinowany jest liczbą obliczeń wartości *prawej strony równania* f na każdym kroku całkowania. Zatem widzimy, że schematy Runge Kutty są wygodne w stosowaniu (schematy otwarte), ale raczej kosztowne.

Zadanie. Wyprowadź wzory dla dwupoziomowego schematu Runge - Kutty, otwartego.

$$u_{k+1} = u_k + h[c_1K_1 + c_2K_2].$$

Ile takich schematów rzędu 2 można zbudować?

Dotychczas, mówiąc o schematach różnicowych, podawaliśmy jako istotną ich cechę **rząd**. Pamiętamy jednak, że najważniejszą cechą schematu jest jego **zbieżność**. Jakie znaczenie dla funkcjonowania schematu ma jego rząd wyjaśnia **teoria zbieżności schematów różnicowych.** Podstawowe fakty z tej teorii, dla przypadku **schematów jednokrokowych** przytoczymy poniżej.

Teoria zbieżności schematów jednokrokowych.

Nierówność Gronwall'a. *Niech ciąg liczb nieujemnych* $\{v_k\}\quad k = 0, 1, \cdots,$ *spełnia nierówność*

$$0 \le v_{k+1} \le Av_k + B, \quad k = 0, 1, \cdots$$

gdzie $A, B > 0$*, to wtedy dla każdego* $k = 0, 1, \cdots$

$$0 \le v_k \le A^k v_0 + \begin{cases} \frac{A^k-1}{A-1}B & gdy\ \ A \ne 1 \\ kB & gdy\ \ A = 1. \end{cases} \tag{6.35}$$

Zadanie. Udowodnij nierówność Gronwall'a. Wskazówka: zastosuj indukcję względem k.

Teraz będziemy rozważać schematy jednokrokowe otwarte [17] postaci

$$u_{k+1} = u_k + h\Phi(h, t_k, u_k), \quad h > 0 \tag{6.35}$$

Zapis ten obejmuje wszystkie rozważane przez nas schematy jednokrokowe otwarte.

[17] Schemat zamknięty, jesli jest stosowalny, musi dać się *rozwikłać* przynajmiej lokalnie. Otrzymamy wtedy jego *lokalny odpowiednik otwarty.*

Konsystentność. *Mówimy, że schemat* (6.35) *jest* **konsystentny**, *jeśli*

- *funkcja* Φ *jest ciągła (względem wszystkich swoich argumentów) w całej swojej dziedzinie,*
- Φ *spełnia warunek Lipschitza względem zmiennej* u*:*

  istnieje stała L*, taka że dla wszystkich* (h, t, u_1), (h, t, u_2) *z dziedziny* Φ

  $$|\Phi(h, t, u_1) - \Phi(h, t, u_2)| \leq L|u_1 - u_2|,$$

  gdzie $|.|$ *oznacza normę w* $\mathbf{R^m}$,
- $\phi(0, t, u) = f(t, u)$, *gdzie rozpatrywane przez nas równanie ma postać*

  $$\frac{d}{dt}u(t) = f(t, u(t)).$$

Rozpatrujemy zagadnienie Cauchy'ego

$$\frac{d}{dt}u(t) = f(t, u(t)),$$

$$u(t_0) = u^0,$$

oraz schemat jednokrokowy dla tego zagadnienia:

$$u_{k+1} = u_k + h\Phi(h, t_k, u_k), \quad u_0 = u^0.$$

Twierdzenie o zbieżności z rzędem schematu jednokrokowego. *Jeśli rozwiązanie* u *zagadnienia Cauchy'ego jest klasy* C^{p+1}, $p > 0$ *w przedziale* $[t_0, t_0 + \alpha]$ $\alpha > 0$, *w którym jest określone, i schemat jest* **konsystentny** *oraz* **rzędu** p, *to schemat* **jest zbieżny** *i ponadto dla każdego ustalonego* $t = t_k \in [t_0, t_0 + \alpha]$

$$|u(t_k) - u_k| \leq Kh^p, \quad gdy \;\; h \to 0, \quad (h = \frac{t_k - t_0}{k}, \quad k \to \infty)$$

gdzie K *jest stałą niezależną od* h.

Dowód. Podstawiając rozwiązanie zagadnienia Cauchy'ego u do schematu różnicowego otrzymamy

$$u(t_{k+1}) = u(t_k) + h\Phi(h, t_k, u(t_k)) + r_k,$$

$$u_{k+1} = u_k + h\Phi(h, t_k, u_k).$$

Odejmując, otrzymamy

$$e_{k+1} = u(t_{k+1}) - u_{k+1} = e_k + h[\Phi(h, t_k, u(t_k)) - \Phi(h, t_k, u_k)] + r_k.$$

Ze względu na rząd schematu

$$|r_k| \leq Kh^{p+1}.$$

Ze względu na warunek Lipschitza otrzymamy:

$$|e_{k+1}| \leq (1 + hL)|e_k| + Kh^{p+1}.$$

Zastosujmy teraz **Nierówność Gronwalla** dla $A = 1 + hL$ i $B = Kh^{p+1}$. Otrzymamy

$$|e_k| \leq (1 + hL)^k |e_0| + \begin{cases} \frac{(1+hL)^k - 1}{hL} Kh^{p+1} & \text{dla} \;\; L \neq 0, \\ kKh^{p+1} & \text{dla} \;\; L = 0. \end{cases}$$

Ale $1 + hL \leq e^{hL}$, i stąd $(1 + hL)^k \leq e^{khL} \leq e^{\alpha L}$ oraz $kKh^{p+1} = khKh^p \leq \alpha Kh^p$. Ponadto przyjęliśmy, że $u_0 = u^0$, więc $e_0 = 0$. Ostatecznie

$$|e_k| \leq \begin{cases} e^{\frac{\alpha L}{L}} Kh^p, & \text{gdy} \;\; L > 0, \\ \alpha Kh^p & \text{gdy} \;\; L = 0 \end{cases} = O(h^p).$$

□

Zadanie. Udowodnij, że z samego założenia **konsystentności** wynika już zbieżność schematu. Jednak nie otrzymujemy oszacowania błędu e_k.

Z udowodnionego twierdzenia widać, jaką rolę odgrywa rząd schematu: **jeśli rozwiązanie u, które aproksymujemy jest dostatecznie gładkie ($u \in C^{p+1}$), oraz jeśli rząd schematu jest równy p, to $|e_k| \leq Kh^p$, gdy $h \to 0$.**

Schematy wielokrokowe

Poznaliśmy już ogólną postać **liniowego schematu q - krokowego**

$$\sum_{j=0}^{q} \alpha_j u_{k+j} = h \sum_{j=0}^{q} \beta_j f_{k+j}. \tag{6.30}$$

Zadanie. Udowodnij, że schemat (6.30) jest rzędu p wtedy i tylko wtedy, gdy

$$c_j = 0, \quad j = 0, 1, 2, \cdots, p$$

$$c_{p+1} \neq 0. \tag{6.36}$$

gdzie

$$c_0 = \sum_{j=0}^{q} \alpha_j,$$

$$c_1 = \sum_{j=0}^{q} j\alpha_j - \sum_{j=0}^{q} \beta_j,$$

$$c_s = \frac{1}{s!} \sum_{j=0}^{q} j^s \alpha_j - \frac{1}{(s-1)!} \sum_{j=0}^{q} j^{s-1} \beta_j, \quad s = 2, 3 \cdots.$$

Wskazówka. Podstaw dostatecznie głdkie rozwiązanie u i rozwiń.

Komentarz. Z treści powyższego zadania wynika, że stwierdzenie jaki jest rząd schematu typu (6.30) jest czynnością czysto mechaniczną. Znając współczynniki α_j i β_j wyliczamy współczynniki c_s rozwinięcia Taylora reszty, aż do znalezienia pierwszego współczynnika niezerowego.

Ze schematem q-krokowym typu (6.30) można związać dwa wielomiany

$$\rho(\lambda) = \sum_{j=0}^{q} \alpha_j \lambda^j, \tag{6.37}$$

$$\sigma(\lambda) = \sum_{j=0}^{q} \beta_j \lambda^j. \tag{6.38}$$

Wielomian ρ odgrywa podstawową rolę w teorii zbieżności schematów wielokrokowych postaci (6.30).

Stabilność. *Schemat* (6.30) *jest* **stabilny**, *jeśli wszystkie pierwiastki wielomianu ρ leżą w kole $|z| \leq 1$ na płaszczyźnie zespolonej, zaś te które leżą na okręgu $|z| = 1$ są* **jednokrotne**.
Silna stabilność. *Schemat* (6.30) *jest* **silnie stabilny**, *jeśli jest* **stabilny** *i jeśli jedynym pierwiastkiem wielomianu ρ o module równym 1 jest 1.*

Ponieważ schematy q - krokowe potrzebują q warunków początkowych, definicja zbieżności podana uprzednio dla schematów jednokrokowych wymaga pewnego rozszerzenia.

Warunek zbieżności schematu. *Przypuśćmy, że dla dowolnego ustalonego $t \in [t_0, t_0 + a]$, siatka punktów $\{t_k\}$ została tak dobrana, że $t = t_k = t_0 + kh$.*
Będziemy uważać rozważany schemat za **zbieżny**, *jeśli warunek*

$$u_k \to u(t), \quad gdy \quad h \to 0 \quad (stąd \quad k = \frac{t - t_0}{h}, \quad k \to \infty). \tag{6.26}$$

zachodzi

- *dla dowolnego rozwiązania u dowolnego równania* (6.30) *z warunkiem początkowym $u(t_0) = u^0$, należącego do klasy równań spełniających założenia Twierdzenia Picard'a - Lindelöf'a,*
- *dla dowolnego rozwiązania $\{u_k\}, \quad k = 0, 1, \cdots$ rozważanego schematu, dla którego wartości startowe $u_j = u_j(h), \quad j = 0, 1, \cdots, q - 1$ spełniają warunek*
$$u_j \to u^0, \quad gdy \quad h \to 0, \quad j = 0, 1, \cdots, q - 1$$

Dla schematów typu (6.30) zachodzi następujące **twierdzenie o zbieżności**, które tu podajemy bez dowodu.

Twierdzenie o zbieżności.

1. *Jeśli schemat jest* **stabilny i ma rząd nie niższy niż 1**, *to jest zbieżny.*

2. *Jeśli rozwiązanie u zagadnienia różniczkowego jest klasy C^{p+1} dla $p > 1$ i schemat jest* **stabilny i rzędu** $p > 1$, *to jest zbieżny i zachodzi następujące oszacowanie szybkości zbieżności*

$$|e_k| = |u(t_k) - u_k| \leq Kh^p, \quad h \to 0,$$

gdzie K jest stałą niezależną od h.

Twierdzenie to mówi, że schematy dobre, to takie, które są **stabilne i rzędu przynajmniej 1.** Im wyższy rząd, tym zbieżność jest szybsza, ale pod warunkiem dostatecznej gładkości rozwiązania, które aproksymujemy.

Rola warunku silnej stabilności jest widoczna przy całkowaniu numerycznym zagadnienia Cauchy'ego **z ustalonym krokiem** $h > 0$, **przy** $k \to \infty$. Ta sprawa nie ma nic wspólnego ze zbieżnością schematu, **bo** h **jest ustalone!**

To co się może dziać, gdy użyjemy schematu stabilnego, ale nie silnie stabilnego ilustruje następujący przykład całkowania schematem "Midpoint"

$$u_{k+2} = u_k + hf_{k+1}.$$

Schemat ten jest rzedu 2 i jest stabilny, ale nie silnie stabilny, jest to zatem schemat zbieżny. Proponowane było poprzednio zadanie w którym całkowało sie tym schematem zagadnienie Cauchy'ego

$$\frac{d}{dt}u(t) = -\lambda u(t), \quad \lambda > 0,$$

$$u(0) = 1,$$

którego rozwiązaniem jest $u(t) = e^{-\lambda t}$.

Zadanie. Przeprowadź analizę tego co dzieje sie z rozwiązaniem równania różnicowego $u_{k+2} = u_k + hf_{k+1}$ dla $f(t, u) = -\lambda u, \ \ \lambda > 0$, gdy h jest ustalone, zaś $k \to \infty$.

Wskazówka. Zauważ, że otrzyma się równanie różnicowe liniowe o stałych współczynnikach, rzędu 2. Wypisz **wielomian charakterystyczny** i znajdź jego pierwiastki. Zauważ, że pierwiastki te są w przybliżeniu równe $e^{-\lambda h}$ i $-e^{\lambda h}$. Znajdź postać rozwiązania u_k w zależności od tych pierwiastków. Jedna ze składowych będzie sensownie przybliżać funkcję $e^{-\lambda t_k}$, zaś druga będzie generować *pasożytnicze oscylacje* rosnące wykładniczo wraz z k. Zjawisko to nie ma nic wspólnego ze zbieżnościa. Schemat jest zbieżny! Zauważ, że tego efektu nie byłoby, gdyby było $\lambda < 0$. Zauważ również, ze pasożytnicze oscylacje powstają jedynie z tego powodu, ze wielomian ρ ma pierwiastek -1.

Pozostaje nam jeszcze wyjaśnienie sprawy sensowności używania schematów zamkniętych. Tę kwestię najlepiej wyjaśnić w związku z tak zwaną *własnością sztywności* pewnych układów równań różniczkowych.

Weźmy pod uwagę **zagadnienie modelowe**; będzie to układ równań liniowych jednorodnych o stałych współczynnikach

$$\frac{d}{dt}u(t) = Au(t), \tag{6.37}$$

z warunkiem początkowym

$$u(0) = u^0, \tag{6.38}$$

gdzie A jest macierzą symetryczną wymiaru $m \times m$ o różnych wartościach własnych, przyczym wszystkie wartości własne mają **ujemne części rzeczywiste**. Ponadto wśród wartości własnych macierzy A są takie, które mają **duże** i **małe** moduły.

Zadanie modelowe (6.37), (6.38) jest wyidealizowanym **układem sztywnym**. Ze zjawiskiem sztywności możemy mieć do czynienia w przypadku zupełnie innych, nieliniowych równań różniczkowych, które *lokalnie mają cechy zbliżone do naszego zadania modelowego.*

Na podstawie tego, co już wiemy, potrafimy łatwo rozwiązać nasze zadanie modelowe. Ponieważ macierz A ma różne wartości własne zatem jest ona diagonalizowalna. Możemy więc znaleźć taka nieosobliwą macierz T, że $A = T\Lambda T^{-1}$, gdzie Λ jest macierzą diagonalną, mającą na diagonali wartosci własne $\lambda_1, \lambda_2, \cdots, \lambda_m$ macierzy A. Pomnóżmy lewostronnie równanie (6.37) i warunek (6.38) przez macierz T^{-1}, oznaczając jednocześnie $v(t) = T^{-1}u(t)$ i $v^0 = T^{-1}u^0$, gdzie $v(t) = [v_1(t), v_2(t), \cdots, v_m(t)]^T$. Dla funkcji $v_j, \quad j = 1, 2, \cdots, m$ otrzymamy **układ** m **niezależnych od siebie równań różniczkowych liniowych**

$$\frac{d}{dt}v_j(t) = \lambda_j v_j(t),$$

z warunkami początkowymi

$$v_j(0) = v_j^0,$$

dla $j = 1, 2, \cdots, m$. Mamy zatem

$$v_j(t) = e^{\lambda_j t} v_j^0 \quad j = 1, 2, \cdots, m.$$

Składowe v_j rozwiązania u które odpowiadają wartościom własnym **o dużych modułach** (części rzeczywiste są ujemne!) zanikają bardzo szybko i ich wpływ na rozwiązanie jest znikomy, natomiast charakter rozwiązania jest determinowany przez te składowe, które odpowiadają wartościom własnym o niewielkich modułach. Jednak te składowe szybkozanikające sprawiają kłopoty numeryczne - (wielkie stałe Lipschitza!), wymuszając, na przykład, stosowanie bardzo małych kroków całkowania. Do całkowania takich zadań potrzebujemy więc schematów **odpornych na takie trudności**. Miarą sztywności zadania modelowego jest **współczynnik sztywności**

$$\sigma(A) = \frac{\max_j |\lambda_j|}{\min_j |\lambda_j|}.$$

Sprobujemy odpowiedzieć na pytanie, jakie schematy typu (6.30) nadają sie do całkowania zagadnień o dużym współczynniku sztywności. W tym celu rozpatrzymy **skalarne zadanie modelowe**

$$\frac{d}{dt}u(t) = \lambda u(t), \quad u(0) = 1, \tag{6.39}$$

gdzie $\lambda \in \mathbf{C}$ jest liczbą zespoloną. Nas będą interesowały głównie wartości λ takie, że $\Re(\lambda) < 0$.

Jeśli do zadania (6.39) zastosujemy schemat (6.30) to otrzymamy równanie różnicowe liniowe o stałych współczynnikach

$$\sum_{j=0}^{q} \alpha_j u_{k+j} = h\lambda \sum_{j=0}^{q} \beta_j u_j,$$

którego wielomian charakterystyczny jest postaci

$$\pi(z, \bar{h}) = \rho(z) - \bar{h}\sigma(z), \tag{6.40}$$

gdzie $\bar{h} = \lambda h$, oraz jak poprzednio

$$\rho(z) = \sum_{j=0}^{q} \alpha_j z^j,$$

$$\sigma(z) = \sum_{j=0}^{q} \beta_j z^j.$$

Ponieważ dla $\Re(\lambda) < 0$ nasze zadanie modelowe (6.39) **ma jedynie rozwiązania ograniczone** rozsądne jest wymaganie od schematu różnicowego tego, aby **jego rozwiązania były również ograniczone, gdy $k \to \infty$.** Ponieważ rozwiązanie ogólne dla naszego schematu jest postaci

$$u_k = \sum_{s=1}^{m} C_s \zeta_s(\bar{h})^k,$$

gdzie $\zeta(\bar{h})_s$ są pierwiastkami wielomianu (6.40), zaś $C_s, \quad s = 1, 2, \cdots, m$ są dowolnymi stałymi, warunkiem koniecznym sensownego funkcjonowania schematu dla zadań sztywnych jest to aby $|\zeta(\bar{h})| \leq 1$ dla możliwie szerokiego zakresu liczb zespolonych $\bar{h}$ takich, że $\Re(\bar{h}) < 0$. Prowadzi to do pojęcia

Obszar stabilności absolutnej schematu (6.30). *Obszar stabilności absolutnej schematu* (6.30) *jest to zbiór* $\Omega(\pi)$ *wszystkich takich liczb zespolonych* $\bar{h}$*, dla których wszystkie pierwiastki* $\zeta(\bar{h})$ *wielomianu (6.40)* $\pi(z, \bar{h})$ *mają moduły nie większe od 1.*

Schematy idealne do całkowania zadań sztywnych, to takie, których obszar stabilności absolutnej zawiera całą półpłaszczyznę $\Re(z) \leq 0$, gdyż teoretycznie pozwalają one na całkowanie zagadnień o dowolnie dużym współczynniku sztywności $\sigma(\pi)$ przy użyciu dowolnego kroku h. Zatem ograniczeniem jest tylko dokładność. Takie schematy nazywają sie **A-stabilne**.

Znajdźmy obszary stabilności absolutnej dla kilku prostych schematów.

1. **Schemat otwarty Eulera.**

$$u_{k+1} = u_k + hf_k.$$

$$\pi(z, \bar{h}) = z - 1 - \bar{h}.$$

Stąd $\zeta(\bar{h}) = \bar{h} + 1$ i punkty $\bar{h}$ należące do $\Omega(\pi)$ spełniają nierówność

$$|\bar{h} + 1| \leq 1.$$

Jest to tarcza koła na płaszczyźnie zespolonej o środku w -1 i promieniu 1. Obszar jest bardzo mały. **Metoda nie nadaje sie do całkowania zadań sztywnych.**

2. **Schemat zamknięty Eulera.**

$$u_{k+1} = u_k + hf_{k+1},$$

$$\pi(z, \bar{h}) = z - 1 - \bar{h}z.$$

Stąd $\zeta(\bar{h}) = \frac{1}{1-\bar{h}}$. Zatem obszar stabilności absolutnej dla schematu zamkniętego Eulera to zbiór wszystkich takich $\bar{h}$, dla których zachodzi nierówność

$$|\bar{h} - 1| \geq 1.$$

Jest to **obszar zewnętrzny** w stosunku do tarczy koła o promieniu 1 i środku 1. **Obszar stabilności absolutnej** jest ogromny i zawiera całą półpłaszczyznę $\Re(z) \leq 1$. Schemat jest **A-stabilny**.

3. **Schemat trapezów.**

$$u_{k+1} = u_k + \frac{h}{2}(f_k + f_{k+1}),$$

$$\pi(z, \bar{h}) = z - 1 - \frac{\bar{h}}{2}(z + 1),$$

stąd $\zeta(\bar{h}) = \frac{1+\frac{\bar{h}}{2}}{1-\frac{\bar{h}}{2}}$. Niech $\bar{h} = a + ib$, a więc $|\zeta(\bar{h})|^2 = \frac{(2+a)^2+b^2}{(2-a)^2+b^2}$. Zatem warunek $|\zeta(\bar{h})| \leq 1$ zachodzi, gdy $a = \Re(\bar{h}) \leq 0$ Oznacza to, że

$$\Omega(\bar{h}) = \{z \in \mathbf{C} | \Re(z) \leq 0\}.$$

To znaczy, ze metoda trapezów jest **A-stabilna.**

Widzimy stąd, że schemat zamknięty Eulera jest znacznie lepszy od schematu Eulera otwartego, jeśli chodzi o zastosowanie do zadań sztywnych. Okazuje się, ze jest to ogólna reguła: wszystkie schematy zamknięte mają odszar stabilności absolutnej większy niż ich odpowiedniki otwarte. Jednak żaden ze schematów typu (6.30), za wyjątkiem schematów Eulera zamkniętego i schematu trapezów **nie jest A-stabilny**. Można pokazać, że wśród schematów A-stabilnych, schemat trapezów jest optymalny w tym sensie, że ma rząd 2 (najwyższy możliwy!) i ma najmniejszy możliwy wspólczynnik rozwinięcia reziduum c_3. [18]

---

[18] Patrz wzór (6.36).

Kilka uwag na koniec.

- Schematy wielokrokowe stosowane w trybie **PREDICTOR - CORRECTOR** przy małej liczbie iteracji są szybsze niż schematy typu Runge-Kutty. Schematy obu typów mogą mieć dowolnie wysoki rząd. Schematy typu Runge - Kutty mogą służyć do wyznaczania punktów startowych. Wadą schematów wielokrokowch w przedstawionej tu *prymitywnej* postaci jest trudność dokonania zmiany kroku *w biegu*. Istnieją jednak algorytmy opracowane na podstawie schematów wielokrokowych dla których sprawa zmiany kroku calkowania nie jest problemem (na przykład tak zawna *Metoda Geara*).
- Należy unikać stosowania schematów, które nie są **silnie stabilne**.

Dobre schematy do zadań nie sztywnych, to schematy Adamsa.

- Schemat otwarty **Adamsa - Bathforth'a** - może służyć jako **PREDICTOR**.

$$u_{k+q} = u_{k+q-1} = h \sum_{j=0}^{q-1} \beta_j f_{k+j}.$$

Współczynniki β_j

q/j	0	1	2	3	4	5	rząd
1	1	-	-	-	-	-	1
2	-1/2	3/2	-	-	-	-	2
3	5/12	-16/12	23/12	-	-	-	3
4	-9/24	37/24	-59/24	55/24	-	-	4
5	251/720	-1274/720	2616/720	-2774/720	1901/720	-	5
6	-425/1440	2627/1440	-6798/1440	9482/1440	-7673/1440	4227/1440	6

- Schemat zamknięty **Adamsa - Moultona** może służyć jako **CORRECTOR**. Należy w pary predictor - corrector łączyć schematy tego samego rzędu.

$$u_{k+q} = u_{k+q-1} + h \sum_{j=0}^{q} \beta_j f_{k+j}.$$

Współczynniki β_j

q/j	0	1	2	3	4	5	rząd
1	1/2	1/2	-	-	-	-	2
2	-1/12	8/12	5/12	-	-	-	3
3	1/24	-5/24	19/24	9/24	-	-	4
4	-19/720	106/720	-264/720	646/720	251/720	-	5
5	27/1440	-173/1440	482/1440	-798/1440	1427/1440	475/1440	6

Rozdział 7
O RÓWNANIACH RÓŻNICZKOWYCH O POCHODNYCH CZĄSTKOWYCH

Będziemy dalej używać terminu *równania różniczkowe cząstkowe* zamiast *równania różniczkowe o pochodnych cząstkowych*. Omówimy tu tylko dwa bardzo proste przykłady, pokazujące dwa najważniejsze typy zagadnień rozpatrywanych najczęściej dla równań różniczkowych cząstkowych

- **Zagadnienia Stacjonarne,**
- **Zagadnienia Ewolucyjne.**

Należy podkreślić, że teoria równań różniczkowych cząstkowych jest nieporównywalnie bardziej złażona niż teoria równań różniczkowych zwyczajnych. Rozpatrując równania zwyczajne, mieliśmy do czynienia tylko z operatorem różniczkowym jednego rodzaju

$$u \to \frac{du}{dt},$$

gdzie $u : [t_0, t_0 + a] \to \mathbf{R}^n$. Operatory różniczkowe typu cząstkowego, mogą być bardzo różnorodne. Oto bardzo typowe, proste przykłady

- $$\Delta u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2},$$

  gdzie $u : \Omega \to \mathbf{R}$, $\Omega \subset \mathbf{R}^2$. Operator Δ nazywa się *Laplasjanem.*

- $$Hu = \frac{\partial^2 u}{\partial x^2} - \frac{\partial^2 u}{\partial y^2},$$

  gdzie $u : \Omega \to \mathbf{R}$, $\Omega \subset \mathbf{R}^2$;

- $$\frac{\partial u}{\partial t} + \alpha \frac{\partial u}{\partial x},$$

  gdzie $u : [0, T] \times [a, b] \to \mathbf{R}$, $\alpha \in \mathbf{R}$.

Każdy z tych operatorów ma zupełnie inne własności! Oczywiście, możemy mieć do czynienia z o wiele bardziej skomplikowanymi operatorami różniczkowymi, operatorami zależnymi od większej liczby zmiennych i.t.p.

Często spotykane w różnego rodzaju zastosowaniach jest **zagadnienie brzegowe Dirichleta dla równania Poissona.** Jest to typowe *zagadnienie stacjonarne*. Niech $\Omega \subset \mathbf{R}^2$. Poszukujemy funkcji $u : \bar{\Omega} \to \mathbf{R}$, ciągłej na domknięciu $\bar{\Omega}$ zbioru otwartego Ω, takiej że

$$-\Delta u(p) = f(p), \quad p = (x, y) \in \Omega, \tag{7.1}$$

$$u(p) = \phi(p), \quad p \in \partial\Omega. \tag{7.2}$$

Funkcja $f : \Omega \to \mathbf{R}$, jest *prawą stroną* równania Poissona, zaś $\phi : \partial\Omega \to \mathbf{R}$, *prawą stroną* warunku brzegowego Dirichleta, postawionego na brzegu $\partial\Omega$ obszaru Ω. Funkcje te, oraz obszar Ω określają nasze zagadnienie. Nie mamy tu do czynienia z zaleznością poszukiwanej funkcji u od czasu, przedstawianego zwykle zmienną niezależną t. Mówimy, że nie ma tu *ewolucji rozwiązania w czasie - zagadnienie jest stacjonarne.* Trzeba podkreślić, że *kształt* obszaru Ω odgrywa bardzo ważną rolę w teorii i numeryce tego zagadnienia. Jeśli funkcje f i ϕ są dostatecznie regularne, to zagadnienie (7.1)(7.2) ma jednoznaczne rozwiązanie w obszarze wypukłym Ω o dostatecznie gładkim brzegu. Jeśli $\phi = 0$, to zagadnienie Dirichleta nazywa się *jednorodne*. Zagadnienie (7.1)(7.2) ma wiele interpretacji fizycznych. Jedną z nich (gdy $\phi = 0$, jest jest opis kształtu membrany umocowanej na brzegu $\partial\Omega$, na którą działa siła opisana funkcją f.

Bardzo typowym przykładem **zagadnienia ewolucyjnego** jest

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0, \quad t > 0, \quad x > 0, \quad c > 0, \tag{7.3}$$

$$u(0, x) = \phi(x) \;\; x \geq 0, \;\; \textit{warunek pocztkowy}, \tag{7.4}$$

$$u(t, 0) = \psi(t) \;\; t \geq 0, \;\; \textit{warunek brzegowy}. \tag{7.5}$$

Poszukujemy $u : (0, \infty) \times (0, \infty) \to \mathbf{R}$. Jest to *zagadnienie mieszane, początkowo - brzegowe*. Zmienna x, to *zmienna przestrzenna*. Zmienną t interpretujemy jako czas.

- **warunek początkowy** podaje wartość rozwiązania w chwili $t = 0$
- **warunek brzegowy** określa, co dzieje się z u w czasie t na osi $x = 0$

Dla równania (7.3) rozważa się również *zagadnienie początkowe - zagadnienie Cauchye'go*

$$\frac{\partial u}{\partial t} + c\frac{\partial u}{\partial x} = 0 \quad x \in \mathbf{R}, \quad t \geq 0 \tag{7.6}$$

$$u(0, x) = \phi(x), \quad x \in \mathbf{R}. \tag{7.7}$$

Zagadnienie (7.6)(7.7) łatwo jest rozwiązać, jeśli założyć, że funkcja ϕ jest różniczkowalna. Zauważmy bowiem, że

$$u(t, x) = \phi(x - ct). \tag{7.8}$$

Istotnie
$$u(0, x) = \phi(x),$$
zaś
$$u_t = -\phi^{'}(x - ct)c,$$
$$u_x = \phi^{'}(x - ct),$$
skąd
$$u_t + cu_x = -c\phi^{'}(x - ct) + c\phi^{'}(x - ct) = 0.$$

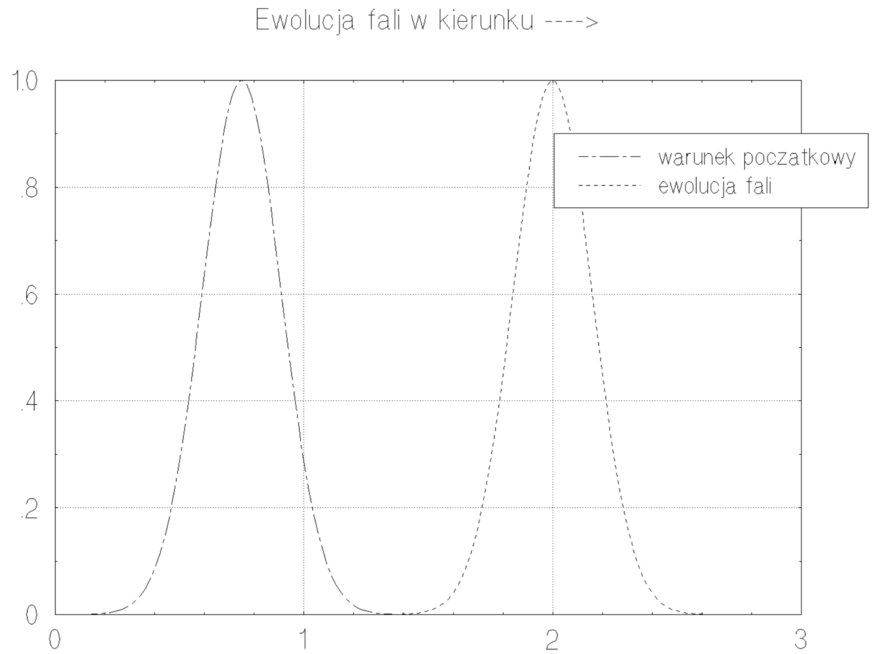
Dla zagadnienia mieszanego (7.3)-(7.5) można także napisać wzór na rozwiązanie

$$u(t, x) = \begin{cases} \phi(x - ct) & \text{dla } x - ct \geq 0, \\ \psi(t - \frac{x}{c}) & \text{dla } x - ct \leq 0. \end{cases} \tag{7.9}$$

Aby wzór (7.9) określał rozwiązanie, powinny zachodzić równości

$$\phi(0) = \psi(0), \quad \phi_+^{'}(0) = \frac{1}{c}\psi_+^{'}(0)$$

zapewniające ciągłość rozwiązania wraz z pierwszymi pochodnymi. W każdym razie z powyższych rozważań wynika, że *jeśli funkcje ϕ i ψ są ograniczone, to i rozwiązanie u też jest ograniczone.*



Rozwiązanie u równania (7.3) można interpretować jako bardzo prymitywną ewolucję fali w czasie. Kształt fali określa funkcja warunku początkowego ϕ. Ewolucja, w tym przypadku polega na *przesuwaniu* niezmienionej fali wzdłuż osi x.

Zadanie 7.1 Przeprowadź analizę tego co dzieje się z rozwiązaniem zagadnienia początkowego i początkowo - brzegowego dla $c > 0$ i dla $c < 0$. Jaka jest prędkość i kierunek przesuwania fali?

Przyjmijmy

(7.9) $$\phi(x) = e^{i\alpha x} = \cos\alpha x + i\sin\alpha x,$$

gdzie $\alpha \in \mathbf{R}$. Wybierając we właściwy sposób wartości α_j i kładąc $\phi_j(x) = e^{i\alpha_j x}$, możemy zapisać *szereg Fouriera* jako $\sum_{j=-\infty}^{\infty} a_j\phi_j(x)$ $a_j \in \mathbf{C}$. Stąd wynika, że przy pomocy *kombinacji liniowych* funkcji ϕ_j można aproksymować bardzo szeroką klasę funkcji, które chcielibyśmy przyjmować jako warunki początkowe dla naszego równania. Ma więc sens rozważanie rozwiązań równania (7.3) następującej postaci

$$u(t,x) = e^{i\alpha(x-ct)} = e^{-i\alpha ct}e^{i\alpha x}. \tag{7.10}$$

Rozwiązanie (7.10) ma postać *rozdzielonych zmiennych - to znaczy iloczynu funkcji zaleznej tylko od t i funkcji zależnej tylko od x.*

Prosty podręcznik teorii równań różniczkowych cząstkowych - patrz [12].

O METODACH NUMERYCZNYCH

Zajmiemy się tu tylko *metodami różnicowymi* rozwiązywania przybliżonego równań cząstkowych, dla dwóch opisanych tu przykładów: zadania stacjonarnego i zadania ewolucyjnego. Nie będzie to szczegółowa analiza problemu. Naszym celem jest wskazanie pewnych istotnych cech zagadnienia. Należy tu wspomnieć, że bardzo ważną rolę w tej dziedzinie numeryki odgrywają również inne typy metod numerycznych, wśrod których należy na pierwszym miejscu wymienić *metody elementu skończonego* (patrz na przykład [10], [11]).

Najpierw zajmiemy się krótko **zagadnieniem stacjonarnym** (7.1) (7.2). Niech obszar Ω będzie prostokątem

$$\Omega = [0,a] \times [0,b].$$

Na prostokącie Ω zbudujemy *siatkę punktów*

$$x_k = kh_1, \quad y_j = jh_2, \quad k = 0,1,\cdots,N, \quad j = 0,1,\cdots,M, \quad h_1 = \frac{a}{N}, \quad h_2 = \frac{b}{M}.$$

Metody różnicowe polegają na konstrukcji *równań różnicowych - schematów różnicowych*, których rozwiązania aproksymują poszukiwane przez nas rozwiązania równań różniczkowych, gdy $h \to 0$, gdzie $h = \max\{h_1, h_2\}$. Jest wiele możliwości konstrukcji takich równań dla zagadnienia (7.1) (7.2), nie wszystkie jednak *muszą* mieć wymagane własności aproksymacyjne. Okazuje się, że dobrą metodę różnicową otrzmamy, na przykład, zastępując pochodne w równaniu (7.1) różnicami dzielonymi

$$-\frac{u_{k-1,j} - 2u_{k,j} + u_{k+1,j}}{h_1^2} - \frac{u_{k,j-1} - 2u_{k,j} + u_{k,j+1}}{h_2^2} = f_{k,j} = f(x_k, y_j) \tag{7.11}$$

dla $0 < k < N, \quad 0 < j < M$, zaś

$$u_{0,j} = \phi_{0,j} = \phi(0, y_j), \quad u_{N,j} = \phi_{N,j} = \phi(a, y_j),$$

$$u_{k,0} = \phi_{k,0} = \phi(x_k, 0), \quad u_{k,M} = \phi_{k,M} = \phi(x_k, b). \tag{7.12}$$

Tutaj $u_{k,j}$ oznacza wartość *funkcji siatkowej w węźle siatki* (x_k, y_j). Funkcja ta jest rozwiązaniem układu równań (7.11)(7.12), i a priori, nic nie można powiedzieć o związku $u(x_k, y_j)$ oraz $u_{k,j}$. Zwróćmy uwagę na to, ze chodzi tu o porównanie funkcji działających w zupełnie innych przestrzeniach. Dowodzi się (patrz na przykład [10], [11]), że istotnie, rozwiązanie równań (7.11)(7.12) mają wymagane własności aproksymacyjne.

Przyjrzyjmy się bliżej równaniom (7.11)(7.12). Jeśli utworzymy wektor

$$\underline{u} = [u_{1,1}, u_{1,2}, \cdots, u_{N-1,M-1}]^T,$$

to łatwo zauważymy, że równania te dadzą się zapisać jako układ równań liniowych algebraicznych

$$A\underline{u} = \underline{g}, \tag{7.13}$$

gdzie macierz A jest *pięcio - diagonalna* wymiaru $(N-1)(M-1) \times (N-1)(M-1)$, zaś składowe wektora $\underline{g}$, wyrażają się poprzez wartości funkcji f i ϕ w punktach siatki. Układ ten służy do obliczania przybliżonego rozwiązania naszego problemu różniczkowego.

Układ (7.13) *jest źle uwarunkowany.* Jego współczynnik uwarunkowania $cond(A)$ jest rzędu $\max\{\frac{1}{h_1^2}, \frac{1}{h_2^2}\}$ i uwarunkowanie układu pogarsza się wraz z zagęszczaniem siatki - to jest wraz z *polepszaniem aproksymacji.* W przypadku, gdy *siatka jest kwadratowa* to znaczy, gdy $h_1 = h_2$, macierz A jest symetryczna i dodatnio określona. Dobrze więc tu stosować metody CGMR lub CGME z odpowiednim *preconditingiem.*

Zadanie 7.2
Dla kwadratu $\Omega = [0, a] \times [0, a]$, oraz dla siatki kwadratowej, gdy $N = M = 10$ rozpisz macierz A układu (7.13). Przyjrzyj się strukturze macierzy w zależności od uporządkowania punktów siatki.

Zajmiemy się teraz **zagadnieniem ewolucyjnym**. Zbudujemy siatkę o stałych krokach h i τ w kierunku osi x i osi t odpowiednio. Oznaczmy rozwiązanie równania różnicowego w punkcie siatki $x_k = kh$, $t_n = \tau n$ przez u_k^n. Pochodne zastąpimy przez różnice dzielone

$$u_t(t, x) \to \frac{u(t+\tau, x) - u(t, x)}{\tau},$$

$$u_x(t,x) \to \frac{u(t,x+h)-u(t,x)}{h}.$$

Niech $\lambda = \frac{\tau}{h}$. Ze względu na kierunek ruchu fali, narzuca sie następujący sposób konstrukcji schematu różniowego

$$u_{k+1}^{n+1} - u_{k+1}^{n} + \lambda c(u_{k+1}^{n} - u_{k}^{n}) = 0, \quad c > 0, \tag{7.14}$$

lub

$$u_{k+1}^{n+1} - u_{k+1}^{n} + \lambda c(u_{k+1}^{n+1} - u_{k}^{n+1}) = 0, \quad c > 0. \tag{7.15}$$

Są to tak zwane schematy *upwind*. Pierwszy ze schematów jest *otwarty*, drugi *zamknięty* (patrz rozdział o równaniach zwyczajnych). Zatem schemat (7.15) wymaga rozwiązywania układu równań liniowych na każdym *kroku czasowym*. Zauważmy, że oba schematy nadają się do rozwiązywania zagadnienia brzegowego (7.1) - (7.3). Natomiast schematem otwartym (7.14) można rozwiązywać tylko zagadnienie początkowe (dla czego?). Oto *stencil* tych schematów.

Dla schematu otwartego:

$$\begin{array}{ccccc}
 & . & . & . & . \\
n+1 & . & . & * & . \\
n & . & * & * & . \\
 & . & . & . & . \\
 & & k & k+1 &
\end{array}$$

Dla schematu zamniętego:

$$\begin{array}{ccccc}
 & . & . & . & . \\
n+1 & . & * & * & . \\
n & . & . & * & . \\
 & . & . & . & . \\
 & & k & k+1 &
\end{array}$$

Kształt przypominający żagiel jaki ma *stencil* schematu otwartego uzasadnia nazwę schematów *upwind*.

Sprobujmy przeprowadzić nieco dokładniejszą analizę tych dwóch schematów. Posłużymy sie w tym celu *metodą Fouriera*. Przez analogię ze wzorem (7.10) możemy sprobować poszukiwać rozwiązań równań (7.14) i (7.15) w postaci

$$u_k^n = \gamma^n e^{i\alpha k}, \tag{7.16}$$

gdzie $\gamma \in \mathbf{C}$, zaś $\alpha \in \mathbf{R}$. Dla dowolnego $\alpha \in \mathbf{R}$, będziemy starali się wyznaczyć $\gamma(\alpha)$, tak aby ciąg $\{u_k^n\}$ spełniał, dla każdego n i k odpowiednie równanie (7.14) lub (7.15). Zauważmy, że $u_k^n = \gamma^n e^{i\alpha k}$ spełnia **ograniczony** warunek początkowy $u_k^0 = e^{i\alpha k}, \quad \alpha \in \mathbf{R}$. Biorąc pod uwagę opisane wyżej własności rozwiązań rozważanych równań widzimy, że jeśli wykażemy, że wzór (7.16) określa rzeczywiście rozwiązanie schematów (7.14) i (7.15), to warunkiem koniecznym *dobroci* naszych schematów będzie to, że $|\gamma(\alpha)|^n$ nie rośnie do nieskończoności, gdy $n \to \infty$, gdyż *ograniczone rozwiązanie równania różniczkowego nie może być poprawnie aproksymowane funkcją nieograniczoną*! Powyższy warunek jest spełniony, gdy

$$|\gamma(\alpha)| \leq 1, \tag{7.17}$$

Można wykazać,że (7.17) jest warunkiem dostatecznym *stabilności* rozważanych schematów. Stąd zaś wynika *zbieżność*. Można o tym przeczytać w [13].

Zbadamy teraz (metodą Fouriera) stabilność schematów (7.14) i (7.15). Podstawiając najpierw wzór (7.16) do schematu (7.14), po łatwych rachunkach otrzymamy warunek dla $\gamma(\alpha)$

$$\gamma(\alpha) = 1 - \lambda c + \lambda c e^{-i\alpha},$$

i stąd

$$|\gamma(\alpha)|^2 = 1 - 2\lambda c(1 - \lambda c)(1 - \cos\alpha).$$

Widać stąd, że jeśli $1 - \lambda c \leq 0$, lub inaczej, jeśli

$$\frac{\tau}{h} = \lambda \leq \frac{1}{c}, \tag{7.18}$$

to $|\gamma(\alpha)|^2 \leq 1, \quad$ dla każdego $\alpha \in \mathbf{R}$

Schemat otwarty (7.14) jest zatem *stabilny - a więc zbieżny* jeśli kroki siatki spełniają następującą nierówność

$$\tau \leq \frac{h}{c}, \quad c > 0. \tag{7.19}$$

Mówimy, że schemat otwarty (7.14) jest *warunkowo stabilny.*

Zbadamy jeszcze schemat (7.15)

$$u_{k+1}^{n+1} - u_{k+1}^n + \lambda c(u_{k+1}^{n+1} - u_k^{n+1}) = 0,$$

Podstawiając $u_k^n = \gamma^n e^{i\alpha k}$, otrzymamy

$$\gamma(\alpha) = \frac{1}{1 + \lambda c - \lambda c e^{-i\alpha}}$$

stąd

$$|\gamma(\alpha)|^2 = \frac{1}{(1+\lambda c)^2 + \lambda^2 c^2 - 2\lambda c(1+\lambda c)\cos\alpha},$$

Zatem warunkiem stabilności jest

$$1 + 2\lambda c(1+\lambda c)(1-\cos\alpha) \geq 1.$$

Ze względu na to, że $1-\cos\alpha \geq 0$ i że $\lambda c \geq 0$ warunek stabilności jest zawsze spełniony. Mówimy więc, że schemat zamknięty (7.15) jest *bezwarunkowo stabilny*. A więc łatwy do stosowania schemat otwarty wymaga, aby kroki siatki spełniały warunek $\tau \leq \frac{h}{c}$. Trudniejszy do stosowania schemat zamknięty nie wymaga żadnych dodatkowych warunków - *jest zawsze stabilny*.

Zadanie 7.3

1. Zbadaj przy pomocy metody Fouriera schemat z parametrem $0 \leq \kappa \leq 1$

$$u_{k+1}^{n+1} - u_{k+1}^n + \kappa\lambda c(u_{k+1}^n - u_k^n) + (1-\kappa)\lambda c(u_{k+1}^{n+1} - u_k^{n+1}) = 0,$$

   gdy $c > 0$. Co zrobić, jeśli $c < 0$?

2. Zbadaj przy pomocy metody Fouriera schematy

   (a)
$$u_k^{n+1} - \frac{1}{2}(u_{k-1}^n + u_{k+1}^n) + \lambda\frac{c}{2}(u_{k+1}^n - u_{k-1}^n) = 0,$$
   dla $c \leq 0$ i $c \geq 0$.

   (b)
$$u_k^{n+1} - u_k^n + \lambda\frac{c}{2}(u_{k+1}^n - u_{k-1}^n) = 0,$$
   również dla $c \leq 0$ i $c \geq 0$.

ZALECANA LITERATURA ZWIĄZANA Z TEMATEM SKRYPTU

1. **P.M. Prenter** "Splines and Variational Methods"

2. **Gantmacher** "Matrix theory" (Oryginał rosyjski)

3. **S. Paszkowski** "Zastosowania numeryczne wielomianów i szeregów Czebyszewa"

4. **V.I. Lebedev, S.A. Finogenov** "O probleme vybora iteracionnych parametrov ..." Žurnal vyčislitelnoi matematiki i matematiczeskoi fiziki T11 Nr 2 1971

5. **G.H. Golub and C.F. van Loan** "Matrix computations"

6. **A.Kiełbasinski H.Schwetlick** "Numeryczna algebra liniowa"

7. **NBS 10.11.1954** "Tables of functions and zeros of functions"

8. **N.S. Bahvalov** "Čislennye Metody" Tom I. Nauka Moskva 1973

9. **A.Palczewski** "Równania różniczkowe zwyczajne" WNT 1999

10. **J.Jankowska, M.Jankowski, M.Dryja** "Przegląd metod i algorytmów numerycznycznych" T.1 i 2

11. **P.G. Ciarlet** "The finite element methods for elliptic problems" North Holland

12. **Fitz John** "Partial differential equations" Springer Verlag

13. **G.A. Sod** "Numerical methods in fluid dynamics" Cambridge Univ. Press