Matematyka obliczeniowa, II rok Matematyki (2020/2021)

Metody numeryczne, III rok Informatyki, (2013/2014)

1. Rozwiązywanie równań nieliniowych
2. Arytmetyka zmiennopozycyjna
3. Błędy w obliczeniach. Uwarunkowanie zadania. Numeryczna poprawność i stabilność algorytmu
4. Rozwiązywanie układów równań liniowych.
5. Liniowe zadania najmniejszych kwadratów
6. Algebraiczne zagadnienie własne
7. Interpolacja wielomianowa
8. Interpolacja funkcjami sklejanymi
9. Interpolacja trygonometryczna. Algorytm FFT
10. Aproksymacja funkcji
11. Numeryczne obliczanie całek
12. Wybrane środowiska i biblioteki dla obliczeń numerycznych

Zasady zaliczania przedmiotu Matematyka Obliczeniowa

Na zaliczenie przedmiotu składają się: zaliczenie ćwiczeń (w tym laboratorium i kolokwium) i zdanie egzaminu. Na końcową ocenę składają się

- punkty, którymi prowadzący ćwiczenia ocenił prace domowe, tj. rozwiązania zadań na kartce,
- punkty za rozwiązania zadań z komputerem,
- punkty z kolokwium,
- punkty zdobyte na egzaminie pisemnym.

Przed przystąpieniem do egzaminu należy zaliczyć ćwiczenia na co najmniej 50% punktów. Propozycje ocen będą złożone po egzaminie pisemnym na podstawie sumy ważonej zdobytych punktów, w której zadania domowe, zadania komputerowe, kolokwium i egzamin pisemny mają udziały odpowiednio 20%, 10%, 20% i 50%, przy czym na ocenę dostateczną na egzaminie pisemnym też trzeba zdobyć co najmniej 50% punktów. Wynik między 33% i 50% punktów z egzaminu daje szansę otrzymania oceny dostatecznej na egzaminie ustnym. Poza tym otrzymaną propozycję oceny co najmniej dostatecznej można przyjąć lub próbować zmienić na egzaminie ustnym.

Zasady zaliczania przedmiotu Metody Numeryczne

Na zaliczenie przedmiotu składają się: zaliczenie ćwiczeń i zdanie egzaminu. Połowa ćwiczeń ma miejsce w laboratorium, pozostałe ćwiczenia są w sali przy tablicy. Na końcową ocenę składają się

- punkty, którymi prowadzący ćwiczenia ocenił prace domowe, tj. rozwiązania zadań na kartce,
- punkty za rozwiązania zadań programistycznych,
- punkty zdobyte na egzaminie pisemnym.

Po egzaminie pisemnym będą wystawione propozycje ocen, w których zadania domowe, zadania programistyczne i egzamin pisemny mają udziały odpowiednio 25%, 25% i 50%, przy czym z każdego z tych elementów trzeba zdobyć co najmniej 25% punktów, a w sumie co najmniej 50%. Otrzymaną propozycję oceny uczestnik zajęć może przyjąć, lub wystawić na ryzyko zmiany na egzaminie ustnym.

Wyjaśnienie

Różnice między przedmiotami Matematyka Obliczeniowa (dla kierunku Matematyka) i Metody Numeryczne (dla kierunku Informatyka) są konsekwencją innej wiedzy początkowej i innej roli tych przedmiotów w dalszych studiach; niemniej, 90% materiału przedmioty te mają wspólne, dlatego zdecydowałem się napisać jeden skrypt.

Zakładana wiedza początkowa jest na obu kierunkach inna głównie z uwagi na znacznie węższy program (i mniejszą liczbę godzin) Algebry Liniowej na Informatyce — program ten nie obejmuje m.in. algebraicznego zagadnienia własnego, który to brak trzeba (w ramach tego przedmiotu) uzupełnić.

Program Metod Numerycznych obejmuje iteracyjne metody rozwiązywania (wielkich) układów równań liniowych. Większość studentów Informatyki w trakcie dalszych studiów się z nimi nie zetknie, ale informatycy w pracy zawodowej mogą dostawać zamówienia na implementacje. Z kolei studenci Matematyki w trakcie dalszych studiów będą mieć okazje do *znacznie dokładniejszego* zgłębienia tych metod, w ramach przedmiotów takich jak Matematyka Obliczeniowa II, Obliczenia Naukowe i Numeryczne Równania Różniczkowe. To dlatego metody iteracyjne rozwiązywania układów równań liniowych są w wykładzie dla nich pominięte.

Program Metod Numerycznych jest zatem nieco szerszy, natomiast na Matematyce Obliczeniowej więcej uwagi jest poświęcone analizie metod i dowodom stosownych twierdzeń. Treść wykładów z obu przedmiotów jest na slajdach (innych dla każdego przedmiotu), dzięki czemu można się zorientować *co kogo obowiązuje*. Ale *wszystkich* studentów (z obu kierunków) zachęcam do przestudiowania *całości* skryptu.

Literatura

- Kincaid D., Cheney W.: *Analiza numeryczna*, WNT, Warszawa, 2006.
- Krzyżanowski P.: *Obliczenia inżynierskie i naukowe*, PWN, Warszawa, 2011.
- Jankowska J., Jankowski M., Dryja M: *Przegląd metod i algorytmów numerycznych* cz. 1 i 2, WNT, Warszawa, 1988.
- Dahlquist G., Björck Å: *Metody numeryczne*, PWN, Warszawa, 1983.

Egzamin z Metod Numerycznych, III rok Inf.

(*Ściśle tajne* przed godz. 14:30 28 stycznia 2012.)

Proszę bardzo uważnie przeczytać treść zadań. Na ocenę bardzo duży wpływ będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Wykonaj dwie iteracje metody Newtona dla układu równań

$$\begin{cases} x^3 + x - xy - 2y^2 = -4, \\ \qquad\qquad y^3 - y = -6, \end{cases}$$

dla punktu startowego $(x_0, y_0) = (2, -2)$.

2. Wartość wyrażenia $w = a^3 - b^3$ została obliczona przy użyciu następującego algorytmu, zrealizowanego za pomocą arytmetyki zmiennopozycyjnej:

```
x1 = a*a+b*b;
x2 = a+b;
x3 = 0.5*(x1+x2*x2);
w = x3*(a-b);
```

Napisz wyrażenie, którego wartością jest błąd (bezwzględny) otrzymanego wyniku, jeśli w żadnym z działań nie wystąpił nadmiar ani niedomiar.

3. Wartości $f_1, \ldots, f_N$ pewnej funkcji rzeczywistej f są podane w punktach $x_1, \ldots, x_N$. Funkcja ta ma być przybliżona przez wielomian w stopnia co najwyżej $n < N$ tak, aby wyrażenie $\sum_{i=1}^{N}(f_i - w(x_i))^2$ było jak najmniejsze. Napisz układ równań liniowych, taki że rozwiązanie powyższego zadania aproksymacji można sprowadzić do liniowego zadania najmniejszych kwadratów dla tego układu. Podaj algorytm rozwiązywania tego zadania za pomocą odbić Householdera. Jaki jest koszt tego algorytmu w zależności od liczb n i N?

4. Skonstruuj odpowiednią bazę Newtona i rozwiąż przy użyciu algorytmu różnic dzielonych zadanie interpolacyjne Hermite'a dla danych przedstawionych w tabelce:

x_i	1	3
$f(x_i)$	-4	-8
$f'(x_i)$	-10	30
$f''(x_i)$	-8	

5. Rozważamy konstrukcję interpolacyjnej funkcji sklejanej drugiego stopnia, $s(x) = \sum_{i=0}^{N-3} d_i N_i^2(x)$, której węzły są liczbami naturalnymi, $u_i = i$ dla $i = 0, \ldots, N$, reprezentowanej za pomocą funkcji B-sklejanych N_i^2. Warunki interpolacyjne (tj. wartości funkcji, $s_k = s(v_k)$) są zadane w punktach $v_0 = 2$, $v_k = k + 1\frac{1}{2}$ dla $k = 1, \ldots, N-4$ i $v_{N-3} = N-2$. Wiedząc, że $N_i^2(i+x) = N_0^2(x)$ dla każdego $x \in \mathbb{R}$ oraz $i \in \{0, \ldots, N-3\}$, a ponadto $N_0^2(x) = 0$ jeśli $x \leqslant 0$ lub $x \geqslant 3$, oraz $N_0^2(\frac{1}{2}) = N_0^2(2\frac{1}{2}) = \frac{1}{8}$, $N_0^2(1) = N_0^2(2) = \frac{1}{2}$ i $N_0^2(1\frac{1}{2}) = \frac{3}{4}$, napisz układ równań, którego rozwiązanie jest wektorem współczynników d_i poszukiwanej funkcji.

6. Które z podanych na wykładzie metod rozwiązywania układów równań liniowych mogą być użyte do rozwiązania układu równań liniowych:

   a) Z poprzedniego zadania.

   b) Układu równań normalnych dla regularnego liniowego zadania najmniejszych kwadratów z liczbą niewiadomych nie przekraczającą 100.

   c) Układu równań z wielką macierzą ($n \times n$, gdzie $n > 10^4$) symetryczną i dodatnio określoną, która ma w każdym wierszu mniej niż 20 niezerowych współczynników rozmieszczonych nieregularnie.

   W każdym przypadku napisz, z uzasadnieniem, która z tych metod wydaje się najbardziej odpowiednia.

7. Podaj najmniejsze n, takie że błąd aproksymacji jednostajnej funkcji $f(x) = \sin x$ w przedziale $[-4\pi, 4\pi]$ przez optymalnie dobrany wielomian stopnia n jest mniejszy niż 1. Odpowiedź uzasadnij, powołując się na stosowne twierdzenie.

8. Całkę

$$I(f) = \int_{-1}^{1} f(x)\,dx,$$

   chcemy przybliżać kwadraturą o postaci

$$Q(f) = A_0\bigl(f(-1) + f(1)\bigr) + A_1\bigl(f(-a) + f(a)\bigr).$$

   Dobierz liczbę a i współczynniki A_0, A_1 tak, aby otrzymać kwadraturę o największym rzędzie. Podaj oszacowanie błędu tej kwadratury, jeśli funkcja f ma w przedziale $[-1, 1]$ ciągłą pochodną czwartego rzędu i istnieje stała M_4, taka że dla każdego $x \in [-1, 1]$ zachodzi nierówność $|f^{(4)}(x)| \leqslant M_4$.

Egzamin poprawkowy z Metod Numerycznych, III rok Inf.

(*Ściśle tajne* przed godz. 15:15 1 marca 2012.)

Proszę bardzo uważnie przeczytać treść zadań. Na ocenę bardzo duży wpływ będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Metodą odbić Householdera rozwiąż liniowe zadanie najmniejszych kwadratów dla układu równań $A\mathbf{x} = \mathbf{b}$, gdzie

$$A = \begin{bmatrix} 1 & 1 \\ -1 & -1 \\ 1 & -1 \\ 1 & -1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 2 \\ 1 \\ 2 \end{bmatrix}.$$

2. Podaj odpowiednią bazę Newtona i znajdź metodą różnic dzielonych wielomian interpolacyjny dla danych w tabelce:

x_i	0	1
$f(x_i)$	0	0
$f'(x_i)$	1	0
$f''(x_i)$	6	

3. Znajdź liczbę a i współczynniki A_0, A_1, takie że kwadratura

$$Q(f) = A_0 f(-a) + A_1 f(0) + A_0 f(a),$$

przybliżająca całkę

$$I(f) = \int_{-1}^{1} f(x)dx$$

ma maksymalny rząd. Podaj oszacowanie błędu tej kwadratury, przy założeniu, że funkcja f ma ciągłą pochodną rzędu r, który jest rzędem tej kwadratury.

4. Znajdź wielomian h stopnia co najwyżej 1, który jest optymalnym rozwiązaniem zadania aproksymacji jednostajnej dla funkcji $f(x) = \sin x$ w przedziale $[0, \pi]$. Uzasadnij poprawność rozwiązania, powołując się na odpowiednie twierdzenie.

5. Oblicz wskaźniki uwarunkowania macierzy

$$A_1 = \begin{bmatrix} 1 & 0 & 0 \\ 10 & 1 & 0 \\ 100 & 0 & 1 \end{bmatrix}$$

w normach $\|\cdot\|_1$ i $\|\cdot\|_\infty$. Na jaką dokładność wyniku rozwiązywania układu równań liniowych z tą macierzą można liczyć, jeśli współczynniki wektora prawej strony są znane z błędem nie większym niż 0.01%?

6. Niech f oznacza funkcję wypukłą klasy C^2, która ma w przedziale $[a, b]$ miejsce zerowe α o krotności 2. Jak, mając do dyspozycji podprogram obliczania wartości funkcji i pochodnej, można znaleźć liczbę α i która z metod: Newtona, siecznych, czy bisekcji, jest do tego odpowiednia. Odpowiedź uzasadnij.

7. Macierz A jest symetryczna i ma wartości własne w przedziale $[1, 9]$ (w szczególności liczby 1 i 9 też są wartościami własnymi tej macierzy). Dobierz parametr τ tak, aby zbieżność metody Richardsona rozwiązywania układu równań $A\mathbf{x} + \mathbf{b}$:

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \tau(A\mathbf{x}_k - \mathbf{b})$$

była najszybsza.

8. Jaki jest koszt rozwiązywania metodą eliminacji Gaussa układów równań liniowych z nieosobliwą macierzą $n \times n$

   a) trójdiagonalną ($a_{ij} = 0$ dla $|i - j| > 1$),
   b) Hessenberga ($a_{ij} = 0$ dla $i - j > 1$),
   c) blokowo-diagonalną, zbudowaną z bloków $k \times k$, gdzie n/k jest liczbą naturalną,
   d) jak wyżej, blokowo-diagonalną, przy czym wszystkie bloki diagonalne są jednakowe.

Egzamin z Metod Numerycznych, III rok Inf.

(*Ściśle tajne* przed godz. 14:30 2 lutego 2013.)

Proszę bardzo uważnie przeczytać treść zadań. Na ocenę bardzo duży wpływ będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Wykonaj dwie iteracje metody Newtona dla układu równań
$$\begin{cases} x^3 - y^2 = 0, \\ y^3 + yz = 0, \\ z^2 + z = 2 \end{cases}$$
dla punktu startowego $(x_0, y_0, z_0) = (1, 1, -2)$.

2. Znajdź wyrażenia, których wartości są wskaźnikami uwarunkowania zadania obliczania liczby $w = a^4 - b^4$ ze względu na dane a, b.
Dla jakich danych zadanie to jest dobrze, a dla jakich źle uwarunkowane?

3. Wartości $f_1, \ldots, f_M$ pewnej funkcji rzeczywistej f są podane w punktach $x_1, \ldots, x_M$. Należy skonstruować taką kubiczną funkcję sklejaną
$$s(x) = \sum_{i=0}^{N-4} d_i N_i^3(x),$$
aby wyrażenie $\sum_{i=1}^{M} (f_i - s(x_i))^2$ było jak najmniejsze.
   a) Napisz układ równań liniowych, taki że powyższe zadanie aproksymacji jest równoważne liniowemu zadaniu najmniejszych kwadratów dla tego układu równań. Co można powiedzieć o macierzy tego układu?
   b) Podaj warunek, który musi spełniać ciąg węzłów użyty do zdefiniowania funkcji B-sklejanych, aby zadanie było regularne.
   c) Jakich metod można użyć do rozwiązania tego LZNK, przy założeniu, że liczby M i N są rzędu 10^2?

4. Podaj najmniejsze n, takie że błąd aproksymacji jednostajnej funkcji $f(x) = \sin x$ w przedziale $[0, \pi/2]$ przez wielomian interpolacyjny z węzłami Czebyszewa jest mniejszy niż 10^{-3}. W oszacowaniu możesz skorzystać z nierówności $\pi^2 < 10$ (dokładniej, jest $\sqrt{10} \approx 3.162$).

5. Skonstruuj odpowiednią bazę Newtona i rozwiąż przy użyciu algorytmu różnic dzielonych zadanie interpolacyjne Hermite'a dla danych przedstawionych w tabelce:

x_i	0	1	3
$f(x_i)$	-2	-3	7
$f'(x_i)$	0		33
$f''(x_i)$	2		

6. Które z podanych na wykładzie metod rozwiązywania układów równań liniowych mogą być użyte do rozwiązania układu równań liniowych:

   a) Układu z macierzą o postaci $I - 2\boldsymbol{v}\boldsymbol{v}^T$, gdzie $\boldsymbol{v}$ jest danym wektorem spełniającym warunek $\|\boldsymbol{v}\|_2 = 1$.

   b) Układu dualnych równań normalnych dla dualnego liniowego zadania najmniejszych kwadratów z liczbą równań nie przekraczającą 100.

   c) Układu równań z wielką macierzą ($n \times n$, gdzie $n > 10^4$) niesymetryczną i diagonalnie dominującą, która ma w każdym wierszu mniej niż 20 niezerowych współczynników rozmieszczonych nieregularnie.

   W każdym przypadku napisz, z uzasadnieniem, która z tych metod wydaje się najbardziej odpowiednia.

7. Wiedząc, że dla każdego $n \in \{0, 1, 2, 3, \ldots\}$

$$\int_0^\infty x^n e^{-x}\, dx = n!,$$

   znajdź pierwsze cztery wielomiany ortogonalne Laguerre'a, tj. wielomiany stopnia 0, 1, 2, 3, ortogonalne w sensie iloczynu skalarnego

$$\langle f, g \rangle = \int_0^\infty f(x)g(x)e^{-x}\, dx,$$

   za pomocą ortogonalizacji Grama-Schmidta lub formuły trójczłonowej.

8. Korzystając ze wskazówki i wyników poprzedniego zadania, znajdź węzły i współczynniki kwadratury Gaussa-Laguerre'a czwartego rzędu.

Egzamin z Metod Numerycznych, III rok Inf.

(*Ściśle tajne* przed godz. 14:30 3 lutego 2014.)

Proszę uważnie przeczytać treść zadań. *Bardzo* **duży wpływ** na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi oraz użycie dobrych praktyk przedstawionych na wykładzie.

1. Przy założeniu, że punkt startowy $x_0 \neq 2$ leży w kuli zbieżności rozwiązania $\alpha = 2$ dla każdego z równań

   a) $f(x) = 0$, gdzie $f(x) \stackrel{\text{def}}{=} x^3 - 5x^2 + 8x - 4$,

   b) $g(x) = 0$, gdzie $g(x) \stackrel{\text{def}}{=} x^3 - 4x^2 + 5x - 2$,

   jaka będzie szybkość zbieżności metody Newtona? Odpowiedź uzasadnij.

2. Dane są wektory $\mathbf{b}_1, \mathbf{b}_2, \mathbf{c}_1, \mathbf{c}_2, \boldsymbol{v} \in \mathbb{R}^n$, liczba rzeczywista d, oraz macierz trójdiagonalna T o wymiarach $n \times n$, symetryczna i dodatnio określona. Wektor $\boldsymbol{v}$ jest jednostkowy, tj. $\|\boldsymbol{v}\|_2 = 1$, i określa macierz $H = I - 2\boldsymbol{v}\boldsymbol{v}^T$. Podaj algorytm, który kosztem proporcjonalnym do n rozwiąże układ równań liniowych

$$\begin{bmatrix} H & 0 & \mathbf{c}_1 \\ 0 & T & \mathbf{c}_2 \\ \mathbf{c}_1^T & \mathbf{c}_2^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \\ z \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ d \end{bmatrix}$$

   (z niewiadomymi $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $z \in \mathbb{R}$) lub stwierdzi, że rozwiązanie nie istnieje lub jest niejednoznaczne.

3. Niech A będzie macierzą 4×2, taką że

$$H_1 A = \begin{bmatrix} 2 & 3 \\ 0 & -1 \\ 0 & 2 \\ 0 & -2 \end{bmatrix}, \text{ przy czym } H_1 = I - \boldsymbol{v}_1 \frac{2}{\boldsymbol{v}_1^T \boldsymbol{v}_1} \boldsymbol{v}_1^T, \text{ gdzie } \boldsymbol{v}_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}.$$

   Oblicz:

   a) macierz A,

   b) macierz trójkątną górną R wymiaru 4×2 i wektor $\boldsymbol{v}_2$ wyznaczający odbicie Householdera o macierzy H_2 takiej, że macierze $Q = H_1 H_2$ i R są czynnikami rozkładu QR macierzy A,

   c) Rozwiązanie LZNK dla układu z macierzą A i wektorem prawej strony $\mathbf{b} = [4, -2, 1, -3]^T$.

4. Niech $\varepsilon > 0$. Oblicz wskaźnik uwarunkowania w normie $\|\cdot\|_1$ macierzy

$$M_\varepsilon = \begin{bmatrix} \varepsilon & \varepsilon & \varepsilon \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

5. Zbadaj, które z wektorów: $[1,1,-1]^T$, $[1,1,0]^T$, $[-1,1,0]^T$, $[0,-1,1]^T$, $[0,0,1]^T$, są wektorami własnymi macierzy

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

Znajdź wszystkie jej wartości własne. Czy (prosta) metoda potęgowa rozwiązywania zagadnienia własnego byłaby dla tej macierzy skuteczna?

6. Należy znaleźć wielomian interpolacyjny Hermite'a dla trzech węzłów, za pomocą algorytmu różnic dzielonych. Ktoś utworzył i częściowo wypełnił następującą tabelkę różnic dzielonych:

$$\begin{array}{c|cccccc}
0 & 1 & & & & \\
 & & 1 & & & \\
1 & 2 & & 3 & & \\
 & & -2 & & -1 & \\
1 & 2 & & 2 & & ? \\
 & & -2 & & ? & \\
1 & 2 & & 0 & & \\
 & & -2 & & & \\
2 & 0 & & & &
\end{array}$$

Odczytaj z tabelki nałożone na wielomian warunki interpolacyjne.

Sprawdź, czy obliczone (a nie wpisane!) elementy tabelki zostały obliczone poprawnie i jeśli nie, to je skoryguj. Wypełnij tabelkę do końca, a następnie podaj współczynniki wielomianu Hermite'a w bazie Newtona związanej z tymi węzłami (z krotnościami i w kolejności z tabelki).

7. Funkcję $f(x) = \sin x$ na odcinku $[0, \pi]$ przybliżamy interpolacyjną funkcją sklejaną pierwszego stopnia, $s(x)$, opartą na $N+1$ równoodległych węzłach (będących zarówno węzłami interpolacyjnymi, jak i węzłami funkcji sklejanej): $x_k = k\pi/N$ dla $k = 0, \ldots, N$. Jakie N wystarczy, aby błąd aproksymacji jednostajnej funkcji f przez s był mniejszy niż 10^{-3}?

8. Udowodnij, że jeśli kwadratura interpolacyjna przybliżająca całkę

$$I(f) = \int_{-a}^{a} f(x)\rho(x)\,dx$$

z parzystą funkcją wagową ρ ma *nieparzystą* liczbę węzłów rozmieszczonych w przedziale $[-a, a]$ symetrycznie względem zera, to rząd tej kwadratury jest większy od liczby węzłów.

Podaj (przynajmniej dwa) przykłady takich kwadratur.

Kolokwium z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 14:15 24 kwietnia 2014.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Wykaż, że funkcja
$$\varphi(x) \overset{\text{def}}{=} \frac{1}{x+2} + \frac{1}{2}x + 100\pi$$
ma w przedziale $[0, 2104]$ dokładnie jeden punkt stały i że można ten punkt znaleźć za pomocą metody iteracji prostej. Jaki jest wykładnik zbieżności tej metody?

2. Pierwiastki trójmianu kwadratowego $f(z) = z^2 + az + b$ ustawiamy w wektor $[z_1, z_2]^T \in \mathbb{C}^2$. Oblicz wskaźnik uwarunkowania w normie maksimum zadania znajdowania tego wektora ze względu na zaburzenia danej a, jeśli $a = -2$, $b = 3$.

3. Jeśli istnieje macierz trójkątna dolna L, taka że macierz
$$A = \begin{bmatrix} 9 & -3 & 3 \\ -3 & 5 & 3 \\ 3 & 3 & 6 \end{bmatrix}$$
jest równa iloczynowi LL^T, to znajdź macierz L metodą Choleskiego.
W przeciwnym razie znajdź metodą eliminacji Gaussa z wyborem elementu głównego czynniki trójkątne L i U rozkładu macierzy PA; macierz permutacji P przedstaw w postaci ciągu par numerów kolejno przestawianych wierszy.
Korzystając ze znalezionych czynników rozkładu rozwiąż układ równań liniowych $A\mathbf{x} = \mathbf{b}$, gdzie $b = [12, 0, 9]^T$.
Sprawdź wynik.

4. Współczynniki a_0, a_1 funkcji $f(x) = a_0 + a_1 x$ należy dobrać tak, aby zminimalizować wyrażenie

$$R = \sum_{i=1}^{4} \big(f(x_i) - y_i\big)^2,$$

dla liczb $x_1, \ldots, x_4$ i $y_1, \ldots, y_4$ podanych w tabelce:

x_i	-2	-1	0	1
y_i	1	2	0	1

Postaw i rozwiąż przy użyciu metody odbić Householdera odpowiednie liniowe zadanie najmniejszych kwadratów. Znajdź minimalne R.
Sprawdź wynik, podstawiając go do utworzonego w tym celu układu równań normalnych.

Kolokwium z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 12:15 29 kwietnia 2015.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Rozważamy równanie $10x - \sin(x) = 3$. Czy metoda iteracyjna określona wzorem $x_{n+1} = (\sin(x_n) + 3)/10$ zbiegnie do rozwiązania α dla $x_0 = 0.33$? Jeśli tak, to określ, dla jakich k zachodzi na pewno nierówność $|x_k - \alpha| \leqslant 10^{-16}|\alpha|$ (jeśli w obliczeniach nie ma błędów zaokrągleń).

2. Wartość w funkcji $f(x) = \sqrt{p_1(x)} - \sqrt{p_2(x)}$, takiej że $p_1(x) = x^3 + 1$ i $p_2(x) = x^3 + 4x^2 + 10$, może być obliczona takimi sposobami:

   Algorytm 1: Obliczamy $w_k = \sqrt{p_k(x)}$ dla $k = 1, 2$, a następnie $w = w_1 - w_2$.
   Algorytm 2: Obliczamy $w_k = \sqrt{p_k(x)}$ dla $k = 1, 2$, a następnie $w = -(4x^2 + 9)/(w_1 + w_2)$.

   W obu przypadkach do obliczenia wartości wielomianów używamy schematu Hornera. Zakładamy, że x jest dodatnie i jest duże (rzędu 10^{10}). Który algorytm, realizowany w arytmetyce zmiennopozycyjnej, jest lepszy? Odpowiedź uzasadnij w kilku zdaniach.

3. Metodą eliminacji Gaussa z wyborem elementu głównego w kolumnie rozwiąż układ równań liniowych $A\mathbf{x} = \mathbf{b}$, dla

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 1 & 0 \\ 8 & 2 & 2 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 1 \\ -6 \end{bmatrix}. \qquad \text{Sprawdź wynik.}$$

4. Niech

$$\boldsymbol{v}_1 = \begin{bmatrix} -4 \\ 2 \\ -1 \\ 2 \end{bmatrix}, \quad \boldsymbol{v}_2 = \begin{bmatrix} 0 \\ 12 \\ -3 \\ 4 \end{bmatrix}, \quad \boldsymbol{v}_3 = \begin{bmatrix} 0 \\ 0 \\ 4 \\ 3 \end{bmatrix}, \quad R = \begin{bmatrix} 1 & 1 & -1 \\ 0 & 1 & -1 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

   Macierz A jest równa iloczynowi $QR = H_1H_2H_3R$, w którym czynniki H_1, H_2, H_3 są macierzami odbić symetrycznych w $\mathbb{R}^4$, względem hiperpłaszczyzn prostopadłych odpowiednio do wektorów $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ i $\boldsymbol{v}_3$. Rozwiąż liniowe zadanie najmniejszych kwadratów dla układu równań liniowych z macierzą A i wektorem prawej strony $\mathbf{b} = [41, -50, -72, 96]^T$. Uzasadnij poprawność użytej metody. Oblicz drugą normę residuum.

Kolokwium z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 12:15 20 maja 2016.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. a) Wykonaj dwie iteracje metody Newtona dla układu równań

$$\begin{cases} x^2 + y^2 - 2 &= 0, \\ (x+1)(y^2-1) &= 0, \end{cases}$$

z punktem początkowym $(x_0, y_0) = (2, 2)$.

b) Wybierając odpowiednio punkt początkowy, można otrzymać ciągi zbieżne do wszystkich czterech rozwiązań tego układu: $(-1,-1)$, $(-1,1)$, $(1,-1)$ i $(1,1)$. Co można powiedzieć o wykładniku zbieżności metody Newtona dla każdego z tych rozwiązań? Odpowiedź uzasadnij.

2. Niech funkcja $\mathbf{f}\colon \mathbb{R}^2 \to \mathbb{R}^2$ będzie określona wzorem

$$\mathbf{f}(x,y) = \begin{bmatrix} x^2 + y^2 - 2 \\ (x+1)(y^2-1) \end{bmatrix}.$$

Oblicz wskaźnik uwarunkowania zadania obliczania wartości tej funkcji w punkcie $(x, y) = (2, 2)$, w normie $\|\cdot\|_\infty$, ze względu na zaburzenia argumentu x. Użytą metodę i wykonane rachunki uzasadnij.

3. Niech

$$A = \begin{bmatrix} 9 & 3 & 0 \\ 3 & 5 & 4 \\ 0 & 4 & 5 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 6 \\ 2 \\ 1 \end{bmatrix}.$$

Rozwiąż układ równań $Ax = \mathbf{b}$, stosując metodę Choleskiego, jeśli można ją zastosować do tego układu, lub metodę eliminacji Gaussa w przeciwnym razie.

4. Niech $0 < n < m$ i niech A oznacza rzeczywistą macierz $m \times n$ o kolumnach liniowo niezależnych. Niech $A = QR$; macierz $Q \in \mathbb{R}^{m \times m}$ jest ortogonalna, a macierz $R \in \mathbb{R}^{m \times n}$ — trójkątna górna, tj. jej współczynniki r_{ij} dla $i > j$ są równe 0. Udowodnij, że pierwsze n kolumn macierzy Q jest określone z dokładnością do zwrotu. Jak zmiana zwrotu j-tej kolumny macierzy Q wpływa na czynnik R?

   Czy jeśli kolumny macierzy A są liniowo zależne, to istnieje nieskończenie wiele różnych sposobów wybierania pierwszych n kolumn macierzy Q? Odpowiedź uzasadnij.

Kolokwium poprawkowe
z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 14:15 30 maja 2016.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Niech $f(x) = (x-2)^3 + 7$. Funkcja f ma w przedziale $[0,1]$ miejsce zerowe, które chcemy znaleźć metodą iteracji prostej,

$$x_{k+1} = \varphi(x_k) \quad \text{dla } k = 0, 1, 2, \ldots,$$

przy użyciu funkcji $\varphi(x) = x - \tau f(x)$. Dla jakich wartości parametru τ przedział $[0,1]$ będzie zawarty w kuli zbieżności poszukiwanego miejsca zerowego?

2. Chcąc otrzymać przybliżoną wartość funkcji e^x, możemy obliczyć wartość wielomianu

$$f(x) = \sum_{k=0}^{9} \frac{x^k}{k!}$$

który dla $x \in [-1,1]$ przybliża e^x z błędem (bezwzględnym) mniejszym niż $3.1 \cdot 10^{-7}$. Obliczenia wykonujemy przy użyciu algorytmu

```
f = x;
for ( k = 8; k > 0; k-- )
  f = (f/(k+1.0)+1.0)*x;
f = f+1.0;
```

zrealizowanego przy użyciu arytmetyki zmiennopozycyjnej pojedynczej precyzji (w której $\nu = 2^{-23} \approx 10^{-7}$). Możemy też obliczyć $e^x \approx 1/f(-x)$, po obliczeniu $f(-x)$ tym samym algorytmem. Dla jakich wartości x można oczekiwać, że druga metoda zapewnia lepszą dokładność (tj. mniejszy błąd względny) końcowego wyniku? Odpowiedź uzasadnij.

3. Niech

$$B = \begin{bmatrix} 9 & 3 & 3 \\ 3 & 5 & 1 \\ 3 & 1 & 2 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \end{bmatrix}.$$

Rozwiąż układ równań liniowych

$$\begin{bmatrix} B & 0 \\ C & H \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathbf{d} \end{bmatrix},$$

którego macierz zawiera blok $H = I - \frac{1}{2}\boldsymbol{v}\boldsymbol{v}^T$, a prawa strona składa się z wektorów $\mathbf{c} = [12, 0, 6]^T$ i $\mathbf{d} = [4, -2, 3, -3]^T$. W tym celu podziel układ na podukłady, które można rozwiązać osobno i do każdego podukładu zastosuj najbardziej odpowiedni algorytm spośród podanych na wykładzie.

4. Metodą odbić Householdera rozwiąż liniowe zadanie najmniejszych kwadratów dla układu równań $A\boldsymbol{x} = \mathbf{b}$, w którym

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

Oblicz normę drugą wektora residuum dla znalezionego wektora $\boldsymbol{x}$.

Kolokwium z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 12:15 10 maja 2017.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Niech $f(x) = \ln(e^x - x - 1)$ dla $x > 0$. Wykaż, że funkcja f ma dokładnie jedno miejsce zerowe α. Zaproponuj metodę iteracyjną, za pomocą której można otrzymać ciąg $(x_k)_k$ zbieżny kwadratowo do α. Jak należy wybrać punkt startowy?

2. Niech $\mathbf{y}$ oznacza ustalony niezerowy element pewnej przestrzeni liniowej X wyposażonej w iloczyn skalarny $\langle \cdot, \cdot \rangle$. Znajdź wzór opisujący wskaźnik uwarunkowania zadania obliczania wartości funkcji $f\colon X \to \mathbb{R}$ danej wzorem $f(\mathbf{x}) = \langle \mathbf{x}, \mathbf{y} \rangle$, ze względu na zaburzenia wektora $\mathbf{x}$. Dla jakich danych zadanie to jest dobrze, a dla jakich źle uwarunkowane?

   Uwaga: W przestrzeni X przyjmujemy normę $\|\mathbf{x}\| \stackrel{\text{def}}{=} \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$.

3. Niech A oznacza rzeczywistą macierz $n \times n$ o współczynnikach a_{ij}, symetryczną i dodatnio określoną, i niech $0 \leqslant k < n - 1$. Wykaż, że jeśli macierz A jest $2k + 1$-diagonalna, tj. $a_{ij} = 0$ dla $|i - j| > k$, to macierz trójkątna dolna L o współczynnikach l_{ij}, taka że $A = LL^T$, jest $k + 1$-diagonalna, tj. $l_{ij} = 0$ dla $i - j > k$.

   Jaki jest rząd złożoności obliczeniowej sensownej metody rozwiązywania układu równań liniowych z taką macierzą, jeśli $k \ll n$?

4. Rozważamy liniowe zadanie najmniejszych kwadratów dla układu równań liniowych $A\mathbf{x} = \mathbf{b}$, w którym

$$A = \begin{bmatrix} 3 & 1 \\ 0 & 2 \\ 0 & 1 \\ 4 & -2 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ -3 \\ 1 \\ 6 \end{bmatrix}.$$

   a) Napisz układ równań normalnych dla tego zadania.

   b) Rozwiąż zadanie, stosując odbicia Householdera do macierzy A i wektora $\mathbf{b}$.

   c) Znajdź wektor residuum znalezionego rozwiązania zadania i normę drugą tego wektora.

Kolokwium poprawkowe z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 16:15 31 maja 2017.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Wielomian $p(x) = 16x^4 - 8x^3 + 4x^2 - 1$ ma miejsce zerowe $\alpha = \frac{1}{2}$.

   a) Znajdź (jak najdłuższy) przedział $A = (\frac{1}{2} - a, \frac{1}{2} + a)$, taki że dla każdego $x_0 \in A$ metoda Newtona wytworzy ciąg zbieżny do α.

   b) Jaki jest wykładnik zbieżności metody Newtona w tym przypadku?

2. Wielomian rzeczywisty $p(x)$ stopnia $n > 0$ ma wolny wyraz $a_0 \neq 0$ oraz miejsce zerowe $\alpha \in \mathbb{R}$, które należy znaleźć. Znajdź wyrażenie, którego wartością jest wskaźnik uwarunkowania numerycznego tego zadania ze względu na zaburzenia współczynnika a_0.

   Kiedy to zadanie jest źle uwarunkowane?

   Wskazówka. Rozwiń zaburzoną funkcję p w szereg Taylora wokół punktu α.

3. Niech A oznacza pewną macierz rzeczywistą 5×2 i niech $\mathbf{f} = [6, 0, 0, -2, 2]^T$. W wyniku zastosowania odbicia względem hiperpłaszczyzny prostopadłej do wektora $\boldsymbol{v}_1 = [0, 0, 0, 1, 1]^T$ powstała macierz

$$H_1 A = \begin{bmatrix} 2 & 3 \\ 0 & -1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

   a) Określ, czy LZNK z macierzą A i wektorem $\mathbf{f}$ ma jednoznaczne rozwiązanie.

   b) Wykonaj drugi krok metody Householdera (z macierzą odbicia H_2) i podaj macierz trójkątną górną R, taką że $H_2 H_1 A = R$.

   c) Rozwiąż LZNK z macierzą A i wektorem $\mathbf{f}$ za pomocą rozkładu QR z poprzedniego punktu i oblicz normę wektora residuum $\|A\boldsymbol{x} - \mathbf{f}\|_2$ dla znalezionego rozwiązania.

4. Niech

$$A = \begin{bmatrix} 4 & -4 & 0 & -2 \\ -4 & 5 & 1 & 4 \\ 0 & 1 & 5 & 6 \\ -2 & 4 & 6 & 10 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 8 \\ -7 \\ 5 \\ 4 \end{bmatrix}.$$

Za pomocą metody Choleskiego znajdź macierz trójkątną dolną L, taką że $A = LL^{T}$, a następnie korzystając z macierzy L, rozwiąż układ równań $Ax = \mathbf{b}$.

Egzamin z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 9:00 20 czerwca 2014.)

Proszę bardzo uważnie przeczytać treść zadań. **Bardzo duży** wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Macierz układu równań liniowych $A\mathbf{x} = \mathbf{b}$ ma następującą postać blokową:

$$A = \begin{bmatrix} B & 0 & \dots & 0 & C_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & B & 0 & C_{k-1} \\ 0 & \dots & 0 & B & C_k \\ C_1^T & \dots & C_{k-1}^T & C_k^T & 0 \end{bmatrix}.$$

Blok B o wymiarach $p \times p$ jest pełną macierzą symetryczną, diagonalnie dominującą. Macierz złożona z bloków $C_1, \dots, C_k \in \mathbb{R}^{p \times q}$ (ostatnia kolumna w schemacie blokowym pokazanym wyżej) ma kolumny liniowo niezależne.

a) Wykaż, że układ z taką macierzą i dowolnym wektorem $\mathbf{b} \in \mathbb{R}^{kp+q}$ ma jednoznaczne rozwiązanie.

b) Podaj algorytm rozwiązywania takiego układu równań, o możliwie małym koszcie.

<u>Wskazówka</u>. Wektor prawej strony i rozwiązanie przedstaw w postaci blokowej dostosowanej do przedstawionego podziału blokowego macierzy A.

2. Do znalezienia pewnej wartości i wektora własnego macierzy

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

chcemy użyć odwrotnej metody potęgowej z parametrem $a = -2$ i z wektorem początkowym $\mathbf{x}_0 = [1, 0, 0, 0]^T$. Czy metoda będzie zbieżna i jeśli tak, to jaką parę własną otrzymamy w granicy? Odpowiedź uzasadnij.

3. Skonstruuj odpowiednią bazę Newtona i rozwiąż, metodą różnic dzielonych, zadanie interpolacyjne Hermite'a dla danych w tabelce:

x	-1	1	2
$f(x)$	5	5	2
$f'(x)$	-4		-16
$f''(x)$	-4		
$f'''(x)$	-24		

4. Funkcję $f(x) = \sin x$ w przedziale $[0, \pi]$ przybliżamy

   a) wielomianem interpolacyjnym Lagrange'a stopnia $2n - 1$ opartym na węzłach Czebyszewa w tym przedziale,

   b) wielomianem interpolacyjnym Hermite'a stopnia $2n - 1$ opartym na dwóch węzłach: 0 i π, z których każdy ma krotność n,

   c) interpolacyjną funkcją sklejaną stopnia 1, której węzły dzielą przedział $[0, \pi]$ na $2n - 1$ podprzedziałów o jednakowej długości $\pi/(2n - 1)$; węzły te są jednocześnie węzłami interpolacyjnymi.

   Podaj górne oszacowanie błędu aproksymacji jednostajnej w każdym z tych przypadków, w zależności od n. Możesz w oszacowaniach użyć równości przybliżonej $\pi^2 \approx 10$.

5. Rozważamy iloczyn skalarny

$$\langle f, g \rangle \stackrel{\text{def}}{=} \int_0^1 f(x)g(x)x\,dx.$$

   a) Znajdź pierwsze trzy wielomiany z rodziny wielomianów ortogonalnych w sensie tego iloczynu skalarnego.

   b) Korzystając z wyniku punktu a), skonstruuj kwadraturę Gaussa rzędu 2 przybliżającą całkę

$$I(f) = \int_0^1 f(x)x\,dx$$

   i znajdź oszacowanie błędu tej kwadratury dla funkcji f klasy $C^2[0, 1]$.

Egzamin poprawkowy z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 12 września 2014.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Dane są wektory v_1, v_2, $\mathbf{b}_1$, $\mathbf{b}_2 \in \mathbb{R}^n$, przy czym v_1 i v_2 mają normę drugą równą 1 i spełniają warunek $|v_2^T v_1| \neq \sqrt{2}/2$. Niech $H_1 = I - 2v_1v_1^T$, $H_2 = I - 2v_2v_2^T$ (gdzie I oznacza macierz jednostkową $n \times n$).

   a) Podaj algorytm o koszcie rzędu n, rozwiązujący układ równań liniowych

$$\begin{bmatrix} H_1 & -H_2 \\ H_2 & H_1 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \end{bmatrix}. \qquad (*)$$

   b) Udowodnij, że układ równań (*) ma jednoznaczne rozwiązanie.

2. Niech $n > 1$. Wykaż (powołując się na odpowiednie twierdzenie z wykładu), że dowolny wielomian o postaci

$$w(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

   można przybliżyć (w sensie aproksymacji jednostajnej) w przedziale $[0, 1]$ wielomianem stopnia co najwyżej $n - 1$ z błędem nie przekraczającym $2^{1-2n}|a_n|$ i nie można lepiej.

   Jak należy wybrać węzły interpolacyjne, aby wielomian interpolacyjny stopnia co najwyżej $n - 1$ przybliżał wielomian w w przedziale $[0, 1]$ z takim właśnie błędem?

3. Skonstruuj odpowiednią bazę Newtona i znajdź współczynniki w tej bazie wielomianu interpolacyjnego Hermite'a stopnia co najwyżej 4. Warunki interpolacyjne są podane w tabelce:

x	-1	2
$f(x)$	4	16
$f'(x)$	-5	31
$f''(x)$	12	

4. Ustalona liczba rzeczywista $q > 1$ jest użyta do określenia nieskończonego ciągu węzłów $(u_i)_{i\in\mathbb{Z}}$: $u_i = q^i$ dla każdego i. Rozważamy funkcje B-sklejane N_i^n (stopnia n) określone dla takiego ciągu.

   a) Udowodnij, że $N_{i+1}^n(qx) = N_i^n(x)$ dla każdego $i \in \mathbb{Z}$ oraz $x \in \mathbb{R}$.

   b) Oblicz wartości wszystkich funkcji N_i^3 w punkcie $x = u_0$, jeśli $q = 2$.

   c) Niech $N > 6$ i niech s oznacza funkcję sklejaną trzeciego stopnia z węzłami $u_i = 2^i$ dla $i = 0, \ldots, N$. Przedstaw funkcję s w bazie złożonej z funkcji B-sklejanych określonych dla tych węzłów i napisz w postaci rozwiniętej równanie (z niewiadomymi współczynnikami funkcji s w tej bazie) opisujące warunek interpolacyjny $s(u_i) = a_i$ dla $i \in \{3, \ldots, N-3\}$; liczba a_i jest dana.

   Wskazówka. Możesz skorzystać z formuły Mansfielda-de Boora-Coxa:

$$N_i^0(x) = \begin{cases} 1 & \text{dla } x \in [u_i, u_{i+1}), \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

$$N_i^n(x) = \frac{x - u_i}{u_{i+n} - u_i} N_i^{n-1}(x) + \frac{u_{i+n+1} - x}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x) \quad \text{dla } n > 0.$$

5. Całkę

$$I(f) = \int_{-1}^{1} f(x)\,dx$$

   przybliżamy kwadraturą interpolacyjną

$$Q_a(f) = A_0 f(0) + A_1 f(a),$$

   dla pewnego $a \in [-1, 1]$.

   a) Znajdź współczynniki A_0, A_1 tej kwadratury (w zależności od węzła a).

   b) Podaj oszacowanie błędu kwadratury Q_a dla funkcji podcałkowej $f \in C^2[-1, 1]$.

   c) Czy można wybrać węzeł a tak, aby otrzymać kwadraturę rzędu większego niż 2? Odpowiedź uzasadnij.

Egzamin z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 9:00 22 czerwca 2015.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Macierz A o wymiarach $n \times n$ jest symetryczna. Opisz, jak można stosować odwrotną metodę potęgową do tej macierzy, mając do dyspozycji tylko procedurę znajdowania rozkładu na czynniki trójkątne metodą Choleskiego, procedury rozwiązywania układów równań liniowych z macierzami trójkątnymi i procedurę obliczania iloczynu skalarnego. W szczególności, jakie pary własne macierzy A da się znaleźć przy użyciu odwrotnej metody potęgowej zaimplementowanej przy użyciu tylko tych procedur?

2. Utwórz odpowiednią bazę Newtona i znajdź wielomian w stopnia co najwyżej 4 (tj. współczynniki wielomianu w tej bazie) spełniający warunki interpolacyjne Hermite'a podane w tabelce:

x	-1	3
$w(x)$	-3	73
$w'(x)$	-1	103
$w''(x)$	10	

Następnie, bez przechodzenia do bazy potęgowej, oblicz $w(1)$ oraz $w''(3)$.

3. Niech $f(x) = \max\{0, 2x\}$ dla $x \in [-1, 1]$. Znajdź wielomian stopnia co najwyżej 3, który jest optymalnym przybliżeniem funkcji f w sensie aproksymacji jednostajnej i podaj błąd aproksymacji dla tego wielomianu. Uzasadnij, że znaleziony wielomian jest optymalny, powołując się na odpowiednie twierdzenie z wykładu.

   Wskazówka. Najpierw rozwiąż zadanie aproksymacji dla funkcji $g(x) = |x|$, zaczynając od zrobienia wykresu.

4. Całkę $I(f) = \int_0^2 f(x)x\,dx$ przybliżamy kwadraturą $Q(f) = A_0 f(0) + A_1 f(c)$. Dobierz węzeł c i współczynniki A_0, A_1 tak, aby rząd kwadratury Q, oznaczony literą r, był jak największy. Znajdź ten rząd i podaj oszacowanie błędu kwadratury dla funkcji $f \in C^r[0, 2]$.

5. Niech $u_i = i$ dla $i = 0, \ldots, N$, gdzie $N > 10$. Kubiczną funkcję sklejaną s z węzłami u_i przedstawiamy w bazie B-sklejanej,

$$s(x) = \sum_{i=0}^{N-4} d_i N_i^3(x).$$

Równanie opisujące warunek interpolacyjny $s(u_i) = a_i$ (gdzie $i \in \{3, \ldots, N-3\}$) oraz równanie opisujące warunek brzegowy $\lim_{x \searrow u_3} s'''(x) = 0$ napisz w postaci jawnej, tj. podając liczby, które są współczynnikami w tych równaniach.

Wskazówka. Do obliczenia współczynników możesz użyć wzorów

$$N_i^0(x) = \begin{cases} 1 & \text{dla } x \in [u_i, u_{i+1}), \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

$$N_i^n(x) = \frac{x - u_i}{u_{i+n} - u_i} N_i^{n-1}(x) + \frac{u_{i+n+1} - x}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x)$$

oraz

$$\frac{d}{dx} N_i^n(x) = \frac{n}{u_{i+n} - u_i} N_i^{n-1}(x) - \frac{n}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x).$$

Egzamin poprawkowy z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 9 września 2015.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Niech

$$A = \begin{bmatrix} -5 & 8 & 8 \\ 8 & 16 & -20 \\ 8 & -20 & -11 \end{bmatrix}, \quad \boldsymbol{v} = \begin{bmatrix} 2 \\ -2 \\ 1 \end{bmatrix}$$

i niech $H = I - \frac{2}{\boldsymbol{v}^T\boldsymbol{v}}\boldsymbol{v}\boldsymbol{v}^T \in \mathbb{R}^3$.

a) Oblicz macierz $B = H^{-1}AH$.

b) Korzystając z wyniku obliczenia w punkcie a), znajdź pełne rozwiązanie algebraicznego zagadnienia własnego dla macierzy A.

c) Co można powiedzieć o zbieżności (prostej) metody potęgowej rozwiązywania algebraicznego zagadnienia własnego, zastosowanej do macierzy A?

2. Niech

$$f(x) = \begin{cases} 2x+1 & \text{dla } x < -\frac{1}{2}, \\ 0 & \text{dla } -\frac{1}{2} \leqslant x < \frac{1}{2}, \\ 2x-1 & \text{dla } \frac{1}{2} \leqslant x. \end{cases}$$

Znajdź wielomian p stopnia co najwyżej 2, który jest najlepszym przybliżeniem funkcji f w przedziale $[-1,1]$ w sensie aproksymacji jednostajnej (tj. w normie $\|g\| = \max_{x\in[-1,1]} |g(x)|$). Uzasadnij, że znaleziony wielomian jest tym, który należało znaleźć, powołując się na odpowiednie twierdzenie z wykładu.

<u>Wskazówka:</u> Funkcja f jest nieparzysta, przedział $[-1,1]$ jest symetryczny względem zera.

3. Chcemy znaleźć funkcję sklejaną s stopnia 2, klasy $C^1[a, b]$, która w węzłach $u_0 = a < u_1 < \cdots < u_{N-1} < u_N = b$ przyjmuje zadane wartości $a_0, a_1, \ldots, a_{N-1}, a_N$. Znajdź (metodą różnic dzielonych) wielomiany drugiego stopnia $H_{i,0}, H_{i,1}, H_{i,2}$, takie że

$$
\begin{aligned}
&H_{i,0}(u_i) = 1, \quad &&H_{i,0}(u_{i+1}) = 0, \quad &&H'_{i,0}(u_{i+1}) = 0,\\
&H_{i,1}(u_i) = 0, \quad &&H_{i,1}(u_{i+1}) = 1, \quad &&H'_{i,1}(u_i) \quad = 0,\\
&H_{i,2}(u_i) = 0, \quad &&H_{i,2}(u_{i+1}) = 0, \quad &&H'_{i,2}(u_i) \quad = 1.
\end{aligned}
$$

Wielomian p_i, opisujący funkcję s w przedziale $[u_i, u_{i+1}]$, można przedstawić w tej bazie: $p_i(x) = a_i H_{i,0}(x) + a_{i+1} H_{i,1}(x) + b_i H_{i,2}(x)$. Wyprowadź równanie liniowe z niewiadomymi współczynnikami b_{k-1}, b_k, którego spełnienie zapewnia ciągłość pochodnej funkcji s w węźle u_k.

Czy układ takich równań dla $k = 1, \ldots, N-1$ jest niesprzeczny i czy jest określony? Odpowiedź uzasadnij.

4. Wiedząc, że

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi},$$

oraz

$$\int_{-\infty}^{+\infty} x^{2k} e^{-x^2}\,dx = \frac{(2k-1)!!}{2^k}\sqrt{\pi} \quad \text{dla } k = 1, 2, 3, \ldots$$

($(2k-1)!!$ oznacza iloczyn wszystkich liczb nieparzystych od 1 do $2k-1$), znajdź wielomian stopnia co najwyżej 3, który jest optymalnym przybliżeniem funkcji $g(x) = x^4$ w normie określonej wzorem

$$\|f\| = \left(\int_{-\infty}^{+\infty} f^2(x) e^{-x^2}\,dx\right)^{1/2}.$$

<u>Wskazówka:</u> Funkcje x^4 oraz e^{-x^2} są parzyste.

5. Wiedząc to, co w poprzednim zadaniu, znajdź węzły i współczynniki kwadratury Gaussa-Hermite'a czwartego rzędu, przybliżającej całkę

$$I(f) = \int_{-\infty}^{+\infty} f(x) e^{-x^2}\,dx.$$

Egzamin z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 27 czerwca 2016.)

Proszę bardzo uważnie przeczytać treść zadań. bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Dla macierzy symetrycznej

$$A = \begin{bmatrix} -0.5 & -1 \\ -1 & -0.5 \end{bmatrix}$$

odwrotna metoda potęgowa z parametrem α oraz z punktem startowym $x_0 = \frac{1}{\sqrt{5}}[1, 2]^T$, wytworzyła ciąg wektorów x_k (o normie drugiej równej 1).

Określ, czy dla $\alpha = -3$ i $\alpha = 3$ ciągi ilorazów Rayleigha zbiegną.

Jeśli tak, to określ ich granice i jeśli zbiegną oba, to który szybciej zbiegnie, czy też szybkość zbieżności obu ciągów będzie jednakowa?

2. Znajdź wielomian interpolacyjny stopnia co najwyżej 5 dla danych (warunków interpolacyjnych) podanych w tabelce:

x	0	1	3
$f(x)$	1	3	1
$f'(x)$	0	3	
$f''(x)$	6		

Rozwiązanie zadania przedstaw w odpowiedniej bazie Newtona.

3. Niech $f(x) = \sin x$. Znajdź wielomian p stopnia co najwyżej 2, który jest najlepszym przybliżeniem funkcji f w przedziale $[-\pi, \pi]$ w sensie aproksymacji jednostajnej (tj. w normie $\|g\| = \max_{x\in[-\pi,\pi]} |g(x)|$). Uzasadnij, że znaleziony wielomian jest tym, który należało znaleźć, powołując się na odpowiednie twierdzenie z wykładu.

<u>Wskazówka:</u> Narysuj wykres i skorzystaj z widocznych na nim symetrii.

4. Wiedząc, że

$$\int_0^\infty x^k e^{-x}\,dx = k! \quad \text{dla każdego } k \in \mathbb{N},$$

znajdź pierwsze 3 wielomiany z rodziny wielomianów ortogonalnych w sensie iloczynu skalarnego

$$\langle f, g\rangle = \int_0^\infty f(x)g(x)e^{-x}\,dx.$$

Znajdź wielomian stopnia co najwyżej 2 będący optymalnym przybliżeniem funkcji $f(x) = x^3$ w normie $\|f\| = \sqrt{\langle f, f\rangle}$ i podaj błąd aproksymacji funkcji f przez ten wielomian.

5. Kwadratura przybliżająca całkę $\int_{-1}^1 f(x)\,dx$ jest określona wzorem

$$Q(f) = A\big(f(-1) + f(1)\big) + B\big(f(-a) + f(a)\big)$$

dla pewnego $a \in (0, 1)$.

a) Wykaż (uzasadnij), że dla dowolnego $a \in (0, 1)$ można dobrać współczynniki A i B tak, aby była to kwadratura interpolacyjna.

b) Wykaż (uzasadnij), że rząd takiej kwadratury dla dowolnego $a \in (0, 1)$ jest większy niż 4 i mniejszy niż 8.

c) Wybierz węzeł a i współczynniki A i B tak, aby rząd kwadratury określony podanym wzorem był największy.

Egzamin poprawkowy z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 7 września 2016.)

Proszę bardzo uważnie przeczytać treść zadań. bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Diagonalizowalna macierz rzeczywista A ma widmo $\{-2, 1-2i, 1+2i, +1, -i, +i, +2\}$. Do znalezienia jej rzeczywistych par własnych chcemy użyć odwrotnej metody potęgowej, z rzeczywistym parametrem przesunięcia.

   a) Znajdź zbiór wszystkich liczb rzeczywistych, które przyjęte jako parametr w odwrotnej metodzie potęgowej mogą spowodować brak zbieżności metody.

   b) Wskaż, dla każdej rzeczywistej wartości własnej, taki przedział, że wybranie z niego parametru przesunięcia zapewnia, że błąd przybliżenia odpowiedniego (jednostkowego) wektora własnego przez wektor otrzymany w każdej kolejnej iteracji maleje co najmniej dziesięciokrotnie.

2. Znajdź wielomian interpolacyjny f stopnia co najwyżej 5 dla danych (warunków interpolacyjnych) podanych w tabelce:

x	-1	2
$f(x)$	4	-2
$f'(x)$	-17	4
$f''(x)$	34	16

   Rozwiązanie zadania przedstaw w odpowiedniej bazie Newtona. Oblicz $f(0)$, $f'(0)$ i $f''(0)$, bez obliczania pochodnych wielomianów będących elementami bazy Newtona ani potęgowej.

3. Niech $f(x) = x + \big||2x| - 1\big|$. Znajdź wielomian p stopnia co najwyżej 1, będący optymalnym przybliżeniem funkcji f w sensie aproksymacji jednostajnej w przedziale $[-1, 1]$.

   Jaki jest najmniejszy stopień wielomianu, który jest lepszym jednostajnym przybliżeniem funkcji f w przedziale $[-1, 1]$ niż optymalny wielomain stopnia co najwyżej 1? Odpowiedź uzasadnij, powołując się na odpowiednie twierdzenie z wykładu.

4. Znajdź funkcję sklejaną s stopnia 2 klasy $C^1[-1, 1]$ z węzłami $-1, 0, 1$, która spełnia warunki interpolacyjne $s(-1) = 0$, $s(-1/2) = 0$, $s(1) = 1$, $s'(1) = 0$. Czy istnieje tylko jedna taka funkcja? Odpowiedź uzasadnij.

5. Znajdź kwadratury Gaussa rzędu 4 (tj. oblicz węzły i współczynniki) dla całek określonych wzorami

$$I_1(f) = \int_{-1}^{1} f(x)|x|\,dx \qquad \text{oraz} \qquad I_2(f) = \int_{-1}^{1} f(x)(1 - |x|)\,dx.$$

   Dla jednej (dowolnej) z tych kwadratur wyprowadź wzór na oszacowanie błędu kwadratury, jeśli funkcja podcałkowa jest klasy $C^4[-1, 1]$.

Egzamin z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 19 czerwca 2017.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Macierz układu równań liniowych $A\boldsymbol{x} = \mathbf{b}$ ma następującą postać blokową:

$$A = \begin{bmatrix} B & 0 & \dots & 0 & C_1 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & B & 0 & C_{k-1} \\ 0 & \dots & 0 & B & C_k \\ C_1^T & \dots & C_{k-1}^T & C_k^T & 0 \end{bmatrix}.$$

Blok B o wymiarach $p \times p$ jest pełną macierzą symetryczną i dodatnio określoną. Macierz złożona z bloków $C_1, \dots, C_k \in \mathbb{R}^{p\times q}$ (ostatnia kolumna w schemacie blokowym pokazanym wyżej) ma kolumny liniowo niezależne.

a) Wykaż, że układ z taką macierzą i dowolnym wektorem $\mathbf{b} \in \mathbb{R}^{kp+q}$ ma jednoznaczne rozwiązanie.

b) Podaj algorytm rozwiązywania takiego układu równań, o możliwie małym koszcie.

Wskazówka. Wektor prawej strony i rozwiązanie przedstaw w postaci blokowej dostosowanej do przedstawionego podziału blokowego macierzy A.

2. Niech $n > 1$. Wykaż (powołując się na odpowiednie twierdzenie z wykładu), że dowolny wielomian o postaci

$$w(x) = a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0$$

można przybliżyć (w sensie aproksymacji jednostajnej) w przedziale $[0, 1]$ wielomianem stopnia co najwyżej $n-1$ z błędem nie przekraczającym $2^{1-2n}|a_n|$ i nie można lepiej.

Jak należy wybrać węzły interpolacyjne, aby wielomian interpolacyjny stopnia co najwyżej $n-1$ przybliżał wielomian w w przedziale $[0, 1]$ z takim właśnie błędem?

3. Całkę $I(f) = \int_0^2 f(x)x\,dx$ przybliżamy kwadraturą $Q(f) = A_0 f(0) + A_1 f(c)$. Dobierz węzeł c i współczynniki A_0, A_1 tak, aby rząd kwadratury Q, oznaczony literą r, był jak największy. Znajdź ten rząd i podaj oszacowanie błędu kwadratury dla funkcji $f \in C^r[0, 2]$.

4. Dla macierzy symetrycznej

$$A = \begin{bmatrix} -0.5 & -1 \\ -1 & -0.5 \end{bmatrix}$$

odwrotna metoda potęgowa z parametrem a oraz z punktem startowym $\mathbf{x}_0 = \frac{1}{\sqrt{5}}[1,2]^T$, wytworzyła ciąg wektorów $\mathbf{x}_k$ (o normie drugiej równej 1).

Określ, czy dla $a = -3$ i $a = 3$ ciągi ilorazów Rayleigha zbiegną.

Jeśli tak, to określ ich granice i jeśli zbiegną oba, to który szybciej zbiegnie, czy też szybkość zbieżności obu ciągów będzie jednakowa?

5. Znajdź wielomian interpolacyjny stopnia co najwyżej 5 dla danych (warunków interpolacyjnych) podanych w tabelce:

x	0	1	3
$f(x)$	1	3	1
$f'(x)$	0	3	
$f''(x)$	6		

Rozwiązanie zadania przedstaw w odpowiedniej bazie Newtona.

Egzamin poprawkowy z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 10:00 6 września 2017.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Macierz rzeczywista

$$A = \begin{bmatrix} 0 & 4B \\ B & 0 \end{bmatrix}$$

ma wymiary $2n \times 2n$, a jej blok B jest macierzą symetryczną i dodatnio określoną. Czy można znaleźć parę własną (λ, x) macierzy A z wartością własną λ o największej wartości bezwzględnej za pomocą prostej metody potęgowej i jeśli tak, to jak należy tej metody użyć?

Wskazówka. Zastanów się nad związkiem algebraicznych zadań własnych dla macierzy A i B.

2. Rozwiąż zadanie interpolacyjne Hermite'a w przestrzeni wielomianów stopnia co najwyżej 5 dla danych przedstawionych w tabelce:

x	-2	0	1
$f(x)$	9	1	0
$f'(x)$	-24		0
$f''(x)$	44		

Rozwiązanie przedstaw w bazie Newtona skonstruowanej dla podanych w tabelce węzłów interpolacyjnych. Następnie oblicz pochodną trzeciego rzędu wielomianu interpolacyjnego w punkcie 1, bez obliczania pochodnych wielomianów z bazy Newtona i bez znajdowania reprezentacji rozwiązania w innej bazie.

3. Rozważamy funkcję sklejaną s stopnia 3 z węzłami $0, 1, \ldots, N$, która w tych węzłach przyjmuje zadane wartości $a_0, a_1, \ldots, a_N$, a jej pochodna ma w tych węzłach wartości $b_0, b_1, \ldots, b_N$ dobrane tak, aby funkcja s była klasy C^2. Wyprowadź równanie z niewiadomymi b_0, b_1 opisujące warunek interpolacyjny $s'''(0) = d_0$, w którym liczba d_0 jest dana.

4. Funkcję $f(x) = \cos x$ chcemy przybliżyć w przedziale $[0, \frac{\pi}{2}]$ funkcją interpolacyjną:

a) wielomianem interpolacyjnym stopnia mniejszego niż $2n$ opartym na $2n$ węzłach równoodległych,

b) wielomianem interpolacyjnym stopnia mniejszego niż $2n$ opartym na dwóch węzłach, 0 i $\frac{\pi}{2}$, każdy o krotności n.

c) funkcją sklejaną pierwszego stopnia opartą na $2n$ równoodległych węzłach interpolacyjnych będących jednocześnie węzłami tej funkcji sklejanej.

Znajdź dla każdego z tych przypadków wzór na (możliwie dokładne) oszacowanie błędu aproksymacji w normie maksimum, w zależności od liczby n.

W którym przypadku błąd aproksymacji maleje najszybciej ze wzrostem n? Odpowiedź uzasadnij.

5. Całkę

$$S(f) = \int_{-1}^{1} f(x)\,dx$$

przybliżamy kwadraturą

$$Q(f) = A_0\big(f(-1) + f(1)\big) + A_1\big(f'(-a) + f(a)\big) + A_2 f(0).$$

Dobierz węzeł a oraz współczynniki A_0, A_1, A_2 tak, aby rząd kwadratury Q był największy. Znajdź oszacowanie błędu kwadratury Q dla funkcji f klasy $C^r[-1,1]$, gdzie r oznacza rząd kwadratury.

Egzamin z Matematyki Obliczeniowej, II rok Mat.

(*Ściśle tajne* przed godz. 9:00 16 czerwca 2020.)

Proszę bardzo uważnie przeczytać treść zadań. Bardzo duży wpływ na ocenę będzie miała czytelność rozwiązań i poprawność uzasadnienia każdej odpowiedzi.

1. Niech $a > 0$.

   a) Znajdź wskaźnik uwarunkowania zadania obliczania $\sqrt[k]{a}$ dla ustalonego $k \geqslant 2$.

   b) Podaj funkcję iteracyjną opartą na metodzie Newtona, umożliwiającą otrzymanie ciągu $x_0, x_1, x_2 \ldots$ zbieżnego do $\sqrt[k]{a}$.

   c) Czy dla każdego $x_0 > 0$ ciąg będzie zbieżny do tej samej granicy?

2. Macierz A ma strukturę blokową

$$A = \begin{bmatrix} B & -B & 0 \\ B & B & C \\ 0 & C^T & 0 \end{bmatrix},$$

   przy czym blok B jest macierzą symetryczną i dodatnio określoną, a macierz C ma kolumny liniowo niezależne.

   Podaj algorytm rozwiązywania układu równań z taką macierzą, dostosowany do szczególnej postaci macierzy A, w którego kolejnych krokach są używane metody numeryczne z wykładu, jak najlepiej dopasowane do podzadań rozwiązywanych w tych krokach.

   Wykaż, że układ równań liniowych z macierzą o tej postaci ma jednoznaczne rozwiązanie.

3. Funkcję f klasy $C^2[0,3]$ chcemy przybliżyć funkcją sklejaną s klasy C^1, opisaną przez wielomiany stopnia co najwyżej 2 w przedziałach $[0,1]$ i $[1,3]$, przy czym funkcja ta ma przyjmować te same wartości co f w punktach 0, 1 i 2, a ponadto ma być $s'(1) = f'(1)$.

   a) Znajdź wzory na współczynniki wielomianów opisujących funkcję s w wybranej przez siebie bazie (należy podać tę bazę).

   b) Znajdź górne oszacowanie maksymalnej różnicy funkcji f i s w przedziale $[0,3]$.

4. Rzeczywista macierz symetryczna A ma wartości własne 1, 10, 20 i 100 oraz dodatkowe (nieznane) wartości własne w przedziałach $[2,8]$ i $[30,50]$. Do rozwiązania algebraicznego zagadnienia własnego dla tej macierzy ma być użyta odwrotna metoda potęgowa z parametrem przesunięcia a.

a) Znajdź przedział, z którego należy wybrać parametr a, aby ciąg wektorów wytworzony przez metodę był zbieżny do wektora własnego przynależnego do wartości własnej 10 macierzy A (przy założeniu, że wektor początkowy x_0 nie jest prostopadły do tego wektora).

b) Czy można liczyć na zbieżność ciągu ilorazów Rayleigha (do dowolnej granicy), jeśli $a = 75$? Odpowiedź uzasadnij.

c) Znajdź wskaźnik uwarunkowania macierzy $B = (A - aI)^{-1}$ w normie drugiej dla $a = 19.5$.

5. a) Znajdź wielomian interpolacyjny Hermite'a stopnia co najwyżej 4 dla funkcji f, której wartości i pochodne w pewnych punktach są podane w tabeli:

x	0	2
$f(x)$	1	-7
$f'(x)$	-2	-22
$f''(x)$	6	

Zadanie rozwiąż przy użyciu algorytmu różnic dzielonych, zaczynając od skonstruowania odpowiedniej bazy Newtona. Rozwiązanie przedstaw w tej bazie.

b) Podaj dolne oszacowanie maksymalnej wartości bezwzględnej pochodnej czwartego rzędu funkcji f w przedziale $[0, 2]$ (przy założeniu, że funkcja f jest klasy C^4).

6. Niech funkcja $f \colon \mathbb{R} \to \mathbb{R}$ będzie wielomianem. Niech g będzie wielomianem stopnia co najwyżej n, przyjmującym te same wartości co f w $n+1$ węzłach Czebyszewa skonstruowanych dla tego przedziału.

a) Udowodnij, że jeśli stopień f jest równy $n+1$, to w przestrzeni wielomianów stopnia co najwyżej n najlepszym przybliżeniem funkcji f w normie maksimum jest wielomian g.

b) Znajdź błąd aproksymacji wielomianu f stopnia $n+1$ przez wielomian g.

Rozwiązywanie równań nieliniowych

Rozważamy zadanie znalezienia liczby x, takiej że

$$f(x) = 0,$$

przy czym mamy do dyspozycji podprogram obliczający wartość funkcji f dla argumentu x podanego jako parametr. To dla programu. Natomiast aby taki program napisać, lub wybrać gotowy do rozwiązania konkretnego zadania, zawsze musimy wiedzieć coś więcej o funkcji f. Przede wszystkim trzeba wiedzieć, czy rozwiązanie istnieje. Czy istnieje więcej niż jedno? A może nieskończenie wiele? To oczywiście zależy od funkcji f. Dalej, jeśli rozwiązań jest więcej, to czy mamy znaleźć wszystkie, kilka, czy tylko jedno, obojętnie które, albo spełniające jakiś dodatkowy warunek?

Aby wybrać algorytm, musimy wiedzieć też w jakim zbiorze ta funkcja jest określona i czy jest ciągła, przyda się też wiedza np. czy ciągła jest jej pochodna rzędu 1, 2 i być może dalsze. W niektórych metodach oprócz podprogramu obliczającego $f(x)$ będzie też potrzebny podprogram obliczający $f'(x)$, a nawet dalsze pochodne.

Metoda Newtona

Niech A oznacza ograniczony przedział domknięty, w którym jest określona funkcja rzeczywista f klasy C^2. Chcemy znaleźć w tym przedziale miejsce zerowe funkcji f (założymy, że istnieje i jest tylko jedno, w każdym praktycznym zastosowaniu to założenie oczywiście trzeba sprawdzić).

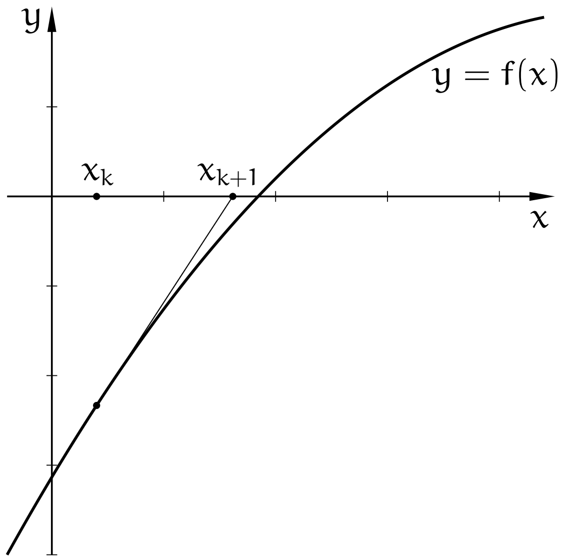
Metoda Newtona (często w literaturze nazywana metodą stycznych lub metodą Newtona-Raphsona, wersja współczesna różni się od wersji podanych przez nich obu) jest następująca: wybieramy liczbę x_0, która jest przybliżeniem miejsca zerowego funkcji f, a następnie konstruujemy rekurencyjnie elementy ciągu $x_1, x_2, \ldots$, w taki sposób: mając x_k, określamy wielomian

$$w_k(x) = f(x_k) + f'(x_k)(x - x_k).$$

Znajdujemy miejsce zerowe wielomianu w_k i przyjmujemy, że jest to x_{k+1}. Mamy stąd formułę

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

Interpretacja geometryczna jest taka: wykres funkcji f jest gładką krzywą, przechodzącą przez punkt $(x_k, f(x_k))$. Konstruujemy prostą styczną do wykresu w tym punkcie i przyjmujemy za x_{k+1} punkt przecięcia stycznej z osią x.



Zbadamy, jakie warunki wystarczy spełnić, aby ciąg $(x_k)_{k\in\mathbb{N}}$ zbiegał do rozwiązania, które oznaczymy literą α. Przede wszystkim zauważmy, że w żadnym punkcie tego ciągu pochodna funkcji f nie może być zerowa. Naturalne jest założenie, że w przedziale A pochodna znaku nie zmienia, co więcej, zachodzi nierówność $|f'(x)| \geqslant K_1$ dla pewnej stałej $K_1 > 0$. Dalej, ponieważ f jest klasy $C^2(A)$, istnieje stała M_2, taka że $|f''(x)| \leqslant M_2$ dla każdego $x \in A$.

Napiszemy wzór Taylora:

$$f(x+h) = \frac{f(x)}{0!} + \frac{f'(x)}{1!}h + \frac{f''(\xi)}{2!}h^2.$$

Rozumiemy go tak: jeśli liczby x oraz $x+h$ należą do przedziału A, w którym funkcja f jest klasy C^2, to istnieje liczba ξ, leżąca pomiędzy x oraz $x+h$, taka że powyższa równość zachodzi.

Oznaczmy $\varepsilon_k = x_k - \alpha$ — jest to (bezwzględny) błąd aproksymacji rozwiązania przez k-ty element ciągu. Na podstawie wzoru Taylora piszemy

$$0 = f(\alpha) = f(x_k) + f'(x_k)(\alpha - x_k) + \frac{1}{2}f''(\xi_k)(\alpha - x_k)^2.$$

Liczba ξ_k leży między α i x_k. Dzielimy strony przez $f'(x_k)$:

$$0 = \frac{f(x_k)}{f'(x_k)} + \alpha - x_k + \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2 = \frac{f(x_k)}{f'(x_k)} + \alpha - x_{k+1} + x_{k+1} - x_k + \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2.$$

Ponieważ $x_{k+1} - x_k = -\frac{f(x_k)}{f'(x_k)}$, mamy stąd

$$\varepsilon_{k+1} = \frac{f''(\xi_k)}{2f'(x_k)}\varepsilon_k^2. \tag{*}$$

Możemy oszacować

$$|\varepsilon_{k+1}| \leqslant \frac{M_2}{2K_1}|\varepsilon_k|^2.$$

Aby zachodziła nierówność $|\varepsilon_{k+1}| < |\varepsilon_k|$, wystarczy, że $\frac{M_2}{2K_1}|\varepsilon_k| < 1$, czyli

$$|\varepsilon_k| < \frac{2K_1}{M_2}.$$

Jeśli błąd przybliżenia rozwiązania przez punkt x_0, z którego zaczynamy, spełnia tę nierówność, to każdy następny błąd ma mniejszą wartość bezwzględną niż poprzedni, co więcej, ciąg błędów zbiega do zera.

Mamy zatem warunek dostateczny zbieżności metody, ale zbadajmy jeszcze szybkość tej zbieżności. Dowolną ustaloną liczbę $b > 1$ przyjmiemy za podstawę logarytmu, którego tu użyjemy. Oznaczmy $a_k = \log|\varepsilon_k|$ oraz $g(k) = \log\left|\frac{f''(\xi_{k-1})}{2f'(x_{k-1})}\right|$. Na podstawie równości (*) możemy napisać równanie różnicowe

$$a_k = 2a_{k-1} + g(k).$$

Niech $G = \log\frac{M_2}{2K_1}$. Jeśli rozważymy równanie uproszczone,

$$\tilde{a}_k = 2\tilde{a}_{k-1} + G,$$

dla którego przyjmiemy $\tilde{a}_0 = a_0 < -G$, to dla każdego k mamy

$$a_k \leqslant \tilde{a}_k = (a_0 + G)\cdot 2^k - G.$$

Ciąg $(\tilde{a}_k)_{k\in\mathbb{N}}$ dąży wykładniczo do $-\infty$, a ciąg $(a_k)_{k\in\mathbb{N}}$ dąży do $-\infty$ co najmniej tak samo szybko.

Takim samym sposobem możemy znaleźć dolne oszacowanie błędów, tj. liczb $|\varepsilon_k|$. Niech zatem M_1 oznacza stałą, taką że dla każdego $x \in A$ zachodzi nierówność $|f'(x)| \leqslant M_1$. Przypuśćmy, że istnieje też stała dodatnia K_2, taka że $|f''(x)| \geqslant K_2$ dla każdego $x \in A$.[1] Mając takie stałe, możemy napisać nierówność

$$|\varepsilon_{k+1}| \geqslant \frac{K_2}{2M_1}|\varepsilon_k|^2,$$

[1]Ponieważ funkcja f jest klasy $C^2(A)$, istnienie takiej stałej oznacza, że funkcja f jest albo wypukła, albo wklęsła.

a następnie oznaczyć $\hat{G} = \log \frac{K_2}{2M_1}$ i oszacować z dołu ciąg $(a_k)_{k\in\mathbb{N}}$ przez ciąg $(\hat{a}_k)_{k\in\mathbb{N}}$ będący rozwiązaniem równania różnicowego

$$\hat{a}_k = 2\hat{a}_{k-1} + \hat{G}.$$

Jeśli zatem funkcja f spełnia *wszystkie* przyjęte założenia i mamy odpowiedni punkt startowy x_0, to zachodzą nierówności

$$(a_0 + \hat{G}) \cdot 2^k - \hat{G} \leqslant a_k \leqslant (a_0 + G) \cdot 2^k - G.$$

Równoważnie, mamy

$$b^{(a_0+\hat{G})\cdot 2^k - \hat{G}} \leqslant b^{a_k} \leqslant b^{(a_0+G)\cdot 2^k - G}.$$

Uporządkowanie wyrażeń powiązanych powyższymi nierównościami kończy dowód następującego twierdzenia:

<u>Twierdzenie.</u> *Jeśli funkcja* f *jest klasy* C^2 *w przedziale* A, *ma w nim miejsce zerowe* α *i istnieją stałe* K_1 *i* M_2, *takie że* $0 < K_1 \leqslant |f'(x)|$ *oraz* $|f''(x)| \leqslant M_2$ *dla każdego* $x \in A$, $x_0 \in A$ *oraz* $|x_0 - \alpha| < \frac{2K_1}{M_2}$, *to metoda Newtona startująca z punktu* x_0 *wytwarza ciąg* $(x_k)_{k\in\mathbb{N}}$ *zbieżny do* α, *przy czym*

$$|x_{k+1} - \alpha| \leqslant \frac{M_2}{2K_1}|x_k - \alpha|^2.$$

Jeśli ponadto istnieją stałe M_1 *i* $K_2 > 0$, *takie że dla każdego* $x \in A$ $|f'(x)| < M_1$ *oraz* $0 < K_2 \leqslant |f''(x)|$, *to*

$$\frac{K_2}{2M_1}|x_k - \alpha|^2 \leqslant |x_{k+1} - \alpha|.$$

<u>Wniosek.</u> *Jeśli założenia twierdzenia są spełnione, to istnieją dodatnie liczby* c, d, C, D, *takie że dla każdego* k *zachodzą nierówności*

$$c(d|x_0 - \alpha|)^{2^k} \leqslant |x_k - \alpha| \leqslant C(D|x_0 - \alpha|)^{2^k}.$$

Z twierdzenia wynika, że jeśli x_k jest przybliżeniem rozwiązania, które ma n cyfr dokładnych, to x_{k+1} będzie mieć w przybliżeniu $2n$ cyfr dokładnych. Zatem zbieżność jest bardzo szybka. Znając oszacowanie $|\varepsilon_0|$ i G oraz tolerancję błędu, można oszacować liczbę iteracji wystarczającą do otrzymania rozwiązania z błędem w granicach tej tolerancji.

<u>Uwaga.</u> Można udowodnić zbieżność metody Newtona przy słabszych założeniach, np. że funkcja f jest tylko klasy C^1, ale jej pochodna spełnia warunek Lipschitza.

Podstawowe pojęcia w numerycznym rozwiązywaniu równań

Równania nieliniowe (i ich układy) są rozwiązywane za pomocą metod iteracyjnych — przykładem jest przedstawiona wyżej metoda Newtona. Powód jest taki, że rozwiązania na ogół nie dają się wyrazić za pomocą czterech działań algebraicznych i to dotyczy nawet równań drugiego stopnia (oczywiście, mamy do dyspozycji pierwiastek kwadratowy, ale jego też oblicza się iteracyjnie, za pomocą odpowiedniego mikroprogramu procesora). Aby mieć ogólne spojrzenie na metody iteracyjne, wprowadzimy kilka pojęć.

Funkcja iteracyjna jest to funkcja, która elementowi x_k, będącemu przybliżeniem rozwiązania, przyporządkowuje kolejne przybliżenie, x_{k+1}. W metodzie Newtona funkcja iteracyjna jest określona wzorem

$$\varphi_N(x) = x - \frac{f(x)}{f'(x)}.$$

Jest jasne, że funkcja iteracyjna powinna być tak skonstruowana, aby rozwiązanie α było jej punktem stałym, tj. aby było $\varphi(\alpha) = \alpha$.

Istnieje nieskończenie wiele możliwości „przerobienia” równania $f(x) = 0$ na równoważne równanie $x = \varphi(x)$. W najprostszym przypadku możemy wziąć

$$\varphi(x) = x - \tau f(x),$$

z jakimś parametrem rzeczywistym τ. Oczywiście, nie zawsze otrzymana w ten sposób funkcja φ prowadzi do otrzymania ciągu zbieżnego. Aby zbieżność miała miejsce, trzeba, by w otoczeniu rozwiązania α funkcja φ była odwzorowaniem zwężającym (może mieć np. pochodną o wartości bezwzględnej mniejszej od 1).

Funkcje iteracyjne dla pewnych metod są bardziej skomplikowane. Po pierwsze, argumentem funkcji iteracyjnej oprócz ostatniego przybliżenia może być także jedno lub więcej poprzednich (czasami takie metody nazywa się metodami z pamięcią). Na przykład w metodzie siecznych, o której będzie mowa dalej, potrzebne są dwa ostatnie przybliżenia, które nie mogą być jednakowe. Funkcja iteracyjna ma w tym przypadku postać

$$\varphi_S(x, y) = x - \frac{f(x)}{f[x, y]}, \quad \text{gdzie} \quad f[x, y] = \frac{f(x) - f(y)}{x - y},$$

zatem w kolejnych iteracjach obliczamy $x_{k+1} = \varphi_S(x_k, x_{k-1})$. Wreszcie, funkcja iteracyjna może w jawny sposób zależeć od numeru iteracji, k — w tym przypadku mówimy o metodzie niestacjonarnej.

Kula zbieżności rozwiązania α jest to największa kula B o środku α (w przypadku równań z jedną niewiadomą jest to przedział symetryczny względem α), taka że jeśli wybierzemy dowolny punkt startowy x_0 wewnątrz tej kuli, to ciąg $(x_k)_{k\in\mathbb{N}}$ zbiega do α. Znalezienie kuli zbieżności jest na ogół bardzo trudne, więc tego nie robimy, ale na podstawie własności funkcji f i definicji metody możemy szacować jej promień r. Na przykład, dla metody Newtona znaleźliśmy oszacowanie $r \geqslant \frac{2K_1}{M_2}$. Jest oczywiste, że jeśli równanie ma kilka rozwiązań, to każde z nich ma własną kulę zbieżności i wszystkie te kule są rozłączne. Kule zbieżności pewnych rozwiązań mogą być zbiorem pustym — może się zdarzyć, że dana metoda nie jest w stanie takich rozwiązań znaleźć. Warto natomiast zwrócić uwagę, że jeśli punkt startowy nie należy do kuli zbieżności żadnego rozwiązania, to metoda może znaleźć rozwiązanie, jeśli otrzymany po pewnej liczbie iteracji punkt „wpadł" do kuli zbieżności. Tylko, że *nie należy* liczyć na taki przypadek.

Twierdzeniem o ogromnym znaczeniu praktycznym, a zwłaszcza w metodach numerycznych, jest twierdzenie Banacha o punkcie stałym: *jeśli zbiór* X *z metryką* ρ *jest zupełną przestrzenią metryczną, a funkcja* $\varphi\colon X \to X$ *jest przekształceniem zwężającym (tj. istnieje stała* $L < 1$*, taka że* $\forall_{a,b\in X}\, \rho(\varphi(a), \varphi(b)) \leqslant L\rho(a, b)$*), to funkcja* φ *ma jeden punkt stały w zbiorze* X. Wykazanie, że metoda działa, tj. wytwarza ciąg zbieżny do rozwiązania, często sprowadza się do znalezienia (wykazania istnienia lub oszacowania promienia) kuli X, zawartej w kuli zbieżności, w której funkcja iteracyjna φ jest przekształceniem zwężającym.

Wykładnik zbieżności metody opisuje asymptotyczną szybkość zbieżności ciągu $(x_k)_{k\in\mathbb{N}}$ do rozwiązania. Przeprowadzony rachunek dla metody Newtona dowiódł, że *jeśli funkcja* f *spełnia uczynione założenia*, to wykładnik zbieżności jest równy 2. Formalna definicja jest taka: wykładnik zbieżności metody jest to największa liczba p, taka że istnieją stałe K i $C < +\infty$, takie że dla każdego $k \geqslant K$ zachodzi nierówność

$$|\varepsilon_{k+1}| \leqslant C|\varepsilon_k|^p, \quad \text{czyli} \quad \log|\varepsilon_{k+1}| \leqslant \log C + p\log|\varepsilon_k|.$$

Wykładnik zbieżności powinien być większy lub równy 1, przy czym jeśli $p = 1$, to oczywiście musi być $C < 1$. Przykładem metody o wykładniku zbieżności 1 jest metoda bisekcji: w każdej iteracji otrzymujemy przybliżenie rozwiązania z oszacowaniem błędu mniejszym o połowę. Również metoda Newtona ma wykładnik zbieżności 1, jeśli nie jest spełnione założenie, że pochodna funkcji f w otoczeniu rozwiązania jest niezerowa. Jeśli $p > 1$, to dla ustalonego K istnieją stałe c, d, r i s, takie że dla każdego $k > K$

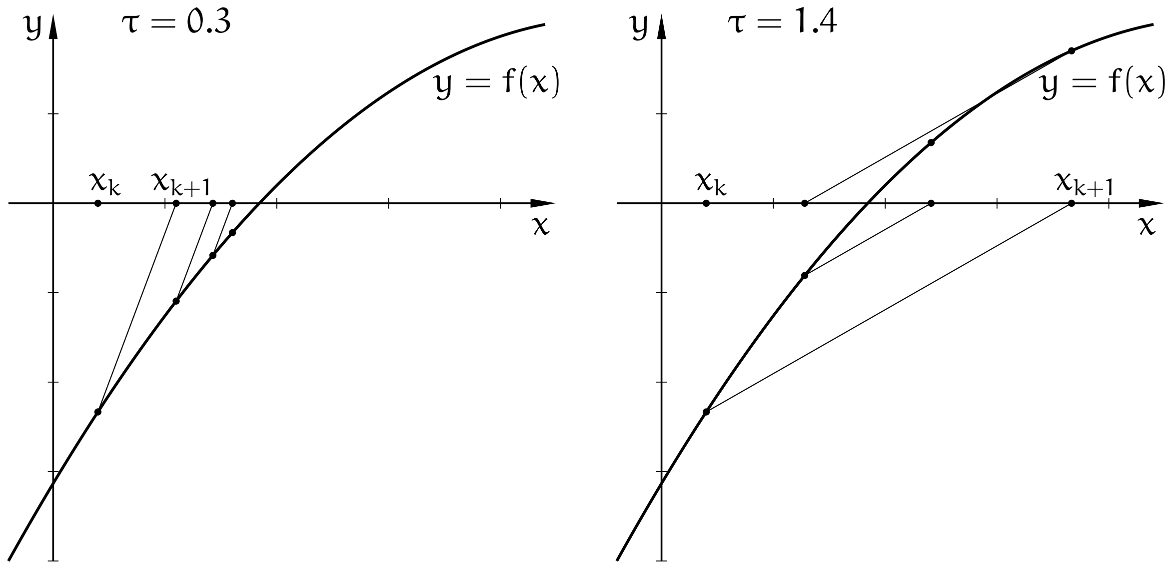
$$\log|\varepsilon_k| \leqslant c + (\log|\varepsilon_K| + d)p^{k-K}, \quad \text{czyli} \quad |\varepsilon_k| \leqslant r(s|\varepsilon_K|)^{p^{k-K}}.$$

Ostatnie przedstawiane tu pojęcie podstawowe to maksymalna graniczna dokładność. Analiza metody Newtona była przeprowadzona przy założeniu, że nie ma błędów w obliczeniach, tj. zarówno wartości funkcji f i pochodnej w x_k są obliczane dokładnie, jak i w końcowych działaniach obliczenia wartości funkcji iteracyjnej nie ma błędów. Obliczenia wykonujemy jednak z błędami zaokrągleń, które ograniczają możliwą do uzyskania dokładność rozwiązania. Wartość funkcji f jest obliczana z pewnym błędem, dalsze działania w obliczaniu funkcji iteracyjnej też są niedokładne. Za rozwiązanie metoda może zatem przyjąć dowolny punkt przedziału, w którym błąd obliczonej wartości funkcji f jest większy lub równy 100%. Jeśli pochodna funkcji jest bliska 0, to ten przedział może być długi.

Metoda iteracji prostej

Metoda iteracji prostej polega na iterowaniu funkcji $\varphi(x) = x - \tau f(x)$. Zatem, przyjmujemy punkt początkowy x_0 i obliczamy

$$x_{k+1} = x_k - \tau f(x_k), \quad k = 0, 1, \ldots .$$



Parametr τ trzeba dobrać tak, aby osiągnąć zbieżność. Załóżmy, że funkcja f jest klasy C^1 w otoczeniu miejsca zerowego α funkcji f i że $f'(\alpha) \neq 0$. Oznaczmy $\varepsilon_k = x_k - \alpha$. Dla każdego k zachodzi równość

$$\varepsilon_{k+1} = \varepsilon_k - \tau f(x_k). \tag{*}$$

Istnieje liczba ξ_k, położona między rozwiązaniem α i jego przybliżeniem x_k, taka że

$$f'(\xi_k) = \frac{f(x_k) - f(\alpha)}{x_k - \alpha} = \frac{f(x_k)}{\varepsilon_k}.$$

Dzieląc strony (*) otrzymamy równość

$$\frac{\varepsilon_{k+1}}{\varepsilon_k} = 1 - \tau f'(\xi_k).$$

Lepsze przybliżenie rozwiązania otrzymamy, jeśli

$$\left|\frac{\varepsilon_{k+1}}{\varepsilon_k}\right| = |1 - \tau f'(\xi_k)| < 1, \quad \text{czyli} \quad 0 < \tau f'(\xi_k) < 2.$$

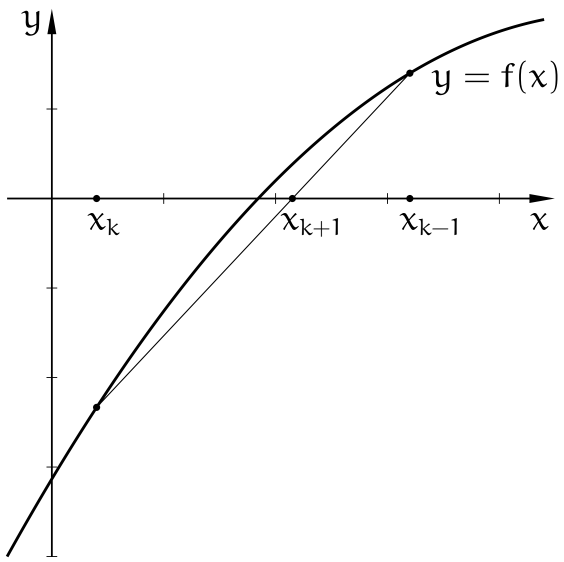
Ponieważ f' jest różna od 0 w rozwiązaniu α i ciągła, istnieje otoczenie A i stałe K_1 i M_1, takie że $K_1 \leqslant |f'(x)| \leqslant M_1$ dla każdego $x \in A$. Jeśli $x_k \in A$, to warunek dostateczny zmniejszenia błędu w kolejnym kroku ma zatem postać

$$\operatorname{sgn} \tau = \operatorname{sgn} f'(x_k) \quad \text{oraz} \quad |\tau| M_1 < 2.$$

Warunek dostateczny zbieżności całego ciągu $(x_k)_k$ to spełnienie powyższej nierówności i na przykład zawieranie $[\alpha - |\varepsilon_k|, \alpha + |\varepsilon_k|] \subset A$. Znajomość stałych K_1 i M_1 umożliwia wybranie „dobrej” wartości parametru τ.

Metoda siecznych

Wadą metody Newtona jest konieczność obliczania pochodnej funkcji f. Metoda siecznych jest modyfikacją metody Newtona, w której pochodna została zastąpiona przez różnicę dzieloną (albo iloraz różnicowy, jak kto woli), czyli pewne przybliżenie pochodnej. Mając *dwa różne* przybliżenia rozwiązania, x_k i x_{k-1}, prowadzimy prostą przez punkty $(x_k, f(x_k))$ i $(x_{k-1}, f(x_{k-1}))$. Prosta ta przecina (siecze) wykres funkcji f w tych punktach, i w tym sensie jest jego sieczną.



Skonstruowana sieczna jest wykresem wielomianu pierwszego stopnia. Punkt x_{k+1} jest miejscem zerowym tego wielomianu. W metodzie siecznych należy podać dwa początkowe przybliżenia rozwiązania, x_0 i x_1, a następnie w każdej iteracji obliczać

$$x_{k+1} = x_k - \frac{f(x_k)}{f[x_k, x_{k-1}]}, \quad \text{gdzie} \quad f[x_k, x_{k-1}] = \frac{f(x_k) - f(x_{k-1})}{x_k - x_{k-1}}.$$

Aby dokonać analizy metody siecznych, użyjemy pewnego uogólnienia wzoru Taylora:

$$f(z) = f(x) + f[x, y](z - x) + \frac{f''(\xi)}{2!}(z - x)(z - y).$$

Wzór ten jest szczególnym przypadkiem wzoru opisującego resztę interpolacyjną Hermite'a, który podam (z dowodem) na jednym z dalszych wykładów. Podany wyżej wzór rozumiemy w ten sposób, że jeśli liczby x, y, z leżą w przedziale A, w którym funkcja f jest klasy C^2, to istnieje $\xi \in A$, takie że podana wyżej równość zachodzi (liczba ξ leży między najmniejszą i największą spośród tych trzech liczb).

Jak poprzednio, α oznacza poszukiwane rozwiązanie, zaś $\varepsilon_k = x_k - \alpha$. Liczymy

$$0 = f(\alpha) = f(x_k) + f[x_k, x_{k-1}](\alpha - x_k) + \frac{f''(\xi_k)}{2}(\alpha - x_k)(\alpha - x_{k-1})$$

i dzielimy stronami przez $f[x_k, x_{k-1}]$:

$$0 = \underbrace{\frac{f(x_k)}{f[x_k, x_{k-1}]}} + \alpha - x_{k+1} + \underbrace{x_{k+1} - x_k} + \frac{f''(\xi_k)}{2f[x_k, x_{k-1}]}(\alpha - x_k)(\alpha - x_{k-1}),$$

skąd, po skróceniu podkreślonych składników, otrzymujemy

$$0 = \alpha - x_{k+1} + \frac{f''(\xi_k)}{2f[x_k, x_{k-1}]}(\alpha - x_k)(\alpha - x_{k-1}),$$

a po uporządkowaniu i uwzględnieniu faktu, że istnieje liczba η_k położona między x_k i x_{k-1}, taka że $f[x_k, x_{k-1}] = f'(\eta_k)$, mamy stąd równość

$$\varepsilon_{k+1} = \frac{f''(\xi_k)}{2f'(\eta_k)}\varepsilon_k\varepsilon_{k-1}. \tag{**}$$

Jeśli, jak poprzednio, możemy oszacować $|f'(x)| \geqslant K_1 > 0$ i $|f''(x)| \leqslant M_2$ dla każdego $x \in A$, to mamy

$$|\varepsilon_{k+1}| \leqslant \frac{M_2}{2K_1}|\varepsilon_k||\varepsilon_{k-1}|.$$

Jeśli oba błędy, ε_k i ε_{k-1}, mają wartości bezwzględne mniejsze niż $\frac{2K_1}{M_2}$, to wartości bezwzględne kolejnych błędów będą coraz mniejsze — w ten sposób mamy oszacowany promień kuli zbieżności. Zbadajmy jeszcze rząd zbieżności. W tym

celu oznaczmy $a_k = \log|\varepsilon_k|$ oraz $g(k) = \log\left|\frac{f''(\xi_{k-1})}{f'(\eta_{k-1})}\right|$ i $G = \log\left|\frac{M_2}{2K_1}\right|$.
Na podstawie (**) możemy napisać równanie różnicowe drugiego rzędu,

$$a_k = a_{k-1} + a_{k-2} + g(k),$$

i jego uproszczoną wersję

$$\tilde{a}_k = \tilde{a}_{k-1} + \tilde{a}_{k-2} + G.$$

Dla ustalonych wyrazów początkowych, $\tilde{a}_0 = a_0$ i $\tilde{a}_1 = a_1$, istnieją liczby c, d, e (mniejsza o dokładne wzory, ale zachęcam do ich znalezienia dla wprawy), takie że

$$\tilde{a}_k = c\lambda_1^k + d\lambda_2^k + e, \quad \text{gdzie} \quad \lambda_1 = \frac{1-\sqrt{5}}{2},\ \lambda_2 = \frac{1+\sqrt{5}}{2},$$

i jeśli liczby a_0 i a_1 są dostatecznie małe, to $d < 0$. Składnik $d\lambda_2^k$ dominuje i wtedy ciąg $(\tilde{a}_k)_{k\in\mathbb{N}}$ zbiega wykładniczo do $-\infty$, zachodzi też nierówność $a_k \leqslant \tilde{a}_k$ dla każdego k. Polecam jako ćwiczenie dokończenie dowodu twierdzenia

Twierdzenie. *Jeśli funkcja f jest klasy C^2 w przedziale A, ma w nim miejsce zerowe α i istnieją stałe K_1 i M_2, takie że $0 < K_1 \leqslant |f'(x)|$ oraz $|f''(x)| \leqslant M_2$ dla każdego $x \in A$, $x_0, x_1 \in A$, $x_0 \neq x_1$ oraz $|x_0 - \alpha|, |x_1 - \alpha| < \frac{2K_1}{M_2}$, to metoda siecznych startująca z punktów x_0, x_1 wytwarza ciąg $(x_k)_{k\in\mathbb{N}}$ zbieżny do α, a ponadto istnieje $H > 0$ takie że dla każdego k zachodzi nierówność*

$$|x_{k+1} - \alpha| \leqslant H|x_k - \alpha|^{\lambda_2}.$$

Jeśli ponadto istnieją stałe K_2 i M_1, takie że $|f'(x)| \leqslant M_1$ oraz $0 < K_2 \leqslant |f''(x)|$ dla każdego $x \in A$, to istnieje $h > 0$, takie że

$$h|x_k - \alpha|^{\lambda_2} \leqslant |x_{k+1} - \alpha|$$

Wniosek. *Jeśli założenia twierdzenia są spełnione, to istnieją stałe dodatnie r, s, R, S, takie że*

$$r(s|x_0 - \alpha|)^{\lambda_2^k} \leqslant |x_k - \alpha| \leqslant R(S|x_0 - \alpha|)^{\lambda_2^k}$$

Tak więc dla dostatecznie dużych k, jeśli przybliżenie x_k rozwiązania α ma n cyfr dokładnych, to przybliżenie x_{k+1} będzie ich miało około $\lambda_2 n$. Wykładnik zbieżności metody siecznych, $\lambda_2 \approx 1.618$, jest ułamkiem.

Metoda siecznych ma mniejszy wykładnik zbieżności niż metoda Newtona, ale jedna jej iteracja jest tańsza — odpada obliczanie pochodnej. Okazuje się, że jeśli zadamy tolerancję ε dopuszczalnego błędu, to metoda siecznych może znaleźć dostatecznie dokładne rozwiązanie szybciej (w większej liczbie iteracji, z których każda zajmuje mniej czasu). Z tego punktu widzenia, jeśli koszt obliczania różnicy dzielonej uznamy za nieistotny, to metoda Newtona jest bardziej opłacalna, gdy koszt obliczania pochodnej nie przewyższa ok. $\frac{\log 2}{\log \lambda_2} - 1 \approx 0.44$ kosztu obliczania wartości funkcji f.

Metoda Newtona dla układu równań

Rozważamy teraz zadanie znalezienia wspólnego miejsca zerowego n rzeczywistych funkcji skalarnych, których argumentami jest n zmiennych rzeczywistych. Możemy zatem napisać układ w postaci rozwiniętej:

$$\begin{cases} f_1(x_1, \dots, x_n) = 0, \\ \qquad\vdots \\ f_n(x_1, \dots, x_n) = 0, \end{cases}$$

lub „zwiniętej”

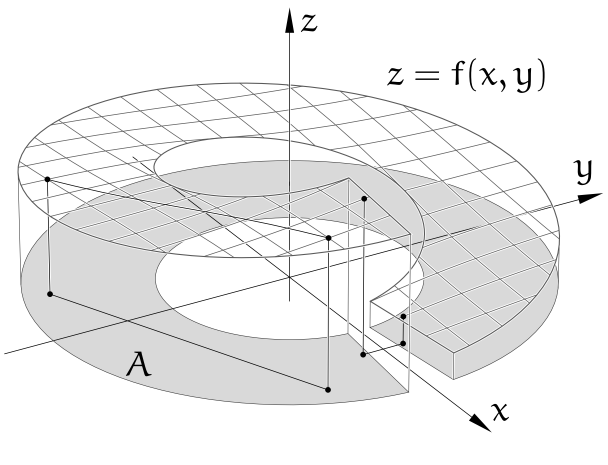
$$\mathbf{f}(\mathbf{x}) = \mathbf{0}.$$

Zakładamy, że funkcja $\mathbf{f}$ jest określona w pewnym obszarze A przestrzeni $\mathbb{R}^n$ i ma wartości w $\mathbb{R}^n$.

Niech $\mathbf{h} = [h_1, \dots, h_n]^T$. Dla funkcji *skalarnej* f_i klasy $C^2(A)$, możemy napisać wzór Taylora:

$$f_i(\mathbf{x} + \mathbf{h}) = \frac{1}{0!} f_i(\mathbf{x}) + \frac{1}{1!} Df_i|_{\mathbf{x}}(\mathbf{h}) + \frac{1}{2!} D^2 f_i|_{\boldsymbol{\xi}_i}(\mathbf{h}, \mathbf{h}).$$

Rozumiemy go tak: *jeśli* obszar A *zawiera odcinek* o końcach $\mathbf{x}$ i $\mathbf{x} + \mathbf{h}$, to istnieje punkt $\boldsymbol{\xi}_i$ na tym odcinku, taki że powyższa równość zachodzi (przykład na rysunku).



Symbol $\mathrm{D}f_i|_x$ oznacza <u>różniczkę</u> funkcji f_i w punkcie $\boldsymbol{x}$, czyli przekształcenie liniowe, które dowolnemu wektorowi $\boldsymbol{h}$ przyporządkowuje liczbę

$$\mathrm{D}f_i|_x(\boldsymbol{h}) = \frac{\partial f_i}{\partial x_1}\Big|_x h_1 + \cdots + \frac{\partial f_i}{\partial x_n}\Big|_x h_n.$$

Wartością tego przekształcenia jest zatem iloczyn skalarny gradientu funkcji f_i w punkcie $\boldsymbol{x}$ i wektora $\boldsymbol{h}$. Symbol $\mathrm{D}^2 f_i|_{\xi_i}$ oznacza <u>różniczkę drugiego rzędu</u>, tj. przekształcenie *dwuliniowe*, którego wartością dla pary wektorów $(\boldsymbol{g}, \boldsymbol{h})$ jest liczba

$$\mathrm{D}^2 f_i|_{\xi_i}(\boldsymbol{g}, \boldsymbol{h}) = \sum_{j=1}^{n}\sum_{k=1}^{n} \frac{\partial^2 f_i}{\partial x_j \partial x_k}\Big|_{\xi_i} g_j h_k.$$

Drobny kłopot (o którym *nie należy* zapominać) jest taki, że punkt $\boldsymbol{\xi}_i$ dla każdego i może być inny, dlatego nie można tak prosto zapisać odpowiedniego wzoru dla funkcji wektorowej $\boldsymbol{f}$. Niemniej, ze wzoru Taylora wynika, że jeśli obszar A zawiera odcinek $\overline{\boldsymbol{x}, \boldsymbol{x}+\boldsymbol{h}}$, to dla wektorowej funkcji $\boldsymbol{f}$ klasy $C^2(A)$ zachodzi równość

$$\boldsymbol{f}(\boldsymbol{x}+\boldsymbol{h}) = \boldsymbol{f}(\boldsymbol{x}) + \mathrm{D}\boldsymbol{f}|_x(\boldsymbol{h}) + \boldsymbol{r}, \tag{$^{**}_{*}$}$$

przy czym $\mathrm{D}\boldsymbol{f}|_x$ jest różniczką przekształcenia $\boldsymbol{f}$ w punkcie $\boldsymbol{x}$, a ponadto istnieje macierz B (zależna od $\boldsymbol{x}$ i $\boldsymbol{h}$) o wymiarach $n \times n$ i współczynnikach *wektorowych* $\boldsymbol{b}_{jl} = \left[\frac{\partial^2 f_1}{\partial x_j \partial x_l}\big|_{\xi_1}, \ldots, \frac{\partial^2 f_n}{\partial x_j \partial x_l}\big|_{\xi_n}\right]^T \in \mathbb{R}^n$, taka że reszta we wzorze ($^{**}_{*}$) jest równa

$$\boldsymbol{r} = \boldsymbol{h}^T B \boldsymbol{h} = \sum_{j=1}^{n}\sum_{l=1}^{n} \boldsymbol{b}_{jl} h_j h_l, \tag{$^{**}_{**}$}$$

i spełnia oszacowanie

$$\|\boldsymbol{r}\| \leqslant \frac{M_2}{2}\|\boldsymbol{h}\|^2$$

dla pewnej stałej M_2 (stała ta jest określona przez pochodne drugiego rzędu funkcji f_i w obszarze A i przez używaną normę).

<u>Metoda Newtona</u> polega na tym, że mając przybliżenie $\boldsymbol{x}_k$ rozwiązania $\boldsymbol{\alpha}$, konstruujemy przekształcenie afiniczne $\mathbb{R}^n \to \mathbb{R}^n$, określone przez pierwsze dwa składniki po prawej stronie wzoru ($^{**}_{*}$), a następnie przyjmujemy za $\boldsymbol{x}_{k+1}$ miejsce zerowe tego przekształcenia. Czyli

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - (\mathrm{D}\boldsymbol{f}|_{x_k})^{-1}\big(\boldsymbol{f}(\boldsymbol{x}_k)\big).$$

Aby obliczyć $\boldsymbol{x}_{k+1}$, należy obliczyć wektor $\boldsymbol{f}_k = \boldsymbol{f}(\boldsymbol{x}_k)$ oraz macierz pochodnych cząstkowych pierwszego rzędu

$$J_k = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}\big|_{x_k} & \cdots & \frac{\partial f_1}{\partial x_n}\big|_{x_k} \\ \vdots & & \vdots \\ \frac{\partial f_n}{\partial x_1}\big|_{x_k} & \cdots & \frac{\partial f_n}{\partial x_n}\big|_{x_k} \end{bmatrix},$$

która reprezentuje różniczkę funkcji $\mathbf{f}$ w punkcie $\mathbf{x}_k$, a następnie rozwiązać układ równań liniowych

$$J_k\boldsymbol{\delta} = -\mathbf{f}_k$$

i obliczyć $\mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\delta}$. Oczywiście, aby to obliczenie było wykonalne, macierz J_k musi być nieosobliwa.

Ilustrację jednego kroku metody dla pewnego układu dwóch równań z dwiema niewiadomymi mamy na rysunkach. Każda z dwóch funkcji skalarnych, f_1 i f_2, jest przybliżana przez wielomian pierwszego stopnia, który w punkcie $\mathbf{x}_k$ ma tę samą wartość i pochodne cząstkowe, co ta funkcja. Punkt $\mathbf{x}_{k+1}$ jest przecięciem zbiorów miejsc zerowych tych wielomianów.

Przyjmiemy założenie, że istnieje taka stała K_1, że dla każdego punktu $\mathbf{x}$ w rozpatrywanym obszarze A różniczka przekształcenia $\mathbf{f}$ spełnia warunek $\|(D\mathbf{f}|_x)^{-1}\| \leqslant K_1^{-1}$. Zatem, dla $\mathbf{x}_k \in A$ jest $\|J_k^{-1}\| \leqslant K_1^{-1}$. Na podstawie wzorów $(^{**}_{\ *})$ i $(^{**}_{**})$, mamy

$$0 = \mathbf{f}(\boldsymbol{\alpha}) = \mathbf{f}(\mathbf{x}_k) + J_k(\boldsymbol{\alpha} - \mathbf{x}_k) + (\boldsymbol{\alpha} - \mathbf{x}_k)^T B_k(\boldsymbol{\alpha} - \mathbf{x}_k),$$

Dalej postępujemy identycznie, jak w przypadku skalarnym. Oznaczamy $\boldsymbol{\varepsilon}_k = \mathbf{x}_k - \boldsymbol{\alpha}$. Strony równości mnożymy[2] przez J_k^{-1}, oraz odejmujemy i dodajemy $\mathbf{x}_{k+1}$ i skracamy:

$$0 = J_k^{-1}\mathbf{f}(\mathbf{x}_k) + \boldsymbol{\alpha} - \mathbf{x}_{k+1} + \mathbf{x}_{k+1} - \mathbf{x}_k + J_k^{-1}\big(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\big) = \boldsymbol{\alpha} - \mathbf{x}_{k+1} + J_k^{-1}\big(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\big).$$

Stąd wielkość błędu kolejnego przybliżenia rozwiązania,

$$\boldsymbol{\varepsilon}_{k+1} = J_k^{-1}\big(\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k\big),$$

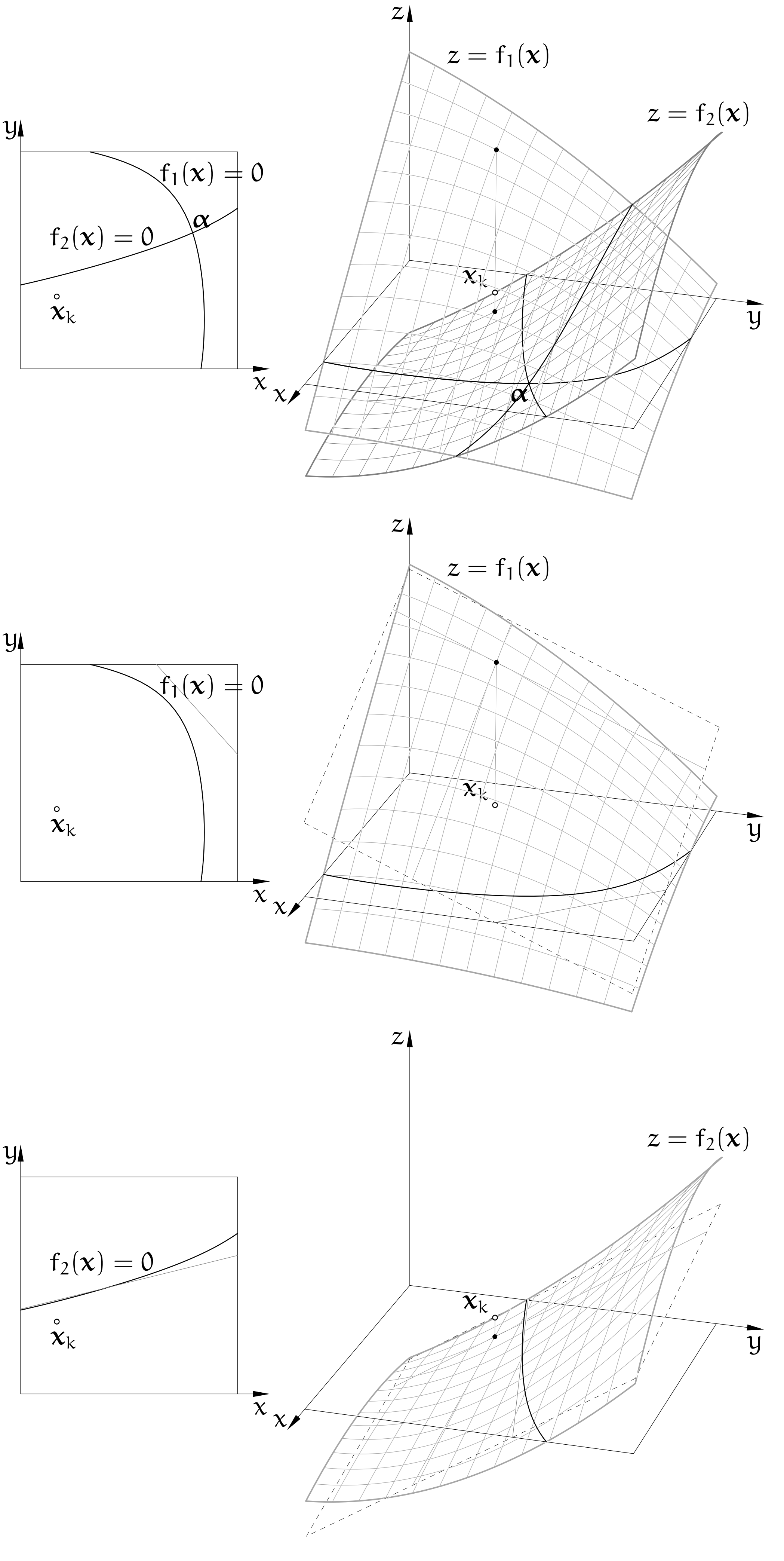
możemy oszacować tak:

$$\|\boldsymbol{\varepsilon}_{k+1}\| \leqslant \frac{M_2}{2K_1}\|\boldsymbol{\varepsilon}_k\|^2.$$

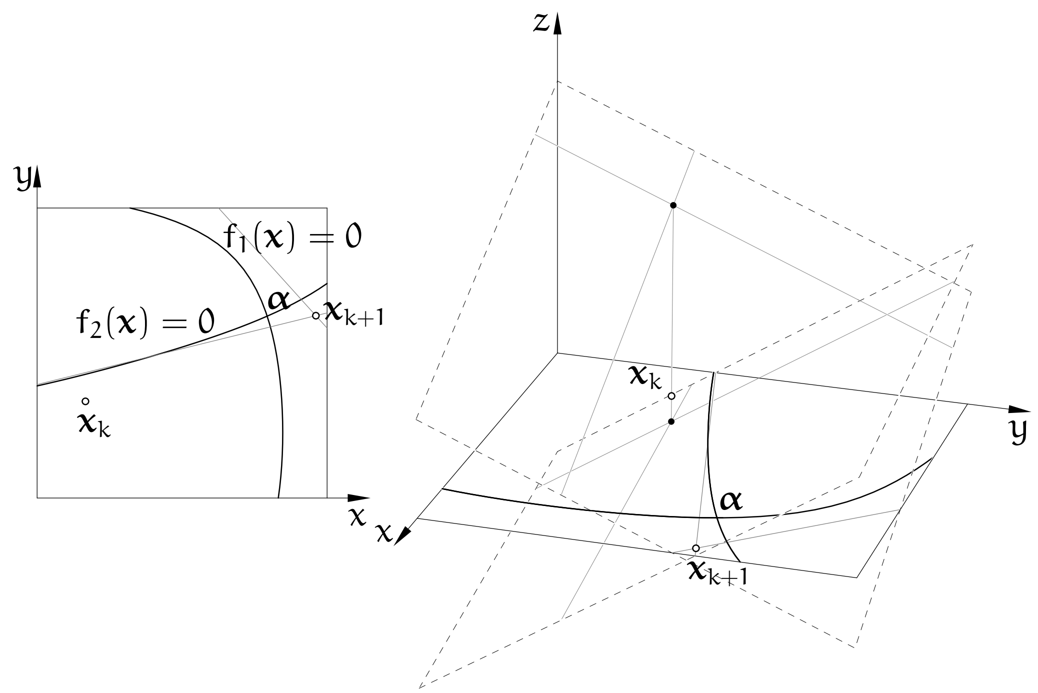
Jeśli funkcja $\mathbf{f}$ spełnia przyjęte założenia, to wykładnik zbieżności metody Newtona jest nie mniejszy niż 2 (a jeśli istnieje stała $K_2 > 0$, taka że $\|\mathbf{r}\| \geqslant \frac{K_2}{2}\|\mathbf{h}\|^2$, zobacz $(^{**}_{**})$, to jest równy 2) — końcowy rachunek (z rozwiązywaniem równania różnicowego) jest identyczny jak dla równania z jedną niewiadomą. Polecam jako ćwiczenie sformułowanie odpowiedniego twierdzenia (analogicznego do twierdzenia dla równania z jedną niewiadomą), ze szczególnym uwzględnieniem *wszystkich* niezbędnych założeń.

---

[2] We wzorach poniżej wyrażenie $\boldsymbol{\varepsilon}_k^T B_k \boldsymbol{\varepsilon}_k$ jest w nawiasie, ponieważ tak zapisane mnożenie macierzy *nie jest* łączne — nie mnożymy $J_k^{-1}\boldsymbol{\varepsilon}_k^T$, bo to nie ma sensu. Ten sam wynik moglibyśmy otrzymać, mnożąc (z lewej strony) wektorowe współczynniki (zobacz wzór $(^{**}_{**})$) macierzy B_k przez J_k^{-1}, a następnie otrzymany iloczyn z lewej i prawej strony odpowiednio przez $\boldsymbol{\varepsilon}_k^T$ i $\boldsymbol{\varepsilon}_k$.



z
z = f1(x)
z = f2(x)
y
f1(x) = 0
α
f2(x) = 0
xk
y
x
x
α
z
z = f1(x)
y
f1(x) = 0
xk
xk
y
x
x
z
z = f2(x)
y
f2(x) = 0
xk
xk
y
x
x



Modyfikacje

Metoda Newtona dla układu równań może być dość kosztowna: oprócz wartości funkcji $\mathbf{f}$, składającej się z n liczb, trzeba obliczyć macierz J_k, tj. w ogólności n^2 liczb, a następnie rozwiązać układ równań liniowych, co może wymagać wykonania $\Theta(n^3)$ działań zmiennopozycyjnych. Ze wzrostem liczby równań i niewiadomych koszty te mogą stać się zaporowe. Dla bardzo dużych n często macierz J_k jest *rzadka*, tj. ma znacznie mniej niż n^2 współczynników niezerowych. W takim przypadku należy po pierwsze obliczać tylko współczynniki niezerowe (ich rozmieszczenie w macierzy należy wyznaczyć zawczasu), a ponadto użyć metody rozwiązywania układu równań liniowych dostosowanej do macierzy rzadkiej.

Często stosuje się rozmaite modyfikacje metody Newtona. Po pierwsze, zamiast obliczać współczynniki macierzy J_k na podstawie dokładnych wzorów, które mogą być znacznie bardziej skomplikowane (czyli droższe) niż wzory opisujące funkcje f_i, można obliczać różnice dzielone; w tym celu trzeba obliczyć wartości funkcji $\mathbf{f}$ w $n+1$ punktach.

Jeśli punkty $\mathbf{x}_{k-n}, \ldots, \mathbf{x}_k$ są w położeniu ogólnym, tj. wektory $\mathbf{x}_j - \mathbf{x}_k$ dla $j = k-n, \ldots, k-1$ są liniowo niezależne, to można obliczyć przybliżenie $\tilde{J}_k$ macierzy J_k na podstawie wartości funkcji $\mathbf{f}$ w tych punktach. W ten sposób powstaje wielowymiarowa metoda siecznych. Różniczka przekształcenia afinicznego $\tilde{\mathbf{f}}\colon \mathbb{R}^n \to \mathbb{R}^n$, które w punktach $\mathbf{x}_{k-n}, \ldots, \mathbf{x}_k$ przyjmuje wartości

$\mathbf{f}_{k-n}, \ldots, \mathbf{f}_k$, jest taka sama w każdym punkcie przestrzeni i spełnia warunek

$$D\tilde{\mathbf{f}}(\mathbf{x} - \mathbf{x}_k) = \mathbf{f}(\mathbf{x}) - \mathbf{f}_k,$$

z którego wynika równość

$$\tilde{J}_k X = F,$$

gdzie $\tilde{J}_k$ oznacza macierz pochodnych przekształcenia $\tilde{\mathbf{f}}$, zaś

$$X = [\mathbf{x}_{k-n} - \mathbf{x}_k, \ldots, \mathbf{x}_{k-1} - \mathbf{x}_k], \quad F = [\mathbf{f}_{k-n} - \mathbf{f}_k, \ldots, \mathbf{f}_{k-1} - \mathbf{f}_k].$$

Jeśli więc macierze X i F są nieosobliwe, to mamy $\tilde{J} = FX^{-1}$ oraz $\tilde{J}^{-1} = XF^{-1}$. W $k+$ pierwszym kroku metody siecznych rozwiązujemy układ równań

$$F\boldsymbol{\beta} = -\mathbf{f}_k,$$

po czym obliczamy

$$\boldsymbol{\delta} = X\boldsymbol{\beta} \quad \text{i} \quad \mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\delta}.$$

Koszt tego obliczenia w ogólnym przypadku jest rzędu n^3 operacji. Wadą wielowymiarowej metody siecznych jest bardzo mały wykładnik zbieżności (bliski 1) dla dużych n.

Kolejna modyfikacja polega na wykorzystaniu macierzy J_k w kilku kolejnych iteracjach. To również obniża wykładnik zbieżności, ale dodatkowe iteracje z tą samą macierzą są bardzo tanie: nie trzeba obliczać pochodnych i można skorzystać z „gotowych" czynników (np. trójkątnych) rozkładu macierzy. Koszt rzędu n^3 w rozwiązywaniu układów równań liniowych jest związany z rozkładaniem macierzy na te czynniki, mając je, można rozwiązać układ kosztem $\Theta(n^2)$ działań.

Istnieją modyfikacje metody Newtona, mające na celu „powiększenie" kuli zbieżności poszukiwanych rozwiązań. Dla nie dość dobrego punktu $\mathbf{x}_k$ często zdarza się, że przyrost $\boldsymbol{\delta}$, otrzymany przez rozwiązanie układu równań $J_k\boldsymbol{\delta} = -\mathbf{f}_k$ jest za duży. Wtedy można przyjąć $\mathbf{x}_{k+1} = \mathbf{x}_k + \beta\boldsymbol{\delta}$, dla odpowiednio wybranego parametru $\beta \in (0, 1)$. Metoda skuteczniejsza, choć bardziej kosztowna, polega na wyznaczeniu przyrostu przez rozwiązanie układu równań

$$(J_k + \lambda I)\boldsymbol{\delta} = -\mathbf{f}_k,$$

z odpowiednio wybranym parametrem λ. Metoda ta może być też skuteczna w pewnych przypadkach, gdy macierz J_k jest osobliwa. Parametr λ dobieramy tak, aby otrzymać jak najmniejsze residuum układu, tj. aby zminimalizować normę wektora $\mathbf{f}_{k+1}$. Po pewnej liczbie iteracji możemy otrzymać przybliżenie rozwiązania należące do kuli zbieżności metody Newtona i od tej chwili przyjmować $\lambda = 0$.

Kryteria stopu

Ważnym elementem obliczeń jest podjęcie decyzji o ich przerwaniu. Na przykład wykonywanie kolejnych iteracji po osiągnięciu maksymalnej granicznej dokładności jest stratą czasu. Dlatego w pętli, realizującej iteracje, musi się pojawić jedna lub więcej instrukcji przerywających obliczenia po spełnieniu pewnego warunku.

Po pierwsze, można dać limit liczby iteracji, np. określony przez parametr procedury. W *wielu* typowych zastosowaniach, jeśli metoda Newtona nie znalazła rozwiązania (z graniczną dokładnością) po siedmiu[3] iteracjach, to już nie znajdzie (bo funkcja nie spełnia warunków koniecznych działania metody, zaczęliśmy od złego przybliżenia startowego, lub w ogóle nie ma rozwiązania).

Drugie kryterium stopu jest residualne. Residuum równania w punkcie x_k jest to liczba $f(x_k)$ (lub wektor $\mathbf{f}(\mathbf{x}_k)$). Jeśli residuum ma dostatecznie małą wartość bezwzględną (lub normę, dla układu równań), na przykład porównywalną z oszacowaniem błędu, z jakim obliczamy wartości funkcji f, to przerywamy obliczenia.

Wreszcie jest kryterium przyrostowe. Obliczenia przerywamy, gdy wartość bezwzględna (lub norma) przyrostu $\delta = x_{k+1} - x_k$ jest mniejsza niż pewna wielkość progowa. Dla wielu metod długość przyrostu w danym kroku jest górnym oszacowaniem błędu rozwiazania przybliżonego x_k (ale to zależy także od funkcji f).

[3]Proszę nie sugerować się, że zawsze taki limit jest dobry!

Zadania i problemy

1. Wykaż, że algorytm

```
w = a[n];
p = 0.0;
for ( i = n-1; i >= 0; i-- ) {
  p = p*x + w;
  w = w*x + a[i];
}
```

   dla $n > 0$ oblicza wartość wielomianu $w(x) = a_n x^n + \cdots + a_1 x + a_0$ i jego pochodnej w punkcie x.

2. Znajdź wzór opisujący funkcję iteracyjną dla obliczania pierwiastka stopnia n z liczby rzeczywistej a przy użyciu metody Newtona. Naszkicuj wykres tej funkcji i oblicz pierwsze trzy przybliżenia liczby $\sqrt{2}$, jeśli $x_0 = 1$.

3. Znajdź wykładnik zbieżności metody iteracyjnej, w której funkcja iteracyjna φ, taka że $\varphi(\alpha) = \alpha$, ma w punkcie α pochodne rzędu $1, \ldots, n$ równe 0 (i niezerową pochodną rzędu $n + 1$).

4. Znajdź wzór opisujący funkcję iteracyjną dla metody Newtona zastosowanej do funkcji $f(x) = a - 1/x$. Czy można obliczyć iloraz $x = a/b$ nie wykonując operacji dzielenia?

5. Niech $f(x) = (x - a)^n$, dla pewnego $n > 1$. Jak działa metoda Newtona w tym przypadku (nie znamy a, startujemy z $x_0 \neq a$)? Jak działa metoda zmodyfikowana (zakładamy, że znamy n), określona wzorem

$$x_{k+1} = x_k - n\frac{f(x_k)}{f'(x_k)} \ ?$$

6. Dla funkcji f jak w poprzednim zadaniu określamy funkcję $g(x) = f(x)/f'(x)$. Jak działa metoda Newtona, zastosowana do funkcji g?

   Napisz funkcję iteracyjną dla tego przypadku wzorem, w którym występują tylko wartości i pochodne funkcji f.

7. Udowodnij zbieżność metody Newtona, użytej do rozwiązania równania $f(x) = 0$, w którym funkcja f jest klasy C^1 w przedziale A zawierającym rozwiązanie, przy czym istnieją stałe dodatnie K_1 i L, takie że $|f'(x)| > K_1$, a ponadto $|f'(x) - f'(y)| \leqslant L|x - y|$ dla każdego $x, y \in A$. Znajdź wykładnik zbieżności.

8. Znajdź wykładnik zbieżności metody Newtona, jeśli funkcja f jest klasy $C^3(A)$, pochodna funkcji f ma wartość bezwzględną nie mniejszą niż pewne $K_1 > 0$ w przedziale A, a ponadto pochodna drugiego rzędu w punkcie α będącym rozwiązaniem, ma wartość 0.

9. Przypuśćmy, że pochodna funkcji f klasy C^3 jest niezerowa. Określamy funkcję g wzorem $g(x) = f(x)/\sqrt{|f'(x)|}$. Oblicz pochodną drugiego rzędu funkcji g w punkcie α, takim że $f(\alpha) = 0$. Co można stąd wywnioskować o zbieżności metody Newtona zastosowanej do funkcji g?

   Zastąpienie funkcji f przez określoną wyżej funkcję g daje metodę zwaną <u>metodą Halleya</u> (tak, tego, który odkrył kometę). Napisz wzór opisujący funkcję iteracyjną dla metody Halleya.

10. Zbadaj, jak zachowuje się metoda Newtona, jeśli funkcja f, której miejsce zerowe należy znaleźć, jest określona wzorem

$$f(x) = \operatorname{sgn} x\sqrt{|x|}$$

   dla dowolnego punktu startowego $x_0 \neq 0$.

11. Przypuśćmy, że w metodzie siecznych początkowe przybliżenia rozwiązania są końcami przedziału, w którym funkcja f zmienia znak. Metodę modyfikujemy w ten sposób, że po wyznaczeniu kolejnego przybliżenia odrzucamy to, w którym funkcja f ma ten sam znak, co w nowym punkcie (ta metoda nazywa się <u>regula falsi</u>). Jaki jest wykładnik zbieżności tej metody, jeśli funkcja f jest ściśle wypukła albo ściśle wklęsła?

12. Zbadaj rząd zbieżności <u>metody Steffensena</u>, w której funkcja iteracyjna jest dana wzorem

$$\varphi(x) = x - \frac{f(x)}{f[x, x + f(x)]},$$

   przy takich samych założeniach, jak dla metod Newtona i siecznych w wykładzie.

   <u>Wskazówka:</u> Użyj wzoru na resztę interpolacyjną (jak dla metody siecznych) i weź pod uwagę fakt, że dla funkcji f klasy $C^2(A)$ istnieje górne oszacowanie wartości bezwzględnej pochodnej pierwszego rzędu.

13. (zadanie z książki Kincaida i Cheneya) Wykonaj jeden lub dwa kroki metody Newtona dla układów równań

$$\left\{\begin{array}{rcl} xy - z^2 &=& 1 \\ xyz - x^2 + y^2 &=& 2 \\ e^x - e^y + z &=& 3 \end{array}\right., \qquad \mathbf{x}_0 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix},$$

$$\left\{\begin{array}{rcl} 4x^2 - y^2 &=& 0 \\ 4xy - x &=& 1 \end{array}\right., \qquad \mathbf{x}_0 = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

$$\left\{\begin{array}{rcl} xy^2 + x^2y + x^4 &=& 3 \\ x^3y^5 - 2x^5y - x^2 &=& -2 \end{array}\right., \qquad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

14. Rozważamy dwie metody rozwiązywania równań nieliniowych o wykładnikach zbieżności $p, q > 1$. Załóżmy, że ciągi błędów kolejnych przybliżeń rozwiązania spełniają odpowiednio warunki $|\tilde{\varepsilon}_{k+1}| = C|\tilde{\varepsilon}_k|^p$ i $|\hat{\varepsilon}_{k+1}| = C|\hat{\varepsilon}_k|^q$ z tą samą stałą C, a także że $\tilde{\varepsilon}_0 = \hat{\varepsilon}_0$. Koszt (czasowy) jednej iteracji pierwszej i drugiej metody jest równy odpowiednio T_1 i T_2. Wykaż, że jeśli rozwiązanie ma być znalezione z *bardzo dużą* dokładnością, tzn. z ustalonym, bardzo małym progiem tolerancji błędu (przyjmujemy, że nie ma błędów zaokrągleń), to pierwsza metoda zużyje na to mniej czasu, jeśli $T_1/T_2 < \log p/\log q$. Wyciągnij stąd podany na wykładzie wniosek dotyczący porównania efektywności metod Newtona i siecznych.
15. Równanie

$$\sin x = 2x - 1$$

rozwiązujemy metodą Newtona. Znajdź dolne oszacowanie promienia kuli zbieżności jedynego rozwiązania.

Przypomnienia

1. Równanie różnicowe liniowe rzędu n o stałych współczynnikach ma postać

$$a_k = c_{n-1}a_{k-1} + \cdots + c_0 a_{k-n} + f(k), \qquad (*)$$

gdzie liczby rzeczywiste $c_0, \ldots, c_{n-1}$ i funkcja f są ustalone (przy czym $c_0 \neq 0$ i $c_{n-1} \neq 0$). Jeśli określimy warunek początkowy, tj. podamy liczby $a_0, \ldots, a_{n-1}$ (albo ogólniej $a_j, \ldots, a_{j+n-1}$ dla jakiegoś j), to powyższe równanie różnicowe określa jednoznacznie ciąg nieskończony $a_n, a_{n+1}, \ldots$ (albo $a_{j+n}, a_{j+n+1}, \ldots$).

Równania różnicowe były użyte w wykładzie do znalezienia rzędów zbieżności metod Newtona i siecznych. Można ich używać do analizy zbieżności innych metod numerycznych, ale są one także wygodnym środkiem do badania kosztów i złożoności wielu algorytmów (nie tylko numerycznych).

Poniżej przypominam ogólny sposób rozwiązywania takich równań z funkcjami f o pewnej postaci — klasa równań, które można rozwiązać tym sposobem, ma wiele zastosowań.

2. Na początek rozważmy równanie jednorodne, tj. z zerową funkcją f:

$$a_k = c_{n-1}a_{k-1} + \cdots + c_0 a_{k-n}. \qquad (**)$$

Jeśli równanie to jest spełnione przez pewne ciągi nieskończone, to jego rozwiązaniem jest też każda kombinacja liniowa tych ciągów. Natomiast znając dowolne rozwiązanie równania (*), możemy otrzymać każde inne jego rozwiązanie, dodając pewne rozwiązanie równania jednorodnego.

3. Każde jednorodne równanie różnicowe o stałych współczynnikach jest spełnione przez pewien ciąg geometryczny, $a_k = \lambda^k$, należy więc znaleźć λ. Podstawiając do równania wyrażenie opisujące taki ciąg, dostajemy

$$\lambda^k = c_{n-1}\lambda^{k-1} + \cdots + c_0\lambda^{k-n}.$$

Po podzieleniu stron przez λ^{k-n} i uporządkowaniu, dostajemy tzw. równanie charakterystyczne

$$\lambda^n - c_{n-1}\lambda^{n-1} - \cdots - c_0 = 0.$$

Każdy wielomian stopnia $n > 0$ ma (rzeczywiste lub zespolone) miejsce zerowe. Ciąg złożony z kolejnych potęg tego miejsca zerowego spełnia równanie (**), co łatwo sprawdzić. Tak więc każdemu pierwiastkowi wielomianu charakterystycznego $w(\lambda) = \lambda^n - c_{n-1}\lambda^{n-1} - \cdots - c_0$ odpowiada ciąg geometryczny, który jest rozwiązaniem.

4. Jeśli wszystkie miejsca zerowe wielomianu charakterystycznego mają krotność 1, tj. jest n różnych liczb $\lambda_1, \ldots, \lambda_n$, takich że $w(\lambda_j) = 0$, to mamy n liniowo niezależnych ciągów geometrycznych, które spełniają równanie jednorodne. Jeśli pewne miejsce zerowe, λ_j, jest liczbą zespoloną, to liczba $\overline{\lambda}_j$ też jest miejscem zerowym. Z dwóch zespolonych ciągów geometrycznych, $(\lambda_j^k)_k$ i $(\overline{\lambda}_j^k)_k$, możemy otrzymać dwa niezależne liniowo ciągi rzeczywiste, które spełniają równanie (**): $(\lambda_j^k + \overline{\lambda}_j^k)_k$ oraz $i(\lambda_j^k - \overline{\lambda}_j^k)_k$.
5. Jeśli pewne miejsce zerowe, λ_j, ma krotność $r > 1$, to rozwiązaniami równania (**) są ciągi $(\lambda_j^k)_k, (k\lambda_j^k)_k, \ldots, (k^{r-1}\lambda_j^k)_k$, zatem dla równania jednorodnego możemy zawsze znaleźć n (niezależnych liniowo) ciągów nieskończonych będących rozwiązaniami. Z rozwiązaniami określonymi przez zespolone miejsca zerowe o krotności większej niż 1 radzimy sobie podobnie, jak z zespolonymi ciągami geometrycznymi.
6. Jeśli funkcja f w równaniu (*) ma postać

$$f(k) = p(k)\mu^k,$$

gdzie $p(k)$ jest wielomianem stopnia s, zaś liczba μ jest pierwiastkiem równania charakterystycznego o krotności m (w szczególności jeśli $w(\mu) \neq 0$, to $m = 0$) ma rozwiązanie o postaci $k^m q(k)\mu^k$, gdzie $q(k)$ jest pewnym wielomianem stopnia s.

Dla funkcji o innej postaci trzeba kombinować, czasami można coś zgadnąć.
7. Metoda rozwiązywania (na papierze) równania (*) z funkcją f jak wyżej:

   - Układamy i rozwiązujemy równanie charakterystyczne (tzn. wyznaczamy wszystkie pierwiastki oraz ich krotności),
   - Znajdujemy dowolne rozwiązanie równania (*). W tym celu określamy m i w miejsce $a_k, \ldots, a_{k-n}$ podstawiamy do równania wyrażenia $k^m(q_s k^s + \cdots + q_1 k + q_0)\mu^k, \ldots, (k-n)^m\bigl(q_s(k-n)^s + \cdots + q_1(k-n) + q_0\bigr)\mu^{k-n}$. Osobno grupujemy składniki różniące się wykładnikiem przy k. Suma składników w każdej grupie jest równa 0. Otrzymujemy stąd układ równań liniowych, którego rozwiązaniem są współczynniki $q_0, \ldots, q_s$ wielomianu q. Obliczamy je, rozwiązując układ.
   - Postać ogólną rozwiązania (tj. np. $k^m q(k)\mu^k + b_1\lambda_1^k + \cdots + b_n\lambda_n^k$, jeśli wszystkie pierwiastki są jednokrotne; w tym przypadku może być $m = 0$ lub $m = 1$), gdzie $q(k)$ jest znalezionym w poprzednim kroku wielomianem, podstawiamy w miejsce a_k do równania (*). Następnie, biorąc $k = 0, \ldots, n-1$ i korzystając z warunku początkowego, piszemy i rozwiązujemy układ równań liniowych z niewiadomymi $b_1, \ldots, b_n$. W ten sposób otrzymujemy współczynniki opisujące ciąg, który spełnia równanie (*) i warunek początkowy.

Uff.

Arytmetyka zmiennopozycyjna

Liczb rzeczywistych jest nieskończenie (a nawet nieprzeliczalnie) wiele, a pamięć choćby największego komputera jest skończona. Dlatego w obliczeniach numerycznych musimy się zadowolić poruszaniem się w pewnym skończonym zbiorze, którego elementy tylko przybliżają wszelkie liczby rzeczywiste, jakie mogłyby się pojawić w tych obliczeniach.

Rzędy wielkości tych liczb mogą być różne i błędy ich przybliżenia też mogą być różne. Zwykle im większa liczba, tym większy jej błąd nam nie przeszkadza. Na przykład, jeśli dowiemy się, że jakiś obiekt ma długość 147 km, to informację, że w rzeczywistości ma o 1 mm mniej, jesteśmy skłonni zignorować. Co innego, jeśli obiekt ma tylko 0.5 mm długości — błąd rzędu 1 mm jesteśmy wtedy skłonni potraktować z całą powagą i stanowczością.

Naturalne jest zatem używanie takiego sposobu reprezentowania liczb, który umożliwia przybliżanie tych liczb rzeczywistych, które nie mają dokładnej reprezentacji, z małym błędem względnym. Na przykład, jeśli długość 147 km jest podana z błędem nie większym niż 0.1%, to wiemy, że błąd bezwzględny jest mniejszy niż 150 m, i taka dokładność nieraz nam wystarczy.

Reprezentacja zmiennopozycyjna

Powszechnie używana reprezentacja zmiennopozycyjna liczb rzeczywistych jest kompromisem między dokładnością i złożonością czasową i pamięciową. Jej głównym celem jest masowe przetwarzanie liczb, czemu służy stosunkowo mała ilość miejsca zajmowanego przez tę reprezentację i możliwość szybkiego wykonywania działań przez specjalnie opracowane w tym celu podukłady procesorów. Błędy tej reprezentacji są dostatecznie małe na potrzeby znakomitej większości zastosowań. Istnieją inne reprezentacje, umożliwiające prowadzenie obliczeń ze znacznie większą dokładnością, ale znacznie wolniej i w większej pamięci. Te inne reprezentacje są poza zakresem tego wykładu. Jeszcze jedno: reprezentacje zmiennopozycyjne mają powszechnie przyjęty standard, który ułatwia m.in. wymianę danych. Reprezentacje niestandardowe tak fajnie nie mają.

Idea reprezentacji zmiennopozycyjnej wiąże się z tzw. półlogarytmicznym zapisem liczby. Każdą dodatnią liczbę rzeczywistą możemy przedstawić za pomocą liczby z przedziału $[1, 10)$ i całkowitej potęgi liczby 10, na przykład

$$27182818 = 2.7182818 \cdot 10^7.$$

W komputerach zamiast podstawy 10 i dziesięciu różnych cyfr, wygodniej jest używać podstawy 2 i bitów.

Podstawowa reprezentacja określona przez standard IEEE-754 (opracowany w 1985 r.) składa się z bitu znaku, s, po którym następuje cecha c i mantysa m:

$$\boxed{s}\underbrace{\boxed{\qquad c \qquad}}_{d}\underbrace{\boxed{\qquad\qquad\qquad m \qquad\qquad\qquad}}_{t}$$

Mantysa jest liczbą rzeczywistą; jeśli reprezentuje ją ciąg bitów $b_{t-1}b_{t-2}\dots b_1 b_0$, to $m = \sum_{k=0}^{t-1} b_k 2^{k-t}$, a zatem zawsze $0 \leqslant m < 1$. Cecha jest liczbą całkowitą (bez znaku), reprezentowaną za pomocą d bitów, która wpływa na sposób interpretacji całego ciągu bitów. Liczba reprezentowana przez taki ciąg, w zależności od cechy, jest równa

$$\begin{aligned}
x &= (-1)^s\, 2^{c-b}\, (1+m) && \text{dla } 0 < c < 2^d - 1,\\
x &= (-1)^s\, 2^{1-b}\, m && \text{dla } c = 0,\\
x &= (-1)^s\, \infty && \text{dla } c = 2^d - 1,\ m = 0,\\
x &= \text{NaN („nie-liczba”)} && \text{dla } c = 2^d - 1,\ m \neq 0.
\end{aligned}$$

Liczby d, t i b są ustalone dla konkretnej reprezentacji. Cechą charakterystyczną reprezentacji z użyciem pierwszego wzoru jest tzw. normalizacja. Mając dowolną liczbę rzeczywistą $x \neq 0$, przedstawioną w układzie dwójkowym, dobieramy cechę c (czyli równoważnie czynnik 2^{c-b}) tak, że czynnik $(1+m)$ w wyrażeniu opisującym x jest liczbą z przedziału $[1, 2)$. Jeśli otrzymana w ten sposób cecha jest za duża (większa lub równa $2^d - 1$), to mamy nadmiar zmiennopozycyjny (ang. *floating point overflow*), czyli niewykonalne zadanie reprezentowania liczby o za dużej wartości bezwzględnej, zwykle będące powodem do przerwania obliczeń. Jeśli nie ma nadmiaru, to pierwszy wzór opisuje liczbę w ten sposób, że najbardziej znacząca jedynka w rozwinięciu dwójkowym nie jest jawnie pamiętana — właśnie to jest normalizacja. Dzięki niej każdy ciąg bitów reprezentuje inną liczbę, co m.in. umożliwia optymalne wykorzystanie bitów do zmniejszenia błędów.

Niech x oznacza dowolną liczbę rzeczywistą. Jej reprezentację, tj. położoną najbliżej niej liczbę zmiennopozycyjną, oznaczymy symbolem $\mathrm{rd}(x)$ (z ang. *rounding*). Jeśli liczbę x możemy przedstawić w postaci

$$x = (-1)^s 2^{c-b}(1+f),$$

dobierając cechę c tak, aby mieć $f \in [0, 1)$ oraz $0 < c < 2^d - 1$, to (z jednym rzadkim wyjątkiem, gdy f trzeba zaokrąglić w górę do jedynki) będziemy mieli

$$\mathrm{rd}(x) = (-1)^s 2^{c-b}(1+m),$$

przy czym $|f - m| \leqslant 2^{-t-1}$. Błąd względny reprezentacji spełnia nierówność

$$\frac{|x - rd(x)|}{|x|} = \frac{|(-1)^s 2^{c-b}(1+f) - (-1)^s 2^{c-b}(1+m)|}{|(-1)^s 2^{c-b}(1+f)|} \leqslant |f - m| \leqslant 2^{-t-1}.$$

Co ciekawe, nierówność ta jest spełniona też w specjalnym przypadku wspomnianym wcześniej (bo w mianowniku $1 + f \approx 2$). Zatem, maksymalny błąd względny reprezentacji zmiennopozycyjnej, jeśli nie ma niedomiaru ani nadmiaru, jest na poziomie 2^{-t-1}, gdzie t jest liczbą bitów mantysy. Jeśli kierunek zaokrąglania wybieramy mniej starannie (np. zawsze obcinamy w kierunku zera), to błąd względny może być dwa razy większy, czyli rzędu $\nu = 2^{-t}$.

Bardziej skomplikowana sytuacja zdarza się w przypadku, gdy cecha jest za mała (tj. gdy w pierwszym wzorze należałoby przyjąć $c \leqslant 0$). Wtedy korzystamy z drugiego wzoru, w którym występuje czynnik m (przypominam, że $m \in [0, 1)$). Jeśli $c = m = 0$, to mamy reprezentację zera; liczba 0 jako jedyna ma dwie reprezentacje, różniące się bitem znaku. Jeśli $c = 0$ i $m \neq 0$, to mamy do czynienia z <u>niedomiarem zmiennopozycyjnym</u>, czyli reprezentowaniem liczby x za pomocą mantysy o mniejszej liczbie bitów istotnych (jeśli w użyciu jest pierwszy wzór, to istotne są wszystkie bity mantysy, jeśli drugi, to tylko bity od pozycji najmniej znaczącej, do najbardziej znaczącej pozycji, na której jest jedynka). Najdokładniejszą reprezentacją liczb o bardzo małej wartości bezwzględnej (mniejszej niż 2^{-b-t}) jest 0. Niedomiar wiąże się zatem ze (stopniową) utratą dokładności reprezentacji. Dla $x \to 0$ błąd względny reprezentacji dąży do 100%, a błąd bezwzględny jest ograniczony. W analizie błędów najczęściej nie bierzemy tego przypadku pod uwagę.

Reprezentacja umożliwia używanie nieskończoności, także w rachunkach (np. wynik dzielenia dowolnej liczby przez nieskończoność jest równy 0).

Nie-liczby są wykorzystywane do sygnalizowania błędów, np. próby obliczenia pierwiastka kwadratowego z liczby ujemnej. Można je też wykorzystać do odpluskwiania programu, np. nadając zmiennym takie wartości początkowe, a następnie śledząc, czy nie ma do nich odwołań przed przypisaniem właściwej wartości liczbowej.

W standardzie IEEE-754 są zdefiniowane formaty liczb pojedynczej i podwójnej precyzji, a także liczb pojedynczej i podwójnej rozszerzonej precyzji. Liczby pojedynczej rozszerzonej precyzji jakoś się nie przyjęły, procesory w komputerach PC ich nie obsługują. Dane na temat standardowych formatów są w tabelce:

	B	d	t	b	M	S	ν	μ
pojedyncza, `float`	32	8	23	127	10^{38}	10^{-38}	10^{-7}	10^{-45}
pojed. rozszerzona —	44	11	31	1023	10^{308}	10^{-308}	10^{-10}	10^{-317}
podwójna `double`	64	11	52	1023	10^{308}	10^{-308}	10^{-15}	10^{-323}
podw. rozszerzona `long double`	80 (96, 128)	15	63	16383	10^{4932}	10^{-4932}	10^{-19}	10^{-4951}

Oznaczenia: B — całkowita liczba bitów, d — liczba bitów cechy, t — liczba bitów mantysy, b — stała odejmowana od cechy w celu otrzymania wykładnika. Stała b jest równa $2^{d-1}-1$, dzięki czemu jeśli liczba x ma reprezentację znormalizowaną, to $1/x$ na ogół też ma. Liczby $M = 2^{2^d-b-2}(2-2^{-t})$ — największa liczba zmiennopozycyjna, $S = 2^{1-b}$ — najmniejsza dodatnia liczba reprezentowana w postaci znormalizowanej (tj. bez niedomiaru), $\nu = 2^{-t}$ — oszacowanie maksymalnego błędu względnego reprezentacji znormalizowanej, oraz $\mu = 2^{1-b-t}$ — najmniejsza zmiennopozycyjna liczba dodatnia, są podane w przybliżeniu (tylko rząd wielkości).

Reprezentacje rozszerzonej precyzji nie wymuszają normalizacji (mantysa ma $t+1$ bitów i jest liczbą z przedziału $[0,2)$, jej najbardziej znaczący bit ma wartość 1), ale wyniki działań, jeśli nie ma niedomiaru, są normalizowane przez procesor. Jeszcze jedno: w 32-bitowych systemach operacyjnych zmienna rozszerzonej podwójnej precyzji zajmuje 96 bitów (12 bajtów), z których 16 (2 bajty) jest nieużywanych. W systemach 64-bitowych taka zmienna zajmuje 128 bitów (16 bajtów), z których 48 (tj. 6 bajtów) jest nieużywanych[4]. To utrudnia m.in. przenoszenie danych między komputerami w postaci binarnej. Jeśli nie ma istotnego powodu, to najlepiej nie używać tej reprezentacji liczb.

Oprócz standardu IEEE-754 istnieje też standard IEEE-854, który definiuje reprezentacje liczb zmiennopozycyjnych z podstawami 2 i 10. Standard ten służy do wymiany danych między komputerami, natomiast określone przezeń reprezentacje nie są przetwarzane bezpośrednio przez jednostki zmiennopozycyjne procesorów (w każdym razie znanych mi). Jeśli nie ma ważnych powodów do używania reprezentacji określonych w tym standardzie, to można się nim nie przejmować.

[4]Powodem jest oczywiście czas dostępu do pamięci — aby go skrócić, kompilator przydziela zmiennym adresy podzielne przez rozmiar słowa maszynowego.

Do celów specjalnych bywają używane reprezentacje niestandardowe; istnieje np. dość rzadko spotykany format poczwórnej precyzji, w którym reprezentacja liczby zajmuje 128 bitów (cecha ma w nim 15 bitów, mantysa 112). Nie słyszałem o procesorach z rejestrami zmiennopozycyjnymi o takiej długości, zatem działania na takich liczbach muszą być wykonywane przez odpowiednie podprogramy, co zabiera sporo czasu. Z drugiej strony, reprezentacje 16- 11- i 10-bitowe (bit znaku może być nieobecny, cecha ma 5 bitów, a mantysa 10, 6 albo 5) są używane przez niektóre karty graficzne podczas wykonywania obrazów, gdy dokładność ma małe znaczenie, zaś najważniejsza jest szybkość obliczeń i oszczędność miejsca. Wspomniane karty graficzne mają specjalizowane podukłady do wykonywania działań na takich liczbach.

Arytmetyka i błędy zaokrągleń

Na potrzeby analizy błędów działanie procesora podczas wykonywania operacji arytmetycznych można sobie wyobrazić tak: dokładny wynik działania jest poddawany normalizacji (tj. dobierana jest cecha), a następnie zaokrągleniu — nieskończony ciąg bitów mantysy jest obcinany i ewentualnie zaokrąglany w górę. Nie wyznacza się oczywiście nieskończonego ciągu bitów mantysy, zamiast tego wykorzystuje się trzy bity dodatkowe („wystające" poza format), z których pierwsze dwa są zwykłe, a trzeci „lepki" — bit ten otrzymuje wartość 1, jeśli dowolny dalszy bit nieskończenie długiej mantysy jest niezerowy. Te trzy bity zawsze wystarczą do poprawnego zaokrąglenia liczby. Wyboru kierunku zaokrąglania można dokonać, ustawiając odpowiednie bity w rejestrze sterującym procesora (zwykle zostawiamy domyślne zaokrąglanie do najbliższej liczby zmiennopozycyjnej).

Istotne jest, że oprócz reprezentacji liczb, standard IEEE-754 określa własności działań, w tym wymagania dotyczące dokładności wyników — dotyczy to czterech działań arytmetycznych, pierwiastka kwadratowego, oraz konwersji reprezentacji całkowitej i zmiennopozycyjnej. Istnieją procesory, które wprawdzie przetwarzają liczby w standardowym formacie, ale realizowane przez nie działania *nie spełniają* wszystkich warunków określonych w standardzie. Najbardziej rozpowszechnionym sprzętem tego rodzaju są karty graficzne, które mogą m.in. nie obsługiwać liczb nieznormalizowanych (tj. zapisanych przy użyciu drugiego wzoru podanego w opisie formatu; w razie niedomiaru wynikiem działania jest zero) lub zaokrąglać wyniki działań w arbitralnie określony sposób (standard nakazuje umożliwiać dokonanie wyboru). Powinien o tym pamiętać każdy, kto zajmuje się tzw. GPGPU (*general purpose GPU* (*graphics processing unit*) *programming*).

Jeśli x jest liczbą rzeczywistą, a $\mathrm{rd}(x)$ jest jej znormalizowanym zmiennopozycyjnym przybliżeniem (bez nadmiaru i niedomiaru), to mamy $|x - \mathrm{rd}(x)| \leqslant |x|2^{-1-t}$, skąd wynika, że istnieje liczba ε, taka że

$$\mathrm{rd}(x) = x(1+\varepsilon) \quad \text{oraz} \quad |\varepsilon| \leqslant 2^{-1-t}.$$

Sposób zaokrąglania (do najbliższej liczby zmiennopozycyjnej, zawsze w stronę zera, zawsze w przeciwną stronę, zawsze w górę albo zawsze w dół) może być ustawiony różnie, przez co błąd względny może być dwa razy większy. Jeśli zatem $\diamond$ oznacza dowolne z czterech działań arytmetycznych, to zamiast wyniku $x = a \diamond b$, po zaokrągleniu, otrzymamy liczbę

$$\tilde{x} = \mathrm{fl}(a \diamond b) = (a \diamond b)(1+\varepsilon),$$

dla pewnego $\varepsilon \in (-\nu, \nu)$ (piszemy $\mathrm{fl}(a \diamond b)$ zamiast $\mathrm{rd}(a \diamond b)$, bo ten ostatni symbol oznacza u nas wynik zaokrąglenia do najbliższej liczby zmiennopozycyjnej).

W superdokładnych analizach błędów używana jest funkcja ulp (ang. *unit in the last position*), która liczbie zmiennopozycyjnej x przyporządkowuje jej odległość od najbliższej innej liczby zmiennopozycyjnej. Mamy

$$\mathrm{ulp}\, x = \begin{cases} 2^{c-b-t} & \text{dla } c > 0, \text{ tj. } x = (-1)^s 2^{c-b}(1+m), \\ 2^{1-b-t} & \text{dla } c = 0, \text{ tj. } x = (-1)^s 2^{1-b} m. \end{cases}$$

Funkcja ta jest wartością bezwzględną przyrostu liczby x spowodowaną zmianą (zanegowaniem) najmniej znaczącego bitu mantysy.

Wyniki działań są najczęściej argumentami dalszych działań, zatem podczas obliczeń numerycznych ma miejsce zjawisko zwane <u>kumulacją błędów</u>.
W szczególnych przypadkach może ono doprowadzić do otrzymania bardzo niedokładnych wyników końcowych, mimo że poszczególne błędy zaokrągleń są małe. Ponadto wskutek zaokrągleń zbiór liczb zmiennopozycyjnych z działaniami dodawania i mnożenia *nie jest* ciałem (z punktu widzenia algebry). Przede wszystkim, nie jest zamknięty ze względu na działania (bo może wystąpić nadmiar) i są w nim dzielniki zera (np. jeśli liczba $|x| \neq 0$ jest dostatecznie mała, to $\mathrm{fl}(x * x) = 0$). Po drugie, dodawanie i mnożenie nie są działaniami łącznymi i dodawanie nie jest rozdzielne względem mnożenia. W konsekwencji, algorytmy oparte na różnych wzorach algebraicznie równoważnych (w ciele $\mathbb{R}$), mogą produkować różne wyniki (czasem bardzo od siebie odległe). Analiza algorytmów ma na celu między innymi badanie, na jaką dokładność wyników obliczeń wykonywanych z błędami zaokrągleń można liczyć (i może się przydać do wybrania najlepszego algorytmu, albo przynajmniej do odrzucenia najgorszego).

Arytmetyka zmiennopozycyjna zespolona

W różnych zadaniach występują liczby zespolone. W obliczeniach ich części rzeczywiste i urojone są reprezentowane w postaci zmiennopozycyjnej. Jeśli zatem zamiast liczby $z = (a, b) \neq 0$ mamy liczbę $\tilde{z} = (\tilde{a}, \tilde{b}) = (a(1+\varepsilon_a), b(1+\varepsilon_b))$, gdzie $|\varepsilon_a|, |\varepsilon_b| < \nu$, to liczbę z reprezentujemy z błędem względnym

$$\frac{|z-\tilde{z}|}{|z|} = \frac{\sqrt{a^2\varepsilon_a^2 + b^2\varepsilon_b^2}}{\sqrt{a^2+b^2}} < \frac{\sqrt{a^2\nu^2 + b^2\nu^2}}{\sqrt{a^2+b^2}} = \nu.$$

Zatem reprezentacja zmiennopozycyjna liczby zespolonej zapewnia równie mały błąd, jak reprezentacja liczby rzeczywistej. Dodawanie i odejmowanie liczb zespolonych wykonujemy na podstawie wzorów będących definicją tych działań, w związku z czym, jeśli nie ma nadmiaru ani niedomiaru, otrzymamy

$$\mathrm{fl}(z_1 \pm z_2) = (z_1 \pm z_2)(1+\varepsilon), \quad \text{gdzie} \quad |\varepsilon| < \nu.$$

Mnożenie też wykonuje się na podstawie definicji tego działania:

$$(a_1, b_1) \cdot (a_2, b_2) = (a_1a_2 - b_1b_2, a_1b_2 + a_2b_1).$$

Zamiast dokładnego wyniku otrzymamy

$$\begin{aligned}\mathrm{fl}((a_1, b_1) \cdot (a_2, b_2)) = ((&a_1a_2(1+\varepsilon_1) - b_1b_2(1+\varepsilon_2))(1+\varepsilon_3),\\ &(a_1b_2(1+\varepsilon_4) + a_2b_1(1+\varepsilon_5))(1+\varepsilon_6)),\end{aligned}$$

przy czym, jeśli w żadnym działaniu nie wystąpił nadmiar ani niedomiar, to wszystkie epsilony mają wartości bezwzględne mniejsze niż ν. Można udowodnić (za pomocą dosyć żmudnego rachunku), że wtedy otrzymany wynik jest równy

$$(a_1, b_1) \cdot (a_2, b_2) \cdot (1+\xi),$$

gdzie ξ jest pewną liczbą zespoloną, taką że $|\xi| < (1+\sqrt{2})\nu$.

Dzielenie zespolone jest bardziej kłopotliwe, bo algorytm musi unikać nadmiaru i niedomiaru (zwróćmy uwagę, że nawet w przypadku mnożenia, wynik działania może mieć reprezentację, zaś wyniki pośrednie mogą jej nie mieć z powodu nadmiaru — w dzieleniu ten problem też występuje). Dzielenie powinno się wykonywać za pomocą algorytmu

```
if ( fabs ( a2 ) >= fabs ( b2 ) ) {
  p = b2/a2;
  q = a2+b2*p;
  wynik = ((a1+b1*p)/q, (b1-a1*p)/q);
```

```
}
else {
  p = a2/b2;
  q = a2*p+b2;
  wynik = ((a1*p+b1)/q, (b1*p-a1)/q);
}
```

Jeśli nie ma nadmiaru ani niedomiaru, to względny błąd zaokrąglenia wyniku nie jest większy niż $(4+\sqrt{2})\nu$.

Zadania i problemy

1. Oblicz oznaczone na wykładzie literami M, S i ν parametry charakteryzujące niestandardowe arytmetyki zmiennopozycyjne: 128-, 16-, 11- i 10-bitową. Przyjmij $b = 2^{d-1} - 1$.
2. Niech x oznacza liczbę zmiennopozycyjną pojedynczej lub podwójnej precyzji, a $i(x)$ liczbę całkowitą bez znaku (odpowiednio 32- lub 64-bitową) reprezentowaną przez ten sam ciąg bitów. Sprawdź, że jeśli $x_2 > x_1 \geqslant 0$, $x_1 > 0 > x_2$ lub $0 \geqslant x_1 > x_2$, to $i(x_2) > i(x_1)$. Wyjaśnij, jak można to wykorzystać do sortowania ciągu liczb zmiennopozycyjnych. Co przeszkadza w sortowaniu w analogiczny sposób liczb rozszerzonej podwójnej precyzji?
3. Znajdź oszacowanie błędu zaokrąglenia iloczynu liczb zespolonych przy założeniu, że nie ma nadmiaru ani niedomiaru i wyniki działań rzeczywistych są zaokrąglane do t bitów mantysy (używaj przy tym oznaczenia ν liczby 2^{-t}).

Błędy w obliczeniach

W obliczeniach numerycznych występują błędy pięciu rodzajów.

Błędy modelu. Model matematyczny dowolnego zjawiska (przyrodniczego, ekonomicznego i w ogóle każdego) jest tego zjawiska uproszczeniem. Na przebieg zjawiska ma wpływ wiele różnych czynników, z których jedne są ignorowane (bo ich wpływ został uznany za pomijalny), a inne nie są znane dostatecznie dokładnie, aby można było napisać całkowicie poprawny wzór. Jeśli model znacznie odbiega od zjawiska, to i wyniki obliczeń mogą bardzo się różnić od tego, co można zaobserwować w rzeczywistości.

Błędy danych wejściowych. Dane wejściowe trzeba zapisać w postaci liczb zmiennopozycyjnych, co powoduje ich zaburzenie. Jeśli wynik od danych zależy (a zwykle tak jest), to nawet gdyby nie było innych błędów, wynik obliczeń może się różnić od wyniku doświadczenia. Ponadto, na ogół dane otrzymujemy z pomiarów, których niedokładności mogą być znacznie większe niż błąd reprezentacji zmiennopozycyjnej. Najdokładniejsze pomiary w fizyce dają kilkanaście cyfr dokładnych, często znamy dane z dokładnością rzędu 1%, a czasami błędy są na poziomie kilkudziesięciu procent. Sygnały lub obrazy mogą być zniekształcone z powodu szumu i bardzo niewyraźne. To wszystko ma bardzo duży wpływ na wynik (albo jego brak, jeśli algorytm nie poradzi sobie z niedokładnymi danymi).

Błędy aproksymacji. W obliczeniach numerycznych stosuje się przybliżenia funkcji, których dokładne obliczenie jest niewykonalne lub zbyt kosztowne. Na przykład, zamiast granicy nieskończonego ciągu zbieżnego, bierze się pewien element tego ciągu. Zamiast sumy szeregu nieskończonego oblicza się sumę iluś początkowych składników. Zamiast całki oblicza się kwadraturę. Równania różniczkowe często zastępuje się równaniami różnicowymi; można podać wiele dalszych przykładów.

Błędy aproksymacji granicy ciągu nieskończonego przez pewien element tego ciągu, lub sumy nieskończonego szeregu przez pewną sumę częściową są często nazywane błędami obcięcia.

Błędy zaokrągleń. Wynik każdego działania wykonanego przez komputer podlega zaokrągleniu. Skutki bardzo często są małe w porównaniu ze skutkami innych błędów, ale czasem mogą zupełnie zmienić wynik.

Błędy grube. To są skutki wszelkich pomyłek, awarii, oraz błędów popełnionych w procesie pozyskiwania danych lub w implementacji algorytmu. Z innych przyczyn można tu też wymienić sabotaż (np. uprawiany przez producentów wirusów komputerowych i przez nierzetelnych autorów oprogramowania).

Uwarunkowanie zadania

Większość zadań numerycznych polega na obliczeniu wartości pewnej funkcji f, której dziedziną jest pewien obszar $D \subset \mathbb{R}^n$. Wynik obliczenia jest wektorem w $\mathbb{R}^m$, przy czym m może być określone przez konkretny argument $x \in D$ — na przykład, gdy trzeba znaleźć wszystkie rzeczywiste miejsca zerowe wielomianu, którego współczynniki są współrzędnymi wektora x. Załóżmy jednak, że m jest ustalone (i znane) dla wszystkich $x \in D$, a funkcja f jest ciągła. Zanim zaczniemy rozpatrywać jakiekolwiek algorytmy obliczania wyniku, zajmiemy się wpływem, jaki zaburzenia danych (które mogą pochodzić z niedokładnych pomiarów i które trzeba zastąpić liczbami zmiennopozycyjnymi) mają na wynik.

Pojęcie numerycznego uwarunkowania zadania określa wrażliwość wyniku na zaburzenia danych; dla zadania dobrze uwarunkowanego niewielkie zaburzenie danych powoduje niewielką zmianę wyniku. Zadanie jest źle uwarunkowane, jeśli po małej zmianie danych otrzymujemy zupełnie inny wynik. W związku ze sposobem reprezentowania liczb (który zapewnia mały błąd względny), bierzemy pod uwagę *względne* zaburzenia danych i spowodowane przez nie zmiany wyniku.

Liczbowa miara uwarunkowania nazywa się wskaźnikiem uwarunkowania zadania. Określa się go wzorem

$$\mathrm{cond}_{f(x)}\, x = \sup_{\|\tilde{x}-x\|<\varepsilon\|x\|} \left(\frac{\|f(\tilde{x}) - f(x)\|}{\|f(x)\|} \Big/ \frac{\|\tilde{x} - x\|}{\|x\|} \right).$$

Symbol cond pochodzi od angielskiego *condition number*; napis po lewej stronie czytamy: „wskaźnik uwarunkowania zadania obliczenia $f(x)$ dla danych x".
W określeniu wskaźnika uwarunkowania używamy jakichś norm (zależnie od zadania) i określamy największą dopuszczalną zmianę (zaburzenie względne) ε danych x. Następnie badamy iloraz względnego zaburzenia wyniku i powodującej to zaburzenie względnej zmiany danych.

Co daje znajomość wskaźnika uwarunkowania? Jeśli dane znamy z błędem względnym nie większym niż ε, to błąd względny wyniku (uwaga: *dokładnego* wyniku dla danych $\tilde{x}$, jakimi dysponujemy, w porównaniu z wynikiem dla nieznanych nam danych dokładnych x) nie jest większy niż $\varepsilon\, \mathrm{cond}_{f(x)}\, x$. Na przykład, jeśli wskaźnik uwarunkowania jest równy 100 (to jeszcze nie jest dużo), a dane reprezentujemy w formacie pojedynczej precyzji, tj. z błędem nie większym niż $\nu \approx 10^{-7}$ (i poza zaokrągleniem nie ma innych błędów), to wiemy, że jesteśmy w stanie otrzymać wynik z pięcioma cyframi dokładnymi. Jeśli jednak pomiar danych ma błąd rzędu 1%, to otrzymany wynik może mieć błąd 100%; na ogół taki wynik jest bezwartościowy. Albo należy wtedy zdobyć dokładniejsze dane, albo zająć się innym zadaniem (być może można jakoś przeformułować problem). Pamiętajmy przy tym, że założyliśmy brak błędów w algorytmie, który może dodatkowo zepsuć wynik.

Często przyjmuje się, że zaburzenia danych są bardzo małe (bo względne błędy reprezentacji zmiennopozycyjnej są bardzo małe), więc dla uproszczenia oblicza się wartość graniczną wskaźnika uwarunkowania, dla $\varepsilon \to 0$ (co ma sens, jeśli wskaźnik jest ciągły w otoczeniu $\boldsymbol{x}$). Jeśli zadanie polega na obliczeniu wartości skalarnej funkcji f, która ma skalarny argument x, przy czym funkcja f ma pochodną, to mamy wtedy

$$\mathrm{cond}_{f(x)}\, x = \left|\frac{x}{f(x)} f'(x)\right|.$$

Błędy reprezentacji wektorów

Niech $\boldsymbol{x} = [x_1, \dots, x_n]^T \in \mathbb{R}^n$ i niech $\tilde{\boldsymbol{x}} = [\tilde{x}_1, \dots, \tilde{x}_n]^T \in \mathbb{R}^n$, przy czym $\tilde{x}_i = x_i(1 + \varepsilon_i)$ dla każdego i. Zamiast rozpatrywać osobno błędy poszczególnych składowych wektora, co mogłoby zbyt być pracochłonne, często błąd opisuje się jedną liczbą, za pomocą jakiejś normy. Najczęściej wykorzystywane są normy Höldera, określone wzorem

$$\|\boldsymbol{x}\|_p = \left(|x_1|^p + \cdots + |x_n|^p\right)^{1/p},$$

dla pewnego $p \geqslant 1$, oraz norma określona jako granica dla $p \to \infty$:

$$\|\boldsymbol{x}\|_\infty = \max_{i\in\{1,\dots,n\}} |x_i|.$$

Za miarę błędu bezwzględnego możemy przyjąć liczbę $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_p$. Jeśli $\boldsymbol{x} \neq 0$ i dla każdego i jest $|\varepsilon_i| \leqslant \nu$, to miara błędu względnego spełnia nierówność

$$\frac{\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_p}{\|\boldsymbol{x}\|_p} = \frac{\left(|x_1\varepsilon_1|^p + \cdots + |x_n\varepsilon_n|^p\right)^{1/p}}{\|\boldsymbol{x}\|_p} \leqslant \frac{\left(|x_1\nu|^p + \cdots + |x_n\nu|^p\right)^{1/p}}{\|\boldsymbol{x}\|_p} = \frac{\|\boldsymbol{x}\|_p \nu}{\|\boldsymbol{x}\|_p} = \nu.$$

Zatem, błąd względny reprezentacji wektora, którego współrzędne zostały zokrąglone do najbliższych liczb zmiennopozycyjnych, mierzony za pomocą dowolnej normy Höldera (także $\|\cdot\|_\infty$), jest na poziomie błędu reprezentacji pojedynczej liczby.

Uwaga: Należy pamiętać, że z nierówności $\frac{\|x-\tilde{x}\|_p}{\|x\|_p} \leqslant \varepsilon > 0$ *nie wynika*, że błędy względne poszczególnych składowych są małe. Jeśli pewna składowa jest równa 0, to dowolne niezerowe jej zaburzenie daje nieograniczony błąd względny. Tak więc, wykonując odpowiednie rachunki, nie należy wyciągać pochopnych wniosków.

Numeryczna poprawność algorytmu

Skutki błędów zaokrągleń w obliczeniach czasem można zinterpretować jako skutki takiego zaburzenia danych, że otrzymany wynik jest dla tych zaburzonych danych dokładny. Jeśli takie hipotetyczne zaburzenie danych jest małe, to mówimy, że algorytm jest numerycznie poprawny. Pewne algorytmy są numerycznie poprawne, inne nie są. W zasadzie numeryczna poprawność „to jest to" — w praktyce niczego lepszego po algorytmach numerycznych spodziewać się nie można.

Tak, jak uwarunkowanie zadania, numeryczną poprawność można mierzyć, badając tzw. stałe kumulacji algorytmu. Algorytm jest tym lepszy, im te stałe są mniejsze. Aby je zdefiniować, wprowadzimy potrzebne oznaczenia.
Niech A oznacza algorytm i niech $A(x)$ oznacza wynik obliczenia, który powinien być jak najbliższy „prawdziwemu" rozwiązaniu zadania, $f(x)$. Obliczony wynik składa się z liczb zmiennopozycyjnych, zatem możemy dopuścić do rozważań jego błąd reprezentacji. Przypuśćmy zatem, że istnieją liczby K_d i K_w, takie że dla każdego $x \in D$ istnieją dane zaburzone $\tilde{x}$, dla których spełnione są nierówności

$$\frac{\|\tilde{x}-x\|}{\|x\|} \leqslant K_d\nu, \quad \text{oraz} \quad \frac{\|f(\tilde{x})-A(x)\|}{\|f(\tilde{x})\|} \leqslant K_w\nu.$$

Mówimy wtedy, że algorytm A jest numerycznie poprawnym algorytmem obliczania wartości funkcji f w dziedzinie (klasie zadań) D, ze stałymi kumulacji (danych) K_d i (wyniku) K_w.

Trzeba podkreślić, że w analizie algorytmu często występuje swoboda wybierania danych lub wyniku, do których „doczepiamy" błędy; z jednej strony to utrudnia analizę, a z drugiej stwarza możliwości pewnej „gimnastyki", wskutek czego pewne oszacowania mogą być poprawione — nieraz jest tak, że algorytm w praktyce działa bardzo dobrze, tj. wytwarza bardzo dokładne wyniki, zaś analiza tego nie potwierdza, bo na przykład daje bardzo grube oszacowania stałych kumulacji. Wspomniana „gimnastyka" czasem pomaga. W analizie błędu zwykle zakłada się, że błędy w poszczególnych działaniach są niezależne (i nieskorelowane), a ich wartości bezwzględne sumują się, tymczasem poszczególne błędy względne mogą być mniejsze niż ν, mogą się też znosić. Czasem analiza błędu pozwala wykryć newralgiczne miejsca i pomaga przeprojektować wzory.

Jeśli stała K_d jest równa 0, to znaczy, że niezależnie od uwarunkowania zadania otrzymany wynik jest bardzo dokładny, tj. otrzymany z dokładnością na poziomie błędu reprezentacji (tj. błąd wyniku jest co najwyżej K_w razy większy). Taka sytuacja występuje w praktyce nadzwyczaj rzadko. Częściej „winę” za niedokładność wyniku można „zwalić” na dane. Takie postępowanie, tj. znalezienie i oszacowanie zaburzenia danych, które prowadzi do otrzymanego wyniku, nazywa się analizą wstecz; jej twórcą był Wilkinson. Jeśli zadanie jest dobrze uwarunkowane i stałe kumulacji są nieduże, to stąd wynika, że obliczony wynik jest dobrym przybliżeniem wyniku poszukiwanego.

Numeryczna stabilność algorytmu

Często się nie udaje udowodnienie numerycznej poprawności algorytmu, tj. znalezienie stałych kumulacji niezależnych od danych w ustalonej dziedzinie D. Wówczas można spróbować zbadać, czy jest on numerycznie stabilny — ta własność jest pewnego rodzaju „minimum przyzwoitości” algorytmu. Aby ją zdefiniować, zbadajmy, jak duży byłby błąd wyniku, gdyby dane zostały zaburzone na poziomie błędu reprezentacji (co musi mieć miejsce — dane do obliczeń są liczbami zmiennopozycyjnymi) i wynik też należałoby zaokrąglić (bo też go reprezentujemy w ten sposób), ale poza zaokrągleniem końcowego wyniku wszystkie obliczenia byłyby wykonywane dokładnie.

Błąd (bezwzględny) wyniku spełniający wymienione warunki można oszacować przez liczbę

$$\|f(x)\|(\mathrm{cond}_{f(x)}\, x + 1)\nu.$$

Względny błąd danych, na poziomie ν, przenosi się na wynik z czynnikiem $\mathrm{cond}_{f(x)}\, x$; do tego wynik trzeba jeszcze zaokrąglić, stąd do wskaźnika uwarunkowania została dodana jedynka. Liczba będąca wartością podanego wyrażenia nazywa się optymalnym poziomem błędu.

Jeśli teraz algorytm zaokrągla wyniki wykonywanych działań, to może wynik zepsuć dodatkowo. Mówimy, że algorytm A jest numerycznie stabilnym algorytmem obliczania funkcji f, jeśli istnieje liczba K (stała kumulacji), taka że dla dowolnych danych $x \in D$ spełniona jest nierówność

$$\|f(x) - A(x)\| \leqslant K\|f(x)\|(\mathrm{cond}_{f(x)}\, x + 1)\nu.$$

Ważne jest też, aby stała kumulacji nie była bardzo duża.

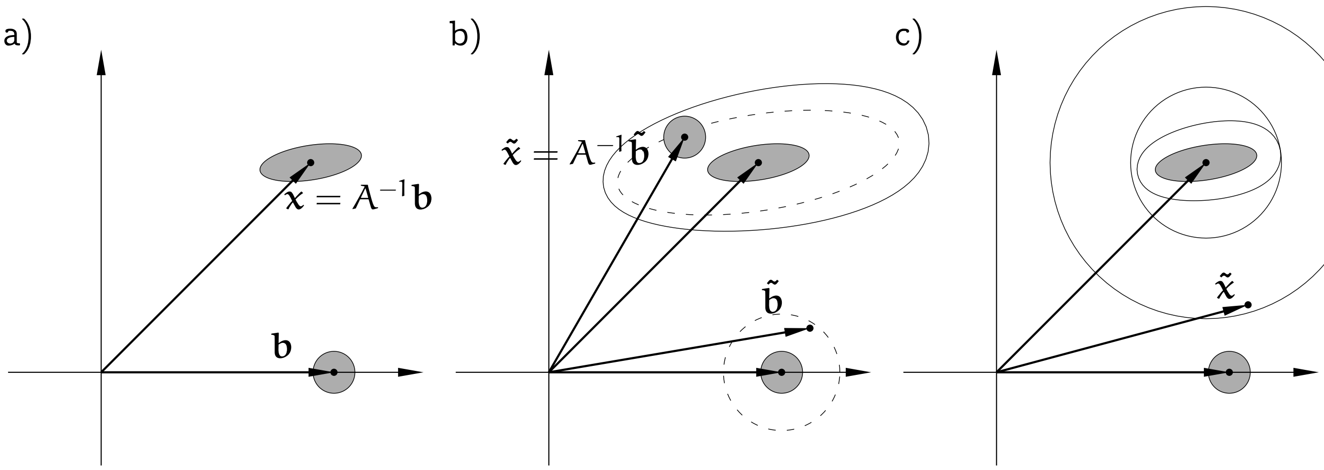
W tym ujęciu analizy błędów nie zajmujemy się tym, czy istnieją takie dane, bliskie danych x, dla których otrzymujemy (ewentualnie zaburzony na poziomie błędu reprezentacji) wynik. Dane takie mogą więc nie istnieć — możemy na przykład otrzymać sinus pewnego kąta rzeczywistego większy niż 1. Istotne jest to, że mając algorytm numerycznie stabilny, możemy dowolnie zmniejszyć skutki błędów zaokrągleń, wykorzystując w obliczeniach dostatecznie dokładną arytmetykę (czyli taką o dostatecznie długiej mantysie: przypominam, że $\nu = 2^{-t}$). Oczywiście, dla zadań źle uwarunkowanych arytmetyki standardowe mogą nie wystarczyć, ale wtedy czy na pewno znamy dane aż tak dokładnie?

Jeśli funkcja f, której wartość należy obliczyć, spełnia warunek Lipschitza, tj. istnieje stała L, taka że

$$\forall_{x,y\in D}\ \|f(x) - f(y)\| \leqslant L\|x - y\|,$$

to każdy algorytm numerycznie poprawny jest też numerycznie stabilny, ale numeryczna stabilność nie gwarantuje numerycznej poprawności.

W opanowaniu opisanych wyżej pojęć może pomóc rysunek, będący ilustracją zadania rozwiązywania układu dwóch równań liniowych $Ax = \mathbf{b}$, z nieosobliwą macierzą A. Rozwiązaniem zadania jest wektor $x = A^{-1}\mathbf{b}$, przy czym ten wzór jest pożyteczny w teoretycznej analizie zadania i algorytmów jego rozwiązywania, ale nie jest dobrym algorytmem numerycznym (i proszę go *nie używać* w tym charakterze). Danymi są współczynniki macierzy A i wektora prawej strony $\mathbf{b}$. Dla ilustracji pojęć rozpatrujemy tylko zaburzenia wektora prawej strony[5]. Wielkość tych zaburzeń jest taka, jak gdyby mantysa miała mniej więcej trzy bity.



[5] W rzeczywistości skutki błędów zaokrągleń w rozwiązywaniu układu równań liniowych przedstawia się jako efekt zaburzenia macierzy A — czego jednak nawet dla układu dwóch równań nie da się przejrzyście zilustrować.

Na rysunku a) mamy ilustrację uwarunkowania zadania. Zaznaczona kula (tj. koło) o środku $\mathbf{b}$ ma promień $\|\mathbf{b}\|\nu$. Zaburzenie danych polega na zastąpieniu wektora $\mathbf{b}$ przez jakiś element tej kuli. Obrazem tej kuli jest elipsoida (elipsa) o środku $\mathbf{x}$. Wskaźnik uwarunkowania zadania (ze względu na zaburzanie wektora $\mathbf{b}$, ale także macierzy A, co będziemy badać na jednym z dalszych wykładów) jest ilorazem długości najdłuższej i najkrótszej osi elipsoidy.

Numeryczna poprawność jest zilustrowana na rysunku b). Algorytm wyprodukował pewien wektor $\tilde{\mathbf{x}}$. Niech $\tilde{\mathbf{b}} = A\tilde{\mathbf{x}}$. Przypuśćmy, że stała kumulacji $K_w = 0$. Wtedy

$$\frac{\|\tilde{\mathbf{b}} - \mathbf{b}\|}{\|\mathbf{b}\|\nu} \leqslant K_d,$$

i mamy gwarancję, że dla otrzymanego wyniku $\tilde{\mathbf{x}}$, który leży w obrębie narysowanej linią przerywaną elipsy, istnieją dane $\tilde{\mathbf{b}}$, które leżą w narysowanym linią przerywaną kole (promień tego koła jest K_d razy większy niż $\|\mathbf{b}\|\nu$). Jeśli zaś weźmiemy $K_w > 0$, to dopuszczamy dodatkowe zaburzenie wyniku; leży on w nieco większym obszarze ograniczonym przez krzywą zobrazowaną przez linię ciągłą (ta krzywa nie jest elipsą). Dla takiego wyniku istnieje bliski punkt leżący w obszarze ograniczonym elipsą, który jest dokładnym wynikiem dla pewnych danych położonych w większym kole o środku $\mathbf{b}$.

Numeryczna stabilność jest przedstawiona na rysunku c). Rozważamy zaburzenia danych $\mathbf{b}$ na poziomie błędu reprezentacji. Dla tak zaburzonych danych wynik leży w obszarze zacienionym, ograniczonym przez elipsę. Ten obszar rozszerzamy, aby uwzględnić błąd reprezentacji wyniku, a następnie opisujemy koło. Promień tego koła jest optymalnym poziomem błędu. Wynik jest punktem koła o promieniu K razy większym. Dla pewnych punktów tego koła, położonych daleko od elipsy, nie istnieją dane $\tilde{\mathbf{b}}$, leżące blisko danych $\mathbf{b}$ i takie, że $\tilde{\mathbf{x}}$ jest dokładnym wynikiem dla danych $\tilde{\mathbf{b}}$.

Zadania i problemy

1. Zadanie polega na obliczeniu sumy n liczb: $s = x_1 + \cdots + x_n$. Dla algorytmu sumowania „po kolei” napisz wyrażenie, którego wartość jest sumą obliczoną z błędami zaokrągleń (przy założeniu, że nie wystąpił nadmiar ani niedomiar zmiennopozycyjny). Skonstruuj wektor $\tilde{x} = [\tilde{x}_1, \dots, \tilde{x}_n]$, taki że wynik obliczenia jest sumą jego współrzędnych i znajdź stałą kumulacji, dowodząc w ten sposób numerycznej poprawności.
2. Zrób to samo dla algorytmu sumowania parami, w którym oblicza się sumy kolejnych par danych liczb, sumy par tych sum, itd., aż do otrzymania sumy wszystkich składników. Porównaj stałe kumulacji tych dwóch algorytmów.
3. <u>Algorytm Kahana</u> sumowania liczb: para zmiennych zmiennopozycyjnych, `(s,c)`, symuluje liczbę zmiennopozycyjną o dwukrotnie dłuższej mantysie, dzięki czemu błędy względne zaokrągleń są na poziomie 2^{-2t}:

```
s = a[0];
c = 0.0;
for ( i = 1; i < n; i++ ) {
  y = a[i]-c;
  t = s+y;
  c = (t-s)-y;
  s = t;
}
```

4. Zadanie polega na obliczeniu przy użyciu schematu Hornera wartości wielomianu $w(x) = a_n x^n + \cdots + a_1 x + a_0$ dla ustalonego x. Skonstruuj współczynniki w bazie potęgowej wielomianu $\tilde{w}(x)$, którego wartość w punkcie x została obliczona i znajdź stałą kumulacji.
5. Zbadaj błędy zaokrągleń wytworzone przez algorytm obliczania wartości wyrażenia $a^2 + ab + b^2$, korzystający ze wzoru

$$w = \frac{1}{2}\big((a+b)^2 + (a^2 + b^2)\big).$$

Mnożenie przez całkowite potęgi liczby 2 (tu przez $\frac{1}{2}$), jeśli nie ma nadmiaru ani niedomiaru, jest wolne od błędów zaokrągleń.

6. (Lab.) Oblicz sumę początkowych n składników szeregu

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \cdots = \frac{\pi}{4}$$

przy użyciu różnych algorytmów sumowania. Porównaj wyniki.

7. (Lab.) k-ta suma częściowa szeregu w poprzednim zadaniu jest opisana wzorem

$$s_k = \sum_{i=0}^{k} (-1)^i \frac{1}{2i+1}.$$

Zbadaj zbieżność ciągu $(s_k)_{k\in\mathbb{N}}$ i porównaj ze zbieżnością ciągu $(t_k)_{k\in\mathbb{N}}$ określonego wzorem

$$t_k = \frac{1}{2}(s_k + s_{k+1}).$$

8. (Lab.) Rozważamy obliczanie kolejnych wyrazów ciągu

$$a_k = \int_0^1 \frac{x^k}{x+7}\,dx.$$

Wszystkie wyrazy tego ciągu są dodatnie; dla $k \to \infty$ ciąg dąży monotonicznie do zera. Dla $k > 0$ mamy

$$a_k + 7a_{k-1} = \int_0^1 \frac{x^k + 7x^{k-1}}{x+7}\,dx = \int_0^1 x^{k-1}\,dx = \frac{1}{k}.$$

Stąd otrzymujemy równanie różnicowe

$$a_k = -7a_{k-1} + \frac{1}{k}.$$

Liczby a_k możemy obliczać kolejno na podstawie tego równania, biorąc $a_0 = \ln 8 - \ln 7$.

Napisz program, który wykonuje to obliczenie, przypisując kolejne obliczone wyrazy ciągu zmiennej stałopozycyjnej pojedynczej, a także podwójnej precyzji; wystarczy obliczyć i wypisać pierwsze 20 wyrazów ciągu. Zbadaj, czy otrzymane wyniki są poprawne i wyjaśnij, co spowodowało pojawienie się takich wyników.

Następnie spróbuj postąpić odwrotnie: przyjmij $\tilde{a}_{30} = 0$ (liczba $\tilde{a}_{30}$ przybliża a_{30} z błędem 100%), a następnie oblicz kolejno

$$\tilde{a}_{k-1} = \frac{1}{7}\Big(\frac{1}{k} - \tilde{a}_k\Big) \quad \text{dla } k = 30, \dots, 1.$$

Ponownie zbadaj, jak dobrym przybliżeniem liczb a_k są otrzymane liczby $\tilde{a}_k$ i wyjaśnij, dlaczego.

Spróbuj wyciągnąć z tego eksperymentu wnioski, które miałyby zastosowanie podczas numerycznego rozwiązywania dowolnych równań różnicowych liniowych o stałych współczynnikach (także równań stopnia większego niż 1).

9. (Lab.) Oblicz wartości funkcji

$$f(x) = 1 - \sqrt{1 - x^2}$$

dla $x = 0.1, 0.01, 0.001, 0.0001, 0.00001, 0.000001$ przy użyciu arytmetyki zmiennopozycyjnej, na podstawie podanego wzoru definiującego funkcję f i na podstawie rozwinięcia Maclaurina z wyrazami do rzędu 4:

$$f(x) = \frac{1}{2!}x^2 + \frac{3}{4!}x^4 + \frac{f^{(5)}(\xi)}{5!}x^5 \approx \frac{1}{2!}x^2 + \frac{3}{4!}x^4$$

i porównaj wyniki. Wiedząc, że

$$f^{(5)}(x) = \frac{3x^3}{(1 - x^2)^{5/2}} + \frac{3x}{(1 - x^2)^{3/2}},$$

oszacuj błąd wzoru przybliżonego, tj. resztę rozwinięcia.

Przypomnienia

1. Przypomnienie pojęcia normy (potrzebnego w dalszych wykładach): norma jest to *dowolny* funkcjonał $\|\cdot\|$ określony w przestrzeni liniowej V nad ciałem liczbowym $\mathbb{K}$ (np. $\mathbb{R}$ lub $\mathbb{C}$; w tym wykładzie rozpatrujemy tylko przestrzenie rzeczywiste) i przyjmujący wartości rzeczywiste, który spełnia warunki:

- dodatniość: $\forall_{x\in V}\ \|x\| \geqslant 0$ oraz $\|x\| = 0 \Rightarrow x = 0$,
- półliniowość: $\forall_{x\in V,\, a\in\mathbb{K}}\ \|ax\| = |a|\|x\|$,
- nierówność trójkąta: $\forall_{x,y\in V}\ \|x+y\| \leqslant \|x\| + \|y\|$.

W przestrzeni $\mathbb{R}^n$ (a także $\mathbb{C}^n$) bardzo często rozważa się normy Höldera: dla dowolnego $p \geqslant 1$ funkcja określona wzorem

$$\|x\|_p = (|x_1|^p + \cdots + |x_n|^p)^{1/p}$$

spełnia podane warunki, jest więc normą. Również w granicy dla $p \to \infty$ otrzymujemy normę

$$\|x\|_\infty = \max_{i\in\{1,\dots,n\}} |x_i|.$$

Szczególnie często rozważa się normy Höldera dla $p = 1$, $p = \infty$ i $p = 2$.

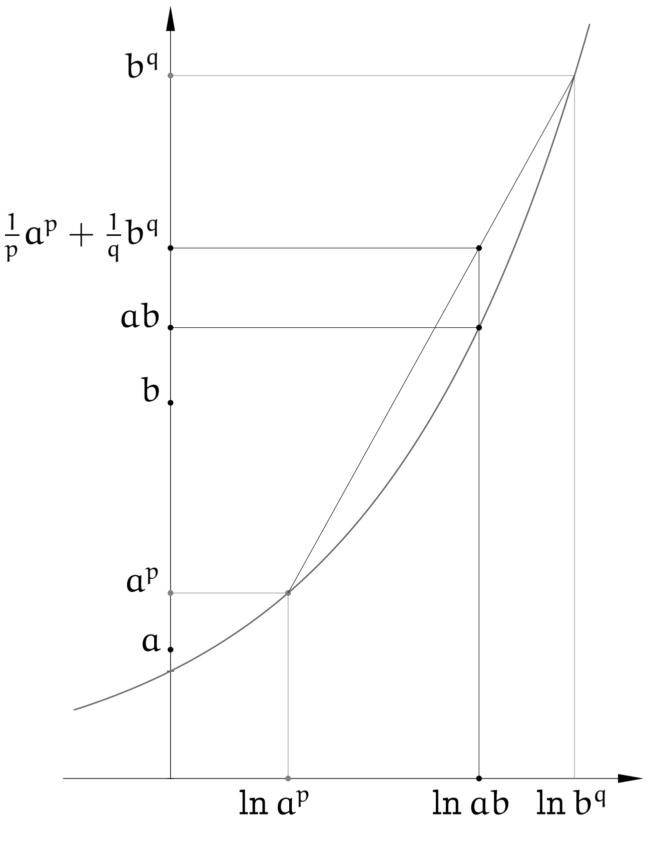
Niżej jest przypomniany dowód tzw. nierówności Minkowskiego, która dla p-tej normy Höldera jest nierównością trójkąta. Najpierw wykażemy dwie nierówności pomocnicze.

2. Nierówność Younga: Jeśli $p, q > 1$, $\frac{1}{p} + \frac{1}{q} = 1$, oraz $a, b \geqslant 0$, to

$$ab \leqslant \frac{1}{p}a^p + \frac{1}{q}b^q.$$

Dowód. Jeśli $a = 0$ lub $b = 0$, to nierówność jest oczywista. Przypuśćmy zatem, że $a, b > 0$. Funkcja $e^x\colon \mathbb{R} \to \mathbb{R}$ jest wypukła, tj. wnętrze każdego odcinka, którego końce leżą na wykresie tej funkcji, leży nad tym wykresem, skąd wynika, że

$$ab = e^{\ln a + \ln b} = e^{\frac{1}{p}\ln a^p + \frac{1}{q}\ln b^q}$$
$$\leqslant \frac{1}{p}e^{\ln a^p} + \frac{1}{q}e^{\ln b^q} = \frac{a^p}{p} + \frac{b^q}{q}. \qquad \square$$



3. Nierówność Höldera: Jeśli $p, q > 1$ i $\frac{1}{p} + \frac{1}{q} = 1$, to

$$\Big|\sum_i x_i y_i\Big| \leqslant \Big(\sum_i |x_i|^p\Big)^{1/p} \cdot \Big(\sum_i |y_i|^q\Big)^{1/q}.$$

Dowód. Najpierw weźmy ciągi $(\tilde{x}_i)_i$ i $(\tilde{y}_i)_i$, takie że $\sum_i |\tilde{x}_i|^p = \sum_i |\tilde{y}_i|^q = 1$. Wtedy

$$\Big|\sum_i \tilde{x}_i \tilde{y}_i\Big| \leqslant \sum_i |\tilde{x}_i \tilde{y}_i| \underset{\substack{\uparrow \\ \text{z nier. Younga}}}{\leqslant} \sum_i \frac{1}{p}|\tilde{x}_i|^p + \frac{1}{q}|\tilde{y}_i|^q = \frac{1}{p}\Big(\sum_i |\tilde{x}_i|^p\Big) + \frac{1}{q}\Big(\sum_i |\tilde{y}_i|^q\Big)$$

$$= \frac{1}{p} + \frac{1}{q} = 1.$$

Dla dowolnych niezerowych ciągów $(x_i)_i$ i $(y_i)_i$ określamy liczby $\tilde{x}_i = x_i/(\sum |x_i|^p)^{1/p}$ oraz $\tilde{y}_i = y_i/(\sum |y_i|^q)^{1/q}$ i mnożymy przez $(\sum |x_i|^p)^{1/p}(\sum |y_i|^q)^{1/q}$ strony nierówności udowodnionej wyżej. Jeśli zaś któryś z ciągów jest zerowy, to nierówność jest oczywista. □

Szczególny przypadek nierówności Höldera, dla $p = q = 2$, nosi nazwę nierówności Schwarza, albo nierówności Cauchy'ego-Schwarza. To z tej nierówności wynika, że iloczyn skalarny dwóch niezerowych wektorów w $\mathbb{R}^n$ spełnia równość

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \|\boldsymbol{x}\|_2 \|\boldsymbol{y}\|_2 \cos\phi$$

dla pewnej liczby rzeczywistej ϕ, która jest miarą kąta między tymi wektorami.

4. Nierówność Minkowskiego: Jeśli $p \geqslant 1$, to

$$\Big(\sum_i |x_i + y_i|^p\Big)^{1/p} \leqslant \Big(\sum_i |x_i|^p\Big)^{1/p} + \Big(\sum_i |y_i|^p\Big)^{1/p}.$$

Dowód. Dla $p = 1$ nierówność jest oczywista. Niech zatem $p > 1$. Wtedy jeśli $\frac{1}{p} + \frac{1}{q} = 1$, to $q = \frac{p}{p-1}$, co wykorzystujemy w obliczeniu

$$\sum_i |x_i + y_i|^p = \sum_i |x_i + y_i| \cdot |x_i + y_i|^{p-1} \leqslant \sum_i \big(|x_i| + |y_i|\big) \cdot |x_i + y_i|^{p-1}$$

$$\underset{\substack{\uparrow \\ \text{z nier. Höldera}}}{\leqslant} \bigg(\Big(\sum_i |x_i|^p\Big)^{1/p} + \Big(\sum_i |y_i|^p\Big)^{1/p}\bigg) \cdot \Big(\sum_i |x_i + y_i|^p\Big)^{\frac{p-1}{p}}.$$

Wystarczy pomnożyć strony powyższej nierówności przez $\big(\sum_i |x_i + y_i|^p\big)^{\frac{1-p}{p}}$. □

Nierówności Höldera i Minkowskiego dotyczą też ciągów nieskończonych i całek — rachunki w dowodach są takie same, ale rozważane szeregi mają być zbieżne.

5. Norma Höldera dla $p = 2$, czyli tak zwana norma euklidesowa, mierzy „zwykłą długość” wektora. Norma ta ma związek z iloczynem skalarnym w $\mathbb{R}^n$, określonym wzorem

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{y}^T \boldsymbol{x},$$

mianowicie dla każdego $\boldsymbol{x} \in \mathbb{R}^n$ zachodzi równość

$$\|\boldsymbol{x}\|_2 = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle}.$$

6. Niech V_a i V_b będą przestrzeniami określonymi nad tym samym ciałem liczbowym $\mathbb{K}$ i niech będą określone w nich jakieś (dowolne) normy, odpowiednio $\|\cdot\|_a$ i $\|\cdot\|_b$. Zbiór $L(V_a; V_b)$ wszystkich przekształceń liniowych $V_a \to V_b$ jest przestrzenią liniową nad $\mathbb{K}$. Funkcjonał określony w przestrzeni $L(V_a; V_b)$ wzorem

$$\|f\| = \sup_{x \neq 0} \frac{\|f(\boldsymbol{x})\|_b}{\|\boldsymbol{x}\|_a} = \sup_{\|x\| \leqslant 1} \|f(\boldsymbol{x})\|_b$$

jest normą. Mówimy, że jest to norma indukowana przez normy $\|\cdot\|_a$ i $\|\cdot\|_b$.
Przestrzeń przekształceń liniowych $\mathbb{R}^n \to \mathbb{R}^m$ możemy utożsamić z przestrzenią $\mathbb{R}^{m \times n}$ macierzy $m \times n$. Jeśli w przestrzeniach $\mathbb{R}^n$ i $\mathbb{R}^m$ przyjmiemy normy p-te Höldera, to normę indukowaną przez nie w $\mathbb{R}^{m \times n}$ będziemy również oznaczać symbolem $\|\cdot\|_p$. W przestrzeni $\mathbb{R}^{m \times 1}$ p-ta norma indukowana jest identyczna z normą p-tą Höldera w $\mathbb{R}^m$, zatem symbol $\|\cdot\|_p$ nie prowadzi do nieporozumień.

Normy indukowane pierwsza i nieskończona są łatwe do obliczenia:

$$\|A\|_1 = \max_{j \in \{1,\ldots,n\}} \sum_{i=1}^{m} |a_{ij}|,$$
$$\|A\|_\infty = \max_{i \in \{1,\ldots,m\}} \sum_{j=1}^{n} |a_{ij}|.$$

Na ćwiczeniach, jeśli będzie czas, warto udowodnić te wzory.

7. Normę w $\mathbb{R}^{m \times n}$ indukowaną przez normy drugie w $\mathbb{R}^n$ i $\mathbb{R}^m$ można obliczyć na podstawie wzoru

$$\|A\|_2 = \sqrt{\rho(A^T A)},$$

w którym $\rho(A^T A)$ oznacza promień spektralny macierzy $A^T A$, tj. największą wartość bezwzględną wartości własnej macierzy $A^T A$ (w tym przypadku każda, więc też największa wartość własna jest liczbą rzeczywistą nieujemną). Jeśli macierz A jest kwadratowa i symetryczna, to zachodzi też równość $\|A\|_2 = \rho(A)$. Jeśli nie znamy wartości własnych, to możemy skorzystać z oszacowania $\|A\|_2 \leqslant \sqrt{\|A^T A\|_\infty}$.

8. Jeśli macierz A jest wierszowa (tj. ma tylko jeden wiersz), to jej norma indukowana przez normę p-tą (dla $1 < p < \infty$) wyraża się wzorem $\|A\|_p = \|A^T\|_q$, gdzie $\frac{1}{p} + \frac{1}{q} = 1$. Macierze wierszowe utożsamiamy z przekształceniami liniowymi $\mathbb{R}^n \to \mathbb{R}$, czyli funkcjonałami liniowymi (określonymi wzorem $f(x) = Ax$).

   Nierówność $\|A\|_p \leqslant \|A^T\|_q$ wynika z nierówności Höldera. Dowód, że także $\|A\|_p \geqslant \|A^T\|_q$, można zacząć od macierzy $\tilde{A} \in \mathbb{R}^{1\times n}$ spełniającej warunek $\|\tilde{A}^T\|_q = 1$. Dla takiej macierzy dobieramy wektor $x \in \mathbb{R}^n$, jednostkowy ($\|x\|_p = 1$) i taki, aby nierówności w dowodzie nierówności Höldera były równościami.
9. Normy indukowane macierzy są submultiplikatywne, tzn. jeśli istnieje iloczyn macierzy A i B, to $\|AB\| \leqslant \|A\|\|B\|$. Jest tak w szczególności dla każdej normy p-tej, ale nie tylko. Własność ta jest prostym wnioskiem z definicji normy indukowanej.
10. Dla $p \geqslant 1$ oraz $p = \infty$ zachodzi też łatwa do wykazania nierówność $\|A\|_p \geqslant \|A_{ij}\|_p$, gdzie A_{ij} jest dowolnym blokiem macierzy A. Ponadto $\|A\|_p = \|PAQ^T\|_p$ dla dowolnych macierzy permutacji P i Q.
11. Jeśli macierz $B \in \mathbb{R}^{n\times n}$ jest nieosobliwa, to wzór

$$\|x\|_B \stackrel{\text{def}}{=} \|Bx\|,$$

   w którym po prawej stronie jest *dowolna* ustalona norma w $\mathbb{R}^n$, też określa normę, co łatwo jest sprawdzić (ćwiczenie).

   Mając daną macierz $A \in \mathbb{R}^{n,n}$, jak można obliczyć jej normę indukowaną przez normę $\|\cdot\|_B$? Zobaczmy:

$$\|A\|_B = \sup_{x\neq 0} \frac{\|Ax\|_B}{\|x\|_B} = \sup_{y\neq 0} \frac{\|AB^{-1}y\|_B}{\|B^{-1}y\|_B} = \sup_{y\neq 0} \frac{\|BAB^{-1}y\|}{\|BB^{-1}y\|} = \sup_{y\neq 0} \frac{\|BAB^{-1}y\|}{\|y\|}$$
$$= \|BAB^{-1}\|.$$

   Jeśli istnieje macierz nieosobliwa B, taka że dana macierz C jest równa BAB^{-1}, to o macierzach A i C mówi się, że są podobne. O relacji podobieństwa można myśleć w ten sposób, że macierze podobne opisują to samo przekształcenie liniowe w różnych bazach; macierz B jest macierzą zmiany bazy.
12. Udowodnij, że norma Frobeniusa, określona w przestrzeni $\mathbb{R}^{m,n}$ wzorem

$$\|A\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |a_{ij}|^2},$$

   jeśli $n > 1$ i $m > 1$, *nie jest* normą indukowaną przez żadną p-tą normę Höldera (ani w ogóle przez żadną normę wektorową).

Rozwiązywanie układów równań liniowych

Zajmujemy się rozwiązywaniem układu równań liniowych

$$A\mathbf{x} = \mathbf{b},$$

w którym dane są: nieosobliwa macierz A o wymiarach $n \times n$ i wektor $\mathbf{b} \in \mathbb{R}^n$. Układ ten ma jednoznaczne rozwiązanie, $\mathbf{x} = A^{-1}\mathbf{b}$, ale ten wzór, poza bardzo szczególnymi przypadkami, nie nadaje się do numerycznego rozwiązywania naszego zadania (ale w rachunkach symbolicznych nie zawahamy się go użyć, do czego mamy pełne prawo).

Zajmiemy się najpierw tzw. metodami bezpośrednimi rozwiązywania układów równań liniowych. Możemy je stosować wtedy, gdy liczba równań i niewiadomych jest mała (co najwyżej rzędu 10^3) lub gdy macierz układu jest „szczególnie łatwa”, np. trójdiagonalna. Metody te, gdyby nie było błędów zaokrągleń, dawałyby dokładny wynik po wykonaniu skończenie wielu działań. Błędy zaokrągleń oczywiście to psują.

Uwarunkowanie układu równań liniowych

Zbadamy, jak zmieni się rozwiązanie układu, jeśli dane, tj. macierz A lub wektor $\mathbf{b}$ zaburzymy. Dla układu równań

$$A\mathbf{x}' = \mathbf{b} + \delta\mathbf{b}$$

otrzymujemy rozwiązanie

$$\mathbf{x}' = A^{-1}\mathbf{b} + A^{-1}\delta\mathbf{b} = \mathbf{x} + A^{-1}\delta\mathbf{b},$$

skąd wynika, że

$$\|\mathbf{x}' - \mathbf{x}\| \leqslant \|A^{-1}\|\|\delta\mathbf{b}\| = \|A^{-1}\|\|\mathbf{b}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} = \|A^{-1}\|\|A\mathbf{x}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|} \leqslant \|A^{-1}\|\|A\|\|\mathbf{x}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

i ostatecznie

$$\frac{\|\mathbf{x}' - \mathbf{x}\|}{\|\mathbf{x}\|} \leqslant \|A\|\|A^{-1}\|\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}.$$

Zaburzymy teraz macierz A, tj. będziemy rozwiązywać układ $(A + \delta A)\mathbf{x}'' = \mathbf{b}$. Mamy

$$A(I + A^{-1}\delta A)\mathbf{x}'' = \mathbf{b}.$$

Musimy założyć, że zaburzenie macierzy A jest na tyle małe, że macierz $(I + A^{-1}\delta A)$ jest nieosobliwa, dzięki czemu możemy ją odwrócić i użyć wzoru przybliżonego

$$(I + A^{-1}\delta A)^{-1} \approx I - A^{-1}\delta A.$$

Dostaniemy wtedy

$$x'' \approx (I - A^{-1}\delta A)A^{-1}\mathbf{b} = (I - A^{-1}\delta A)x = x - A^{-1}\delta A x,$$

skąd wynika przybliżona nierówność

$$\frac{\|x'' - x\|}{\|x\|} \mathrel{\dot{\leqslant}} \|A\|\|A^{-1}\|\frac{\|\delta A\|}{\|A\|}.$$

Zatem, oba zaburzenia względne danych, tj. wektora $\mathbf{b}$ i macierzy A, mogą przenieść się na wynik z czynnikiem co najwyżej $\|A\|\|A^{-1}\|$. Ten czynnik jest <u>wskaźnikiem uwarunkowania zadania</u> rozwiązywania układu równań $Ax = \mathbf{b}$ i bywa też nazywany <u>wskaźnikiem uwarunkowania macierzy</u> A. Jeśli przyjmiemy normę p-tą indukowaną, to mamy <u>wskaźnik uwarunkowania macierzy A w normie p-tej</u>, który oznaczamy symbolem $\mathrm{cond}_p A$ ($\mathrm{cond}_p A = \|A\|_p\|A^{-1}\|_p$).

Normy indukowane $\|\cdot\|_1$ i $\|\cdot\|_\infty$ macierzy A są łatwe do obliczenia. Ponieważ na ogół nie znamy (i nie tracimy czasu na znajdowanie) macierzy A^{-1}, jej normę możemy zwykle tylko oszacować. Jeśli dysponujemy dodatkową informacją o zadaniu, z którego wziął się nasz układ równań, to warto z takiej informacji skorzystać w tym celu. Szacowanie normy macierzy A^{-1} jest też w zasadzie możliwe na podstawie czynników rozkładu znalezionych podczas rozwiązywania układu jedną z metod bezpośrednich.

<u>Uwaga:</u> Znalezione wyżej wskaźniki uwarunkowania dają dość pesymistyczne oszacowania błędów, jakie mogą obciążać wyniki. W oszacowaniu skutków zaburzenia wektora prawej strony korzystamy z nierówności $\|Ax\| \leqslant \|A\|\|x\|$, ale jeśli wektor prawej strony jest taki, że nierówność ta jest ostra, to wskaźnik uwarunkowania zadania z takim wektorem i macierzą A jest mniejszy. Mamy

$$\frac{\|x' - x\|}{\|x\|} \leqslant \|A^{-1}\|\frac{\|\mathbf{b}\|}{\|x\|}\frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|},$$

a stąd wynika wskaźnik uwarunkowania $\|A^{-1}\|\frac{\|\mathbf{b}\|}{\|A^{-1}\mathbf{b}\|}$. Jeśli wektor $\mathbf{b}$ ma taki kierunek, że $\|A^{-1}\mathbf{b}\| = \|A^{-1}\|\|\mathbf{b}\|$, to mamy zadanie ze wskaźnikiem uwarunkowania (ze względu na zaburzenia wektora $\mathbf{b}$) równym 1.

Nie ma podobnie prostego rachunku dla skutków zaburzeń macierzy A, ale zakładaliśmy, że one mogą mieć całkowicie dowolny kierunek. Tymczasem po pierwsze można ograniczyć zaburzenia *względne poszczególnych współczynników* (a więc uznać, że współczynniki o małych wartościach bezwzględnych mogą mieć proporcjonalnie małe zaburzenia), a ponadto dopuścić tylko symetryczne zaburzenia macierzy symetrycznej. Wskaźniki uwarunkowania takich zadań (które należałoby odpowiednio zdefiniować) ze względu na zaburzenia macierzy A mogą być znacznie mniejsze niż $\|A\|\|A^{-1}\|$.

Metoda eliminacji Gaussa: algorytm

Metoda eliminacji Gaussa jest najprostszym i chyba najczęściej używanym algorytmem rozwiązywania układów równań liniowych. Składa się on z dwóch etapów. W pierwszym układ jest przekształcany tak, aby powstał równoważny danemu układ równań liniowych z macierzą trójkątną górną. W etapie drugim, na podstawie kolejnych równań (od końca) obliczamy kolejne niewiadome (też od końca) — w każdym równaniu występuje tylko jedna niewiadoma, której wartość nie została obliczona wcześniej.

W pierwszym etapie (który jest właściwą eliminacją Gaussa), konstruujemy ciąg macierzy $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-1)} = U$, takich że macierz $A^{(k)}$ ma w kolumnach $1, \ldots, k$ współczynniki poniżej diagonali równe 0. Mając macierz $A^{(k-1)} = [a_{ij}^{(k-1)}]_{i,j}$, obliczamy współczynniki macierzy $A^{(k)}$:

$$\left.\begin{array}{rcl} l_{ik} & = & a_{ik}^{(k-1)}/a_{kk}^{(k-1)}, \\ a_{ij}^{(k)} & = & a_{ij}^{(k-1)} - l_{ik}a_{kj}^{(k-1)}, \quad \text{dla } j = k+1, \ldots, n \end{array}\right\} \quad \text{dla } i = k+1, \ldots, n$$

Ponadto $a_{ij}^{(k)} = a_{ij}^{(k-1)}$ dla $i \leqslant k$, oraz $a_{ik}^{(k)} = 0$ dla $i > k$.

Przekształcanie wektora prawej strony polega na skonstruowaniu ciągu wektorów $\mathbf{b}^{(0)} = \mathbf{b}, \mathbf{b}^{(1)}, \ldots, \mathbf{b}^{(n-1)} = \mathbf{y}$. W k-tym kroku eliminacji obliczamy współrzędne wektora $\mathbf{b}^{(k)}$:

$$b_i^{(k)} = b_i^{(k-1)} - l_{ik}b_k^{(k-1)}, \quad \text{dla } i = k+1, \ldots, n,$$

zaś dla $i = 1, \ldots, k$ mamy $b_i^{(k)} = b_i^{(k-1)}$. W wyniku eliminacji otrzymujemy macierz trójkątną

$$U = \begin{bmatrix} u_{11} & u_{12} & u_{13} & \ldots & u_{1n} \\ 0 & u_{22} & u_{23} & \ldots & u_{2n} \\ 0 & 0 & u_{33} & \ldots & u_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 & u_{nn} \end{bmatrix},$$

i wektor $\mathbf{y}$, takie że układ $U\mathbf{x} = \mathbf{y}$ jest równoważny układowi danemu.

Przekształcenie, które z macierzy $A^{(k-1)}$ produkuje macierz $A^{(k)}$ jest liniowe; zachodzi równość $A^{(k)} = L_k^{-1} A^{(k-1)}$, przy czym macierz L_k i jej odwrotność są trójkątne dolne, z jedynkami na diagonali, a poza tym z niezerowymi współczynnikami tylko w k-tej kolumnie:

$$L_k = \begin{bmatrix} \ddots & & & \\ & 1 & & \\ & l_{k+1,k} & 1 & & \\ & \vdots & & \ddots & \\ & l_{n,k} & & & 1 \end{bmatrix}, \quad L_k^{-1} = \begin{bmatrix} \ddots & & & \\ & 1 & & \\ & -l_{k+1,k} & 1 & & \\ & \vdots & & \ddots & \\ & -l_{n,k} & & & 1 \end{bmatrix}.$$

Na końcu otrzymujemy macierz $U = A^{(n-1)} = L_{n-1}^{-1} \dots L_1^{-1} A$, czyli jest $A = L_1 \dots L_{n-1} U$. Iloczyn $L = L_1 \dots L_{n-1}$ macierzy przekształceń wykonanych w kolejnych krokach ma pod diagonalą współczynniki $l_{ik} = a_{ik}^{(k-1)} / a_{kk}^{(k-1)}$:

$$L = \begin{bmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0 \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & 1 \end{bmatrix}.$$

W wyniku eliminacji Gaussa dostajemy więc macierze trójkątne (dolną i górną) L i U, takie że $A = LU$. Przekształcanie wektora prawej strony jest równoważne rozwiązywaniu układu równań $L\mathbf{y} = \mathbf{b}$. Zatem możemy najpierw wyznaczyć tylko macierze L i U, a przetwarzanie wektora prawej strony przenieść do drugiego etapu, w którym trzeba rozwiązać kolejno układy równań z macierzami trójkątnymi, $L\mathbf{y} = \mathbf{b}$ i $U\mathbf{x} = \mathbf{y}$.

Eliminację można wykonać *w miejscu* (po łacinie *in situ*). Po obliczeniu współczynnika l_{ik}, można go zapamiętać na miejscu współczynnika $a_{ik}^{(k-1)}$ (czyli na miejscu zajmowanym początkowo przez a_{ik}). Obliczone współczynniki $a_{ij}^{(k)}$ dla $i \leqslant j$ wpisujemy w miejsce $a_{ij}^{(k-1)}$. W ten sposób otrzymamy tablicę z liczbami

$$\begin{bmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \dots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ l_{n1} & l_{n2} & \dots & l_{n,n-1} & u_{nn} \end{bmatrix},$$

do której może sięgać podprogram rozwiązujący układy trójkątne. Podprogram eliminacji *in situ* „psuje" początkową zawartość tablicy. Jeśli oryginalna macierz A jest potrzebna (często jest), należy ją skopiować i „zepsuć" kopię.

Opisany wyżej algorytm jest zawodny; nieosobliwość macierzy A nie gwarantuje wykonalności dzielenia przez współczynnik $a_{kk}^{(k)}$, który może być zerem. Co więcej, jeśli współczynnik ten ma małą wartość bezwzględną w porównaniu z innymi, to skutki błędów zaokrągleń mogą prowadzić do otrzymania bardzo niedokładnych wyników. Dlatego stosuje się wybór elementu głównego (ang. *pivoting*). Najczęściej stosowany wybór częściowy w kolumnie polega na wyszukaniu w zbiorze $\{a_{kk}^{(k-1)}, \ldots, a_{nk}^{(k-1)}\}$ współczynnika $a_{lk}^{(k-1)}$ o największej wartości bezwzględnej, a następnie (jeśli $l \neq k$) przestawieniu równań l i k. Jeśli macierz A jest nieosobliwa, to któraś z tych liczb nie jest zerem i dzielenie przez nią jest wykonalne. Ponieważ dzielimy przez liczbę o największej wartości bezwzględnej, współczynniki l_{ik} mają wartości bezwzględne nie większe niż 1.

Można dowieść, że skutki takiego przestawiania (tzn. otrzymane na końcu macierze L i U) są takie same, jak gdyby równania zostały poprzestawiane *przed* przystąpieniem do eliminacji. Zatem, po zastosowaniu częściowego wyboru elementu głównego, otrzymamy macierze L i U, takie że $LU = PA$, gdzie P oznacza macierz dokonanej permutacji równań.

Macierz P trzeba jakoś reprezentować, aby można było odpowiednio poprzestawiać współrzędne wektora prawej strony, ponieważ trzeba będzie rozwiązać układy równań $L\boldsymbol{y} = P\boldsymbol{b}$ i $U\boldsymbol{x} = \boldsymbol{y}$. Najprostszy sposób polega na użyciu tablicy liczb całkowitych o długości n; pozycji k-tej przypisujemy indeks l wiersza, który został przestawiony z k-tym (albo k, jeśli nie było przestawienia). Przestawianie liczb zmiennopozycyjnych w tablicy jest operacją wolną od błędów zaokrągleń.

Istnieje też wybór pełny elementu głównego; przestawiamy w nim wiersze i kolumny tak, aby współczynnik $a_{kk}^{(k-1)}$, przez który będziemy dzielić, miał największą wartość bezwzględną w prawej dolnej podmacierzy $n+1-k \times n+1-k$. W ten sposób otrzymujemy rozkład macierzy $LU = PAQ^T$. Do rozwiązania mamy układy $L\boldsymbol{y} = P\boldsymbol{b}$ i $U\boldsymbol{z} = \boldsymbol{y}$, a następnie trzeba obliczyć $\boldsymbol{x} = Q^T\boldsymbol{z}$, czyli odpowiednio poprzestawiać współrzędne rozwiązania. Macierz permutacji Q można reprezentować w taki sam sposób jak P. Pełny wybór elementu głównego jest dosyć kosztowny (choć nie powiększa rzędu złożoności algorytmu) i *bardzo rzadko* zdarza się sytuacja, gdy dokładność wyniku otrzymanego z wyborem częściowym jest za mała, a wybór pełny daje dostatecznie mały błąd.

Metoda eliminacji Gaussa: analiza błędów

Udowodnimy, że metoda eliminacji Gaussa z wyborem elementu głównego jest algorytmem numerycznie poprawnym, tj. istnieje macierz $\tilde{A}$, taka że zachodzi równość $P\tilde{A}Q^T = \tilde{L}\tilde{U}$ dla obliczonych macierzy trójkątnych $\tilde{L}$ i $\tilde{U}$, oraz istnieje liczba F_n, taka że

$$\frac{\|\tilde{A} - A\|}{\|A\|} \leqslant F_n \nu.$$

Zatem, rozpatrujemy macierze $A^{(0)} = PAQ^T$ oraz $A^{(k)} = L_k^{-1}A^{(k-1)}$, składające się z dokładnych wyników wykonanych działań, oraz macierze $\tilde{A}^{(k)}$ i $\tilde{L}_k$ otrzymane w kolejnych krokach eliminacji przy użyciu arytmetyki zmiennopozycyjnej, z błędami zaokrągleń. Weźmy $k = 1$. Zamiast współczynnika $l_{i1} = a_{i1}/a_{ii}$ macierzy L_1 (oraz L) otrzymamy liczbę

$$\tilde{l}_{i1} = \mathrm{fl}(a_{i1}/a_{11}) = (a_{i1}/a_{11})(1 + \alpha_{i1}) = \tilde{a}_{i1}/a_{11},$$

gdzie $\tilde{a}_{i1} = a_{i1}(1 + \alpha_{i1})$, $|\alpha_{i1}| < \nu = 2^{-t}$ (dla układu równań o współczynnikach zespolonych zamiast ν należy przyjąć $(4 + \sqrt{2})\nu$ — zobacz wykład 2; dalej badamy przypadek rzeczywisty).

Następnie, zamiast współczynnika $a_{ij}^{(1)} = a_{ij} - l_{i1}a_{1j}$ macierzy $A^{(1)}$ komputer wyprodukuje liczbę

$$\begin{aligned} \tilde{a}_{ij}^{(1)} &= \mathrm{fl}(a_{ij} - \tilde{l}_{i1}a_{1j}) = \big(a_{ij} - \tilde{l}_{i1}a_{1j}(1 + \beta_{ij})\big)(1 + \gamma_{ij}) \\ &= a_{ij} - \tilde{l}_{i1}a_{1j} + \underbrace{a_{ij}\gamma_{ij} - \tilde{l}_{i1}a_{1j}(\gamma_{ij} + \beta_{ij} + \beta_{ij}\gamma_{ij})}_{b_{ij}} = \tilde{a}_{ij} - \tilde{l}_{i1}a_{1j}. \end{aligned} \tag{*}$$

Liczba b_{ij} jest błędem, który obciąża współczynnik $a_{ij}^{(1)}$ macierzy $A^{(1)}$, ale możemy „doczepić” go do współczynnika a_{ij} macierzy PAQ^T. Zatem, liczba $\tilde{a}_{ij} = a_{ij} + b_{ij}$ jest współczynnikiem pewnej macierzy, takiej że wynik pierwszego kroku eliminacji z błędami zaokrągleń jest dokładnym wynikiem dla tej macierzy. Jest

$$b_{ij} = a_{ij}\gamma_{ij} - \tilde{l}_{i1}a_{1j}(\gamma_{ij} + \beta_{ij} + \beta_{ij}\gamma_{ij}). \tag{**}$$

Na podstawie (*) mamy

$$\frac{\tilde{a}_{ij}^{(1)}}{1 + \gamma_{ij}} = a_{ij} - \tilde{l}_{i1}a_{1j}(1 + \beta_{ij}), \quad \text{czyli} \quad a_{ij} = \frac{\tilde{a}_{ij}^{(1)}}{1 + \gamma_{ij}} + \tilde{l}_{i1}a_{1j} + \tilde{l}_{i1}a_{1j}\beta_{ij}.$$

Wstawiając to do (**), otrzymamy

$$\begin{aligned} b_{ij} &= \left(\frac{\tilde{a}_{ij}^{(1)}}{1 + \gamma_{ij}} + \tilde{l}_{i1}a_{1j} + \tilde{l}_{i1}a_{1j}\beta_{ij}\right)\gamma_{ij} - \tilde{l}_{i1}a_{1j}(\gamma_{ij} + \beta_{ij} + \beta_{ij}\gamma_{ij}) \\ &= \frac{\tilde{a}_{ij}^{(1)}\gamma_{ij}}{1 + \gamma_{ij}} - \tilde{l}_{i1}a_{1j}\beta_{ij}. \end{aligned}$$

Do ostatniego wyrażenia podstawimy $\tilde{l}_{i1}a_{1j} = \tilde{a}_{ij}^{(1)} - \tilde{a}_{ij}$ (porównaj z (*)), otrzymując ostatecznie

$$b_{ij} = \frac{\tilde{a}_{ij}^{(1)}\gamma_{ij}}{1+\gamma_{ij}} - (\tilde{a}_{ij}^{(1)} - \tilde{a}_{ij})\beta_{ij} = \tilde{a}_{ij}\beta_{ij} + \tilde{a}_{ij}^{(1)}\Big(\frac{\gamma_{ij}}{1+\gamma_{ij}} - \beta_{ij}\Big).$$

Ponieważ wartości bezwzględne liczb β_{ij} i γ_{ij} są małe (mniejsze niż ν), możemy pominąć mianownik $1+\gamma_{ij}$ i zastąpić $\tilde{a}_{ij}$ przez a_{ij} w oszacowaniu

$$|b_{ij}| \mathrel{\dot{\leqslant}} \big(|\tilde{a}_{ij}| + 2|\tilde{a}_{ij}^{(1)}|\big)\nu \approx \big(|a_{ij}| + 2|\tilde{a}_{ij}^{(1)}|\big)\nu \quad \text{dla } i,j > 1.$$

Mamy też oszacowanie zaburzenia $b_{i1} = \tilde{a}_{i1} - a_{i1}$:

$$|b_{i1}| \leqslant |a_{i1}|\nu.$$

W pierwszym kroku eliminacji Gaussa otrzymaliśmy zatem macierze $\tilde{L}_1$ oraz $\tilde{A}^{(1)}$, których (dokładny) iloczyn jest równy $PAQ^T + B_0$, przy czym współczynnikami macierzy B_0 są liczby b_{ij}. Na podstawie wykonanych rachunków możemy napisać

$$|B_0| \mathrel{\dot{\leqslant}} \big(|PAQ^T| + 2|M_1|\big)\nu,$$

gdzie macierz M_1 powstaje z macierzy $A^{(1)}$ przez zastąpienie zerami współczynników w pierwszym wierszu i kolumnie. Jeśli używamy normy $\|\cdot\|_1$ lub $\|\cdot\|_\infty$, to (biorąc pod uwagę jawne wzory na normy indukowane podane wcześniej[6]) możemy napisać oszacowanie

$$\|B_0\| \mathrel{\dot{\leqslant}} \big(\|A\| + 2\|M_1\|\big)\nu.$$

W taki sam sposób możemy otrzymać oszacowania błędów we wszystkich krokach eliminacji, dla $k = 2, \ldots, n-1$:

$$\|B_{k-1}\| \mathrel{\dot{\leqslant}} \big(\|M_{k-1}\| + 2\|M_k\|\big)\nu,$$

przy czym $M_0 = PAQ^T$, zaś dla $k > 0$ macierz M_k otrzymujemy z macierzy $\tilde{A}^{(k)}$ przez zamienienie na zero współczynników w początkowych k kolumnach i wierszach. Macierz B_k jest zaburzeniem, które dodane do macierzy $\tilde{A}^{(k)}$ jest równoważne skutkom błędów zaokrągleń popełnionych podczas wyznaczania

[6]Korzystamy tu z własności monotoniczności normy indukowanej $\|\cdot\|_1$ lub $\|\cdot\|_\infty$: jeśli $|A| \leqslant |C|$, to $\|A\| \leqslant \|C\|$ i w szczególności $\||A|\| = \|A\|$. W ogólności normy indukowane przez p-te normy Höldera *nie mają* tej własności. Dla każdego p jest natomiast $\|A\|_p = \|PAQ^T\|_p$, dla dowolnych macierzy permutacji P i Q.

macierzy $\tilde{A}^{(k+1)}$. Całą eliminację Gaussa z błędami zaokrągleń możemy opisać w taki sposób:

$$\tilde{A}^{(0)} = PAQ^T,$$
$$\tilde{A}^{(k)} = \tilde{L}_k^{-1}(\tilde{A}^{(k-1)} + B_{k-1}), \quad k = 1, \dots, n-1.$$

W wyniku otrzymujemy macierz trójkątną górną $\tilde{U} = \tilde{A}^{(n-1)}$. Po rozwinięciu powyższych wzorów otrzymamy

$$\tilde{U} = \tilde{L}_{n-1}^{-1} \dots \tilde{L}_1^{-1}\tilde{A}^{(0)} + \tilde{L}_{n-1}^{-1} \dots \tilde{L}_1^{-1}B_0 + \tilde{L}_{n-1}^{-1} \dots \tilde{L}_2^{-1}B_1 + \dots + \tilde{L}_{n-1}^{-1}B_{n-2}.$$

Mnożąc strony tej równości przez macierz $\tilde{L} = \tilde{L}_1 \dots \tilde{L}_{n-1}$, otrzymamy

$$\tilde{L}\tilde{U} = PAQ^T + B_0 + \tilde{L}_1B_1 + \tilde{L}_1\tilde{L}_2B_2 + \dots + \tilde{L}_1 \dots \tilde{L}_{n-2}B_{n-2} = PAQ^T + E.$$

W początkowych k wierszach macierzy B_k mamy tylko zera, zaś kolumny od $k+1$ do n macierzy $\tilde{L}_1 \dots \tilde{L}_k$ są odpowiednimi kolumnami macierzy jednostkowej. Stąd wynika, że

$$\tilde{L}_1 \dots \tilde{L}_k B_k = B_k,$$

a zatem $E = P(\tilde{A} - A)Q^T = B_0 + \dots + B_{n-2}$. Możemy na tej podstawie oszacować

$$\|E\| \leqslant \sum_{k=0}^{n-2} \|B_k\| \mathrel{\dot{\leqslant}} \Big(\|A\| + 3\sum_{k=1}^{n-1} \|M_k\|\Big)\nu.$$

Przyjmując oznaczenie

$$F_n(A) = \frac{\|A\| + 3\sum_{k=1}^{n-1} \|M_k\|}{\|A\|},$$

możemy przedstawić to oszacowanie w postaci

$$\frac{\|\tilde{A} - A\|}{\|A\|} \mathrel{\dot{\leqslant}} F_n(A)\nu.$$

W oszacowaniu zaburzenia występuje liczba $F_n(A)$, która zależy od konkretnej macierzy; chcielibyśmy znaleźć liczbę F_n, taką że $F_n(A) \leqslant F_n$ dla *każdej* macierzy A o wymiarach $n \times n$. Liczba $F_n(A)$ jest tym większa, im większe wartości bezwględne mają współczynniki macierzy M_k, czyli obliczane w kolejnych krokach współczynniki wszystkich macierzy $A^{(k)}$. Zaburzenie danych, równoważne wytworzonym błędom zaokrągleń, może być bardzo duże, jeśli w trakcie obliczeń będziemy otrzymywać bardzo duże współczynniki. Właśnie temu ma przeciwdziałać wybór elementu głównego.

Jeśli jest stosowany wybór częściowy, to teoretycznie może się zdarzyć, że dla każdego k otrzymamy macierz M_k, taką że $\|M_k\| \approx 2\|M_{k-1}\|$, a wtedy dostaniemy $F_n(A) = F_n \approx 3 \cdot 2^n - 5$. Zazwyczaj to się nie zdarza i współczynniki wszystkich macierzy M_k są zbliżonego rzędu wielkości, a wtedy $F_n(A)$ jest rzędu n.

Obliczenie wektora $\boldsymbol{x}$ przez rozwiązanie układów równań z macierzami trójkątnymi jest również numerycznie poprawnym algorytmem rozwiązywania układu równań liniowych $A\boldsymbol{x} = \boldsymbol{b}$. Dowód tego faktu (szczegółowe rachunki pominę) polega na wykazaniu, że istnieją macierze trójkątne $\hat{L} = \tilde{L} + \delta L$ i $\hat{U} = \tilde{U} + \delta U$, takie że wektory $\hat{\boldsymbol{y}}$ i $\hat{\boldsymbol{x}}$ otrzymane przez rozwiazanie układów równań $\tilde{L}\boldsymbol{y} = P\boldsymbol{b}$ oraz $\tilde{U}Q\boldsymbol{x} = \hat{\boldsymbol{y}}$ są dokładnymi rozwiązaniami układów równań

$$\hat{L}\hat{\boldsymbol{y}} = P\boldsymbol{b} \quad \text{oraz} \quad \hat{U}Q\hat{\boldsymbol{x}} = \hat{\boldsymbol{y}},$$

czyli

$$(\tilde{L} + \delta L)(\tilde{U} + \delta U)Q\hat{\boldsymbol{x}} = P\boldsymbol{b},$$
$$(A + \delta A)\hat{\boldsymbol{x}} = \boldsymbol{b}, \quad \text{gdzie} \quad \delta A = P^T(E + \tilde{L}\delta U + \delta L\tilde{U} + \delta L\delta U)Q,$$

i spełniona jest nierówność

$$\frac{\|\delta A\|}{\|A\|} \mathrel{\dot{\leqslant}} F'_n(A)\nu,$$

z czynnikiem $F'_n(A) \leqslant (n+2)F_n(A)$.

Metoda eliminacji Gaussa: koszt

Koszty algorytmów numerycznej algebry liniowej są zazwyczaj wyrażane w jednostkach opms — 1opms jest to koszt wykonania jednego mnożenia i jednego dodawania albo odejmowania zmiennopozycyjnego, ponieważ najczęściej działania te występują w parach. Za koszt jednego dzielenia też można przyjąć 1opms.

Koszt $T(n)$ rozłożenia pełnej macierzy $n \times n$ na czynniki trójkątne jest sumą kosztu „wytworzenia” zer w pierwszej kolumnie pod diagonalą i kosztu rozłożenia macierzy $n-1 \times n-1$. Stąd mamy równanie różnicowe

$$T(n) = T(n-1) + n(n-1) \quad \text{dla } n > 1.$$

Jego rozwiązanie, spełniające warunek $T(1) = 0$ jest następujące:

$$T(n) = \frac{1}{3}(n^3 - n).$$

Łatwo jest zauważyć, że łączny koszt rozwiązywania układów równań z trójkątnymi czynnikami rozkładu macierzy pełnej jest równy n^2opms. Możliwość rozdzielenia etapów rozkładania macierzy na czynniki i rozwiązywania układów z tymi czynnikami ma duże znaczenie praktyczne, jeśli trzeba rozwiązać wiele układów równań z tą samą macierzą i różnymi prawymi stronami.

W wielu zastosowaniach mamy do czynienia z macierzami rzadkimi, tj. mającymi dużo zerowych współczynników. W takich przypadkach *czynem karalnym*[7] jest użycie ogólnego algorytmu, odpowiedniego dla macierzy pełnych. Jeśli np. macierz jest wstęgowa, tj. istnieje $k \ll n$, takie że $a_{ij} = 0$ dla $|i-j| > k$, to odpowiedni wariant metody eliminacji Gaussa może znaleźć czynniki trójkątne kosztem rzędu nk^2, a koszt rozwiązywania układów równań z tymi czynnikami jest rzędu nk. I tego wariantu należy użyć.

Metoda odbić Householdera

Inną metodą doprowadzenia układu równań liniowych do układu równoważnego z macierzą trójkątną jest metoda odbić Householdera. Przypomnijmy własności odbić symetrycznych w $\mathbb{R}^n$. Niech $\boldsymbol{v}$ oznacza dowolny wektor jednostkowy (w sensie normy drugiej, tj. taki że $\|\boldsymbol{v}\|_2 = 1$). Odbicie symetryczne względem hiperpłaszczyzny prostopadłej do wektora $\boldsymbol{v}$ jest określone wzorem

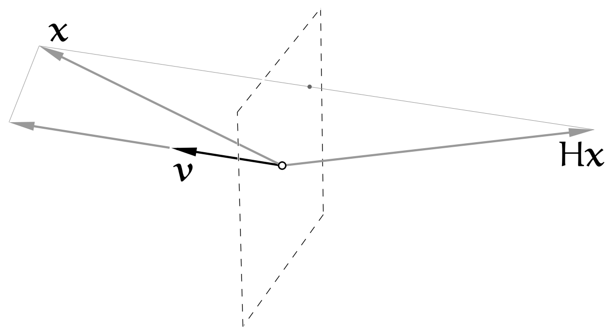
$$H\boldsymbol{x} = \boldsymbol{x} - 2\boldsymbol{v}\boldsymbol{v}^T\boldsymbol{x}.$$

Macierz odbicia jest więc równa

$$H = I - 2\boldsymbol{v}\boldsymbol{v}^T.$$

Odbicie jest inwolucją, tj. swoją własną odwrotnością:

$$H^2 = (I - 2\boldsymbol{v}\boldsymbol{v}^T)(I - 2\boldsymbol{v}\boldsymbol{v}^T) = I - 4\boldsymbol{v}\boldsymbol{v}^T + 4\boldsymbol{v}\underbrace{\boldsymbol{v}^T\boldsymbol{v}}_{=1}\boldsymbol{v}^T = I.$$



[7] Na podstawie §186 Kodeksu Etyki Zawodowej, Kompetencji i Przyzwoitości należy się za to od 1 do 3 lat Kompromitacji, a za notoryczną recydywę grozi Kompromitacja Dożywotnia.

Macierz odbicia jest ponadto symetryczna, a zatem $H^T H = H^2 = I$, czyli macierz H jest ortogonalna. Tak więc odbicie jest izometrią; dla dowolnych wektorów $x, y \in \mathbb{R}^n$ zachodzi równość

$$\langle Hx, Hy\rangle = y^T H^T Hx = y^T x = \langle x, y\rangle,$$

skąd dalej wynika, że dla dowolnego wektora x jest $\|Hx\|_2 = \|x\|_2$. Jeśli wektor v jest niezerowy, ale niekoniecznie jednostkowy, to macierz odbicia symetrycznego względem hiperpłaszczyzny prostopadłej do niego jest określona wzorem

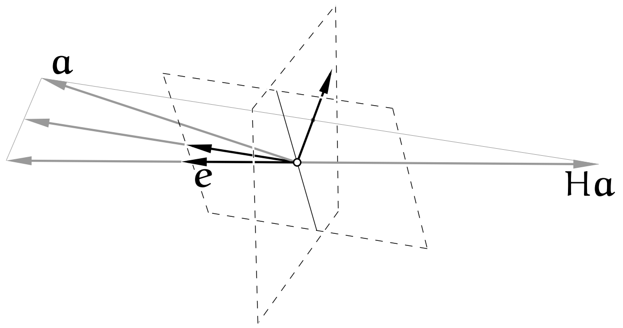
$$H = I - v\frac{2}{v^T v}v^T.$$

W algorytmach numerycznych, jeśli to nie jest wynikiem, który koniecznie musimy otrzymać, nigdy nie wyznaczamy jawnie macierzy H. Mając wektor v, możemy obliczyć liczbę $\gamma = \frac{2}{v^T v}$, a następnie, chcąc obliczyć obraz y dowolnego wektora x w odbiciu, obliczamy kolejno

$$s = v^T x,$$
$$t = \gamma s,$$
$$y = x - tv.$$

W tym obliczeniu należy wykonać $2n + 1$ operacji (mnożeń z dodawaniami), podczas gdy mnożenie wektora x przez macierz H ma koszt n^2 operacji; ponadto metoda z macierzą H reprezentowaną jawnie jest znacznie gorsza ze względu na skutki błędów zaokrągleń. Zauważmy jeszcze jedną własność macierzy odbicia: jeśli k-ta współrzędna wektora v jest równa 0, to k-ty wiersz i k-ta kolumna macierzy H są takie, jak w macierzy jednostkowej. Wtedy k-ta współrzędna wektora Hx jest identyczna, jak k-ta współrzędna wektora x. Zatem każda zerowa współrzędna wektora v (jeśli wiemy, które to są) umożliwia zaoszczędzenie dwóch operacji (2opms) w powyższym obliczeniu.

Niech a oznacza pewien wektor w $\mathbb{R}^n$. Niech e oznacza pewien ustalony wektor jednostkowy (tj. $\|e\|_2 = 1$). Odbicie Householdera jest to odbicie H skonstruowane w taki sposób, aby wektor Ha miał kierunek wektora e. Musi być zatem $Ha = \pm\|a\|_2 e$. To zaś oznacza, że wektor normalny hiperpłaszczyzny odbicia, v, musi mieć kierunek wektora $a - \|a\|_2 e$, albo $a + \|a\|_2 e$. Chcąc zmniejszyć skutki błędów zaokrągleń, należy zawsze wybierać dłuższy z tych dwóch wektorów. Uzasadnienie takiego sposobu wybierania jest na rysunku — znoszenie się składników podczas odejmowania wektorów powoduje większe zaburzenie kierunku krótszego wektora normalnego hiperpłaszczyzny odbicia.



Zastosujmy teraz odbicia do przekształcania układu równań liniowych $Ax = b$. W pierwszym kroku odbijemy kolumny $a_1, \ldots, a_n$ macierzy A i wektor prawej strony b tak, aby obraz $H_1 a_1$ pierwszej kolumny miał kierunek wektora e_1. Powstanie układ $H_1 A x = H_1 b$, czyli $A^{(1)} x = b^{(1)}$, którego macierz ma zera w pierwszej kolumnie pod diagonalą. Odrzucając pierwsze równanie, otrzymalibyśmy podukład, w którym nie występuje niewiadoma x_1. Podukład ten możemy dalej przekształcać w podobny sposób.

Wykonane w krokach $1, \ldots, k-1$ przekształcenia wytworzyły (z założenia indukcyjnego) macierz $A^{(k-1)}$, w której możemy wyróżnić bloki: macierz trójkątną górną $A_{11}^{(k-1)}$ o wymiarach $(k-1) \times (k-1)$, blok $A_{12}^{(k-1)}$, blok zerowy $A_{21}^{(k-1)}$ i macierz kwadratową $A_{22}^{(k-1)}$ o wymiarach $(n+1-k) \times (n+1-k)$ (zobacz schemat na następnej stronie). W k-tym kroku zajmiemy się tą macierzą i blokiem $b_2^{(k-1)} \in \mathbb{R}^{n+1-k}$ przekształcanego wektora prawej strony. Aby je przekształcić, konstruujemy wektor $v_2^{(k)} \in \mathbb{R}^{n+1-k}$, dany wzorem

$$v_2^{(k)} = a_{2k}^{(k-1)} \mp \|a_{2k}^{(k-1)}\|_2 e_1,$$

w którym $a_{2k}^{(k-1)}$ oznacza pierwszą kolumnę macierzy $A_{22}^{(k-1)}$ (czyli „dolną część” k-tej kolumny macierzy $A^{(k-1)}$). Pierwsza współrzędna wektora e_1 jest jedynką, pozostałe $n-k$ to zera. Aby wektor $v_2^{(k)}$ był jak najdłuższy, wybieramy znak „+” jeśli pierwsza współrzędna wektora $a_{2k}^{(k-1)}$ jest dodatnia, a „−” w przeciwnym razie. Następnie obliczamy liczbę $\gamma_k = 2/\big(v_2^{(k)T} v_2^{(k)}\big)$ i poddajemy kolumny macierzy $A_{22}^{(k-1)}$ i wektor $b_2^{(k-1)}$ odbiciu. Nie ma przy tym potrzeby stosowania ogólnego wzoru do odbijania wektora $a_{2k}^{(k-1)}$, bo skądinąd wiemy, co z tego wyjdzie.

Blok $A_{11}^{(k-1)}$ macierzy $A^{(k-1)}$, który jest macierzą trójkątną górną, zostaje blokiem macierzy $A^{(k)}$. Podobnie blok $b_1^{(k-1)}$ złożony z $k-1$ początkowych współrzędnych wektora $b^{(k-1)}$ zostaje blokiem wektora $b^{(k)}$. Przekształcenie kolumn bloku $A_{22}^{(k-1)}$ i wektora $b_2^{(k-1)}$, reprezentowane przez macierz $\overline{H}_k = I - \gamma_k v_2^{(k)} v_2^{(k)T}$, jest równoważne odbiciu *wszystkich* kolumn macierzy $A^{(k-1)}$ i wektora $b^{(k-1)}$ względem hiperpłaszczyzny w $\mathbb{R}^n$, której wektor normalny $v^{(k)}$ składa się z bloku $v_1^{(k)} = 0 \in \mathbb{R}^{k-1}$ i z bloku $v_2^{(k)}$.

Macierz wykonanego w k-tym kroku odbicia przestrzeni $\mathbb{R}^n$ oznaczymy symbolem H_k. Przekształcenia macierzy $A^{(k-1)}$, wektora $\mathbf{b}^{(k-1)}$ i ich bloków możemy zobaczyć na schemacie:

$$
\begin{array}{c}
A^{(k-1)} \mid \mathbf{b}^{(k-1)} \\
\downarrow \quad \downarrow \\
H_k \rightarrow \; A^{(k)} \mid \mathbf{b}^{(k)}
\end{array}
\qquad
\left[\begin{array}{c|c|c}
A_{11}^{(k-1)} & A_{12}^{(k-1)} & \mathbf{b}_1^{(k-1)} \\
\hline
0 & A_{22}^{(k-1)} & \mathbf{b}_2^{(k-1)}
\end{array}\right]
$$

$$
\left[\begin{array}{c|c}
I & 0 \\
\hline
0 & \overline{H}_k
\end{array}\right]
\left[\begin{array}{c|c|c}
A_{11}^{(k-1)} & A_{12}^{(k-1)} & \mathbf{b}_1^{(k-1)} \\
\hline
0 & \overline{H}_k A_{22}^{(k-1)} & \overline{H}_k \mathbf{b}_2^{(k-1)}
\end{array}\right]
$$

Po wykonaniu $n-1$ odbić skonstruowanych w opisany sposób mamy układ

$$R\mathbf{x} = Q^T\mathbf{b},$$

którego macierz

$$R = A^{(n-1)} = H_{n-1} \dots H_1 A = Q^T A$$

jest trójkątna górna. Zachodzi równość

$$A = QR, \quad \text{gdzie} \quad Q = H_1 \dots H_{n-1},$$

ponieważ macierz Q, będąca iloczynem macierzy ortogonalnych, jest ortogonalna. W ten sposób, za pomocą odbić symetrycznych, znaleźliśmy rozkład ortogonalno-trójkątny macierzy A.

Podobnie, jak w eliminacji Gaussa, przekształcanie prawej strony możemy wykonać później, ale w tym celu trzeba zapamiętać wektory $\mathbf{v}_2^{(k)}$ (i, aby nie obliczać ich ponownie, co kosztuje, liczby γ_k). W tym celu możemy użyć miejsc w tablicy początkowo zawierającej współczynniki macierzy A, ale potrzebujemy dla każdego wektora odbicia dwóch dodatkowych miejsc. Dwa możliwe sposoby[8]

[8]Dodatkowo, prostokątną tablicę można w pamięci komputera przechowywać „wierszami" (domyślny sposób w języku C) lub „kolumnami" (jak w języku FORTRAN). Ma to znaczenie, jeśli korzystamy z gotowych procedur, np. z biblioteki LAPACK.

przechowywania wyników obliczeń są takie:

$$
\left[\begin{array}{ccccc}
r_{11} & r_{12} & r_{13} & \dots & r_{1n} \\
v_2^{(1)} & r_{22} & r_{23} & \dots & r_{2n} \\
v_3^{(1)} & v_3^{(2)} & r_{33} & \dots & r_{3n} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
v_n^{(1)} & v_n^{(2)} & \dots & v_n^{(n-1)} & r_{nn} \\
\hline
v_1^{(1)} & v_2^{(2)} & \dots & v_{n-1}^{(n-1)} & \bullet \\
\hline
\gamma_1 & \gamma_2 & \dots & \gamma_{n-1} & \bullet
\end{array}\right]
\qquad
\left[\begin{array}{cccc|c}
v_1^{(1)} & r_{12} & r_{13} & \dots & r_{1n} \\
v_2^{(1)} & v_2^{(2)} & r_{23} & \dots & r_{2n} \\
\vdots & \vdots & \ddots & \ddots & \vdots \\
v_{n-1}^{(1)} & v_{n-1}^{(2)} & \dots & v_{n-1}^{(n-1)} & r_{n-1,n} \\
v_n^{(1)} & v_n^{(2)} & \dots & v_n^{(n-1)} & \bullet \\
\hline
r_{11} & r_{22} & \dots & r_{n-1,n-1} & r_{nn} \\
\hline
\gamma_1 & \gamma_2 & \dots & \gamma_{n-1} & \bullet
\end{array}\right]
$$

Symbole r_{ij} oznaczają tu współczynniki macierzy R, zaś $v_i^{(k)}$ oznaczają współrzędne wektora $\boldsymbol{v}^{(k)}$. Jest też możliwe zmieszczenie wyników obliczenia z wykorzystaniem tylko jednej dodatkowej zmiennej dla każdej kolumny, po przeskalowaniu wektorów $\boldsymbol{v}^{(k)}$. Jak poprzednio, wykonanie obliczeń *in situ* oznacza „zepsucie" tablicy współczynników macierzy A, zatem najlepiej, aby takiemu „zepsuciu" poddać kopię.

Zwróćmy uwagę, że po rozłożeniu macierzy metodą eliminacji Gaussa na czynniki trójkątne, aby rozwiązać układ $A\boldsymbol{x} = \boldsymbol{b}$ rozwiazujemy numerycznie dwa podzadania, tj. układy z macierzami trójkątnymi. Dla każdego p iloczyn wskaźników uwarunkowania tych podzadań, $\operatorname{cond}_p L$ i $\operatorname{cond}_p U$, jest *zawsze* większy lub równy wskaźnikowi uwarunkowania całego zadania, $\operatorname{cond}_p A$. Ponadto dla dowolnych permutacji reprezentowanych przez macierze P i Q mamy $\operatorname{cond}_p A = \operatorname{cond}_p PAQ^T$. Wybór elementu głównego w metodzie eliminacji Gaussa można interpretować jak dążenie do tego, aby iloczyn wskaźników uwarunkowania czynników rozkładu macierzy PA (lub PAQ^T) był możliwie mały.[9]

Koszt wyznaczenia rozkładu QR macierzy $n \times n$ metodą odbić Householdera jest równy $\bigl(\frac{2}{3}n^3 + O(n^2)\bigr)$opms, jest zatem w przybliżeniu dwukrotnie większy niż koszt eliminacji Gaussa. Z drugiej strony, odpadają koszty wybierania elementu głównego, zresztą decydujący wpływ na czas obliczeń ma efektywność wykorzystania pamięci podręcznej (*cache'a*) procesora przez implementację algorytmu. Dlatego nie można powiedzieć z góry, że eliminacja Gaussa działa dwukrotnie szybciej. Natomiast użycie izometrii (tj. przekształceń reprezentowanych przez macierze ortogonalne) daje bardzo dobre własności numeryczne algorytmu. Zauważmy, że oryginalne zadanie zastępujemy dwoma

[9]Istnieją jednak takie macierze nieosobliwe, że iloczyn wskaźników uwarunkowania ich czynników trójkątnych otrzymanych z zastosowaniem wyboru elementu głównego jest *znacznie większy* niż czynników otrzymanych bez wyboru. Dla dokładności rozwiązania istotna jest *postać* zaburzeń będących skutkami błędów zaokrągleń, której wyprowadzenie wzoru na wskaźnik uwarunkowania nie uwzględnia. Dlatego wybór elementu głównego należy stosować.

podzadaniami — układami równań $Qy = b$ i $Rx = y$. Wskaźnik uwarunkowania w normie drugiej macierzy ortogonalnej Q jest równy 1 (bo $\|Q\|_2 = \|Q^{-1}\|_2 = 1$), zaś $\operatorname{cond}_2 R = \operatorname{cond}_2 A$.

Metoda Choleskiego

W wielu zastosowaniach należy rozwiązać układ $Ax = b$, którego macierz jest symetryczna i dodatnio określona. Dla takich macierzy oczywiście można stosować eliminację Gaussa. Jednak symetria macierzy to okazja do zmniejszenia kosztu obliczeń o połowę. Takich okazji nie wypada marnować.

Przypuśćmy więc, że macierz A jest symetryczna i dodatnio określona. Rozważmy eliminację Gaussa dla tej macierzy. Wszystkie współczynniki na jej diagonali, w tym pierwszy, są dodatnie, zatem pierwszy krok eliminacji jest wykonalny bez przestawiania wierszy. Po zrobieniu tego kroku mamy macierze L_1 i $A^{(1)}$, takie że $A = L_1A^{(1)}$. Macierz L_1 ma w pierwszym wierszu jedynkę i $n-1$ zer. Macierz $A^{(1)}$ ma pierwszy wiersz taki jak A i ma zera w pierwszej kolumnie pod diagonalą. Oznaczmy $S^{(1)} = A^{(1)}L_1^{-T} = L_1^{-1}AL_1^{-T}$. Macierz $S^{(1)}$ jest symetryczna; jej wiersze od drugiego do ostatniego są takie same, jak wiersze macierzy $A^{(1)}$, a współczynnik $s_{11}^{(1)}$ jest równy a_{11}. Zauważamy, że macierz $S^{(1)}$ jest też dodatnio określona.

$$
\underbrace{\begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix}}_{A^{(0)} = A}
\to
\underbrace{\begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \end{bmatrix}}_{A^{(1)} = L_1^{-1}A}
\to
\underbrace{\begin{bmatrix} \bullet & & & & \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \end{bmatrix}}_{S^{(1)} = L_1^{-1}AL_1^{-T}}
$$

Dalej możemy rekurencyjnie, dla $k = 2, \ldots, n-1$, określić macierze

$$S^{(k)} = L_k^{-1}S^{(k-1)}L_k^{-T} = L_k^{-1} \ldots L_1^{-1}AL_1^{-T} \ldots L_k^{-T}.$$

Dla każdego k macierz $S^{(k)}$ jest symetryczna i dodatnio określona (ma więc dodatnie współczynniki na diagonali) i jej wiersze o numerach $k+1, \ldots, n$ są takie same, jak wiersze macierzy $A^{(k)}$ — dlatego kolejne kroki eliminacji Gaussa (obliczanie $A^{(k+1)}$ na podstawie $A^{(k)}$) są wykonalne bez przestawiania wierszy. Stąd wynika, że macierz $D = S^{(n-1)}$ jest diagonalna, jej współczynniki na diagonali są dodatnie i macierz A jest następującym iloczynem:

$$A = L'DL'^T,$$

gdzie $L' = L_1 \ldots L_{n-1}$ (macierz trójkątną dolną otrzymaną przez eliminację Gaussa oznaczam L', bo symbolu L za chwilę użyję do czegoś innego).

Niech M oznacza macierz diagonalną, której współczynniki są pierwiastkami kwadratowymi z odpowiednich współczynników macierzy D. Wtedy $D = M^2 = MM^T$. Macierz $L = L'M$ jest trójkątna dolna; jej kolumny są iloczynami odpowiednich kolumn macierzy L' i współczynników diagonalnych macierzy M i ma miejsce równość $A = L'MM^TL'^T = LL^T$. Wykazaliśmy, że jeśli macierz A jest symetryczna i dodatnio określona, to istnieje macierz trójkątna dolna L, taka że $A = LL^T$.

Znając macierz L, możemy rozwiązać układ równań liniowych z macierzą A (w tym celu kolejno rozwiązujemy układy $L\boldsymbol{y} = \boldsymbol{b}$ i $L^T\boldsymbol{x} = \boldsymbol{y}$). Zobaczmy zatem, jak ją znaleźć. Odpowiedni algorytm ma nazwę metody Choleskiego (w polskiej literaturze bywa też nazywany metodą Choleskiego-Banachiewicza).

Współczynniki macierzy L można obliczyć, traktując równość $LL^T = A$ jak układ równań, którego rozwiązanie można obliczyć „po kolei"; ponieważ macierz A jest symetryczna, mamy $\frac{1}{2}n(n+1)$ danych niezależnych współczynników. Tyle samo współczynników ma na diagonali i pod nią poszukiwana macierz L. Zatem, dla i, j takich że $1 \leqslant j \leqslant i \leqslant n$ mamy równania

$$a_{ij} = \sum_{k=1}^{n} l_{ik}l_{jk} = \sum_{k=1}^{j} l_{ik}l_{jk} = \sum_{k=1}^{j-1} l_{ik}l_{jk} + l_{ij}l_{jj}.$$

Wyodrębniony składnik sumy powyżej umożliwia obliczenie współczynnika l_{ij}, jeśli znamy wszystkie pozostałe współczynniki macierzy L występujące w sumowanych iloczynach; można uporządkować równania w takiej kolejności, że to jest możliwe. Mianowicie, możemy obliczać kolejno

$$\left.\begin{aligned} l_{ij} &= \frac{a_{ij} - \sum_{k=1}^{j-1} l_{ik}l_{jk}}{l_{jj}} \quad \text{dla } j = 1, \ldots, i-1, \\ l_{ii} &= \sqrt{a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2}, \end{aligned}\right\} \quad \text{dla } i = 1, \ldots, n.$$

Obliczenie można wykonać *in situ*, wpisując liczby l_{ij} natychmiast po obliczeniu w miejsce a_{ij} (a zatem, „psując" daną macierz A lub jej kopię). Trzeba podkreślić, że macierz A musi być nie tylko symetryczna, ale także dodatnio określona, aby powyższy algorytm był wykonalny (tj. aby istniała nieosobliwa macierz trójkątna dolna L, taka że $A = LL^T$). To jest warunek konieczny i dostateczny, aby wyrażenia, z których należy obliczać pierwiastki kwadratowe, miały dodatnie wartości.

Zauważmy, że jeśli początkowych k współczynników w i-tym wierszu (dla $k < i$) to zera, to również macierz L ma na początku i-tego wiersza k zerowych współczynników. Można to wykorzystać do efektywnego wykorzystania miejsca w pamięci, a także do zmniejszenia kosztu znajdowania rozkładu (np. jeśli macierz A jest wstęgowa). Jeśli macierz A jest pełna, to też można przechowywać tylko dolny jej trójkąt w tablicy o długości $\frac{1}{2}n(n+1)$. Dla macierzy pełnej znalezienie rozkładu wymaga wykonania ok. $\frac{1}{6}n^3$ operacji mnożenia z dodawaniem lub dzielenia (opms) i obliczenia n pierwiastków kwadratowych.

Układy i algorytmy blokowe

Wiele układów równań liniowych rozwiązywanych w praktycznych zastosowaniach ma macierze o specjalnych własnościach, które można wykorzystać do zmniejszenia kosztu rozwiązywania. Bardzo często macierz w naturalny sposób dzieli się na wyróżniające się jakoś bloki. W najprostszej sytuacji, niech

$$A = \begin{bmatrix} B & C \\ D & E \end{bmatrix}.$$

Przypuśćmy, że macierz $A \in \mathbb{R}^{n\times n}$ jest nieosobliwa i blok $B \in \mathbb{R}^{k\times k}$ dla pewnego $k \in \{1, \ldots, n-1\}$ też jest nieosobliwy. Podzielmy również prawą stronę i wektor niewiadomy na bloki:

$$\mathbf{b} = \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}, \qquad \boldsymbol{x} = \begin{bmatrix} \boldsymbol{y} \\ \boldsymbol{z} \end{bmatrix}.$$

Układ $A\boldsymbol{x} = \mathbf{b}$ podzieliliśmy w ten sposób na dwa podukłady:

$$\begin{cases} B\boldsymbol{y} + C\boldsymbol{z} = \mathbf{p}, \\ D\boldsymbol{y} + E\boldsymbol{z} = \mathbf{q}. \end{cases}$$

Znając $\boldsymbol{z}$, moglibyśmy rozwiązać pierwszy podukład; jego rozwiązanie wyraża się wzorem

$$\boldsymbol{y} = B^{-1}(\mathbf{p} - C\boldsymbol{z}).$$

Wstawiamy to wyrażenie do drugiego podukładu; mamy

$$DB^{-1}(\mathbf{p} - C\boldsymbol{z}) + E\boldsymbol{z} = \mathbf{q},$$

czyli

$$(E - DB^{-1}C)\boldsymbol{z} = \mathbf{q} - DB^{-1}\mathbf{p}.$$

Macierz $S = E - DB^{-1}C$ nazywa się macierzą Schura; jeśli macierze A i B są nieosobliwe, to również macierz S jest nieosobliwa. Możemy znaleźć rozwiązanie z układu z tą macierzą, a następnie blok y. Jeśli o macierzy B wiemy tylko tyle, że jest nieosobliwa, to należałoby ją rozłożyć na czynniki trójkątne (lub ortogonalny i trójkątny). Powiedzmy, że mamy macierz permutacji P i macierze trójkątne L i U, takie że $PB = LU$. Wtedy należy wykonać następujący algorytm:

1. Korzystając z macierzy P, L, U, rozwiąż (macierzowy) układ równań $BF = C$, a następnie oblicz macierz $S = E - DF$,
2. Rozwiąż układ równań $B\boldsymbol{v} = \boldsymbol{p}$, a następnie oblicz wektor $\boldsymbol{w} = \boldsymbol{q} - D\boldsymbol{v}$,
3. Rozwiąż układ równań $S\boldsymbol{z} = \boldsymbol{w}$; w tym celu należy wyznaczyć i wykorzystać jakiś użyteczny rozkład macierzy S,
4. Oblicz wektor $\boldsymbol{u} = \boldsymbol{p} - C\boldsymbol{z}$, a następnie, korzystając z macierzy P, L, U, rozwiąż układ równań $B\boldsymbol{y} = \boldsymbol{u}$.
   Alternatywnie, oblicz wektor $\boldsymbol{t} = C\boldsymbol{z}$, a następnie rozwiąż układ $B\boldsymbol{s} = \boldsymbol{t}$ i oblicz $\boldsymbol{y} = \boldsymbol{v} - \boldsymbol{s}$.

Z wyjątkiem ograniczenia możliwości wyboru elementu głównego, nie mamy tu żadnych zmian (w szczególności kosztu) w porównaniu ze zwykłą eliminacją Gaussa (choć pierwsze dwa kroki można wykonać równolegle). Jeśli jednak blok B jest macierzą symetryczną dodatnio określoną, to zamiast eliminacji Gaussa możemy użyć dwukrotnie tańszej metody Choleskiego. Jeśli zaś blok B jest na przykład macierzą diagonalną lub ortogonalną, to niezależnie od tego, jakie są pozostałe bloki macierzy A, układy równań z macierzą B możemy rozwiązywać znacznie mniejszym kosztem. Ponadto, gdyby blok B był macierzą odbicia symetrycznego, reprezentowaną przez wektor normalny hiperpłaszczyzny odbicia (lub iloczynem takich macierzy, reprezentowanych przez odpowiednie wektory), to jawne wyznaczanie współczynników macierzy B, po to by następnie rozwiązać układ równań z tą macierzą, byłoby przejawem *skrajnego niedołęstwa*. Dlatego powtarzam stały apel: *najpierw* należy się dowiedzieć jak najwięcej o zadaniu, a *potem* dobierać algorytm jego rozwiązywania.

Szacowanie błędu i poprawianie rozwiązania

Oznaczmy symbolem $\boldsymbol{\alpha}$ *dokładne* rozwiązanie układu równań $A\mathbf{x} = \mathbf{b}$ (czyli $\boldsymbol{\alpha} = A^{-1}\mathbf{b}$), i niech symbol $\tilde{\mathbf{x}}$ oznacza wynik *numerycznego* rozwiązywania tego układu (jakimś algorytmem z błędami zaokrągleń). Residuum rozwiązania $\tilde{\mathbf{x}}$, tj. wektor $\mathbf{r} = \mathbf{b} - A\tilde{\mathbf{x}}$, jest równe 0 wtedy i tylko wtedy, gdy $\tilde{\mathbf{x}} = \boldsymbol{\alpha}$. Możemy napisać

$$\boldsymbol{\alpha} - \tilde{\mathbf{x}} = A^{-1}\mathbf{b} - A^{-1}A\tilde{\mathbf{x}} = A^{-1}\mathbf{r}.$$

Z równości $\boldsymbol{\alpha} - \tilde{\mathbf{x}} = A^{-1}\mathbf{r}$ oraz $A(\boldsymbol{\alpha} - \tilde{\mathbf{x}}) = \mathbf{r}$ wynikają nierówności

$$\frac{1}{\|A\|_p}\|\mathbf{r}\|_p \leqslant \|\boldsymbol{\alpha} - \tilde{\mathbf{x}}\|_p \leqslant \|A^{-1}\|_p\|\mathbf{r}\|_p.$$

Możemy użyć tych nierówności do oszacowania wielkości błędu rozwiązania, pod warunkiem, że

- umiemy oszacować normę macierzy A^{-1} (dla $p = 1$ lub $p = \infty$ obliczenie $\|A\|_p$ jest łatwe, ale nie chcemy jawnie wyznaczać macierzy A^{-1}),
- umiemy obliczyć wektor $\mathbf{r}$.

Jeśli do rozwiązania układu użyliśmy metody eliminacji Gaussa, to mamy znalezione trójkątne czynniki L, U rozkładu macierzy A (lub PA albo PAQ^T, zależnie od użytego wariantu wyboru elementu głównego). Istnieje algorytm, który na podstawie tych czynników znajduje, kosztem $O(n^2)$ działań, normę macierzy A^{-1} z dokładnością rzędu 50%, co w zastosowaniu do sprawdzania dokładności otrzymanego rozwiązania wystarczy. Znacznie gorzej wygląda kwestia obliczenia residuum — robiąc to przy użyciu arytmetyki zmiennopozycyjnej, dostaniemy inny wektor, $\tilde{\mathbf{r}}$, przy czym w tym obliczeniu występuje silne znoszenie się składników, wskutek czego znaleziona potem liczba $\|\tilde{\mathbf{r}}\|_p$ może mieć bardzo niewiele wspólnego z $\|\mathbf{r}\|_p$.

Z tego powodu podczas obliczania residuum trzeba zadbać o dokładność. Najprostszym sposobem jest użycie *silniejszej arytmetyki*, jeśli na przykład domyślnie używamy pojedynczej precyzji, to residuum powinniśmy obliczyć w precyzji podwójnej. Jeśli do rozwiązania układu użyliśmy precyzji podwójnej, to można sięgnąć po precyzję rozszerzoną, lub użyć algorytmu Kahana. Działania w wyższej precyzji mogą zajmować więcej czasu, ale całe to obliczenie ma złożoność $\Theta(n^2)$, co jest mało istotne w porównaniu ze złożonością eliminacji Gaussa, rzędu n^3.

Jeśli otrzymane oszacowanie błędu jest za duże, to rozwiązanie można poprawić (w praktyce rzadko się to robi, ale *każdy powinien wiedzieć*, jak to zrobić, jeśli pojawi się taka konieczność). W tym celu wystarczy rozwiązać układ równań

$$A\boldsymbol{\delta} = \tilde{\mathbf{r}},$$

a następnie obliczyć poprawione rozwiązanie

$$\hat{\boldsymbol{x}} = \tilde{\boldsymbol{x}} + \boldsymbol{\delta}.$$

Koszt tego postępowania jest rzędu n^2, ponieważ do rozwiązania układu równań z wektorem prawej strony $\tilde{\mathbf{r}}$ wykorzystujemy trójkątne czynniki rozkładu znalezione wcześniej (jeśli do rozwiązania układu używamy innego rozkładu macierzy A, np. ortogonalno-trójkątnego wyznaczonego metodą odbić Householdera, to też otrzymamy koszt poprawiania rzędu n^2).

Rozwinięciem tego postępowania jest iteracyjne poprawianie rozwiązania, w którym po obliczeniu nowego przybliżenia, $\hat{\boldsymbol{x}}$, obliczamy jego residuum i w razie potrzeby poprawiamy je dalej. Iteracje przerywamy, jeśli norma residuum jest dostatecznie mała, lub jeśli nie jest jest istotnie mniejsza od normy residuum poprzedniego przybliżenia — to oznacza osiągnięcie maksymalnej granicznej dokładności. Kluczowym dla dokładności elementem tego postępowania jest dokładność obliczania wektorów residuum.

Metody iteracyjne

Jeśli liczba n równań i niewiadomych jest wielka, to koszt metod bezpośrednich rozwiązywania takich układów jest zbyt duży. Często w zastosowaniach pojawiają się układy z n rzędu tysięcy lub milionów. Bardzo często macierze układów w takich zastosowaniach są rzadkie, np. mają tylko $O(n)$ niezerowych współczynników, ale rozmieszczenie tych współczynników uniemożliwia stosowanie metod bezpośrednich (np. znalezienie metodą różnic skończonych rozwiązania przybliżonego równania różniczkowego Poissona w kwadracie sprowadza się do rozwiązania układu równań liniowych, którego macierz jest wstęgowa, przy czym wstęga ma szerokość rzędu $\sqrt{n}$, a w każdym wierszu jest co najwyżej 5 współczynników niezerowych).

Do rozwiązywania wielkich układów równań liniowych stosuje się metody iteracyjne. Mając początkowe przybliżenie $\boldsymbol{x}_0$ rozwiązania $\boldsymbol{\alpha}$, metoda konstruuje elementy ciągu, $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots$, zbieżnego do $\boldsymbol{\alpha}$. Obliczenia przerywa się na podstawie kryteriów stopu podobnych, jak dla równań nieliniowych, tj. ustalonego limitu liczby iteracji, kryterium residualnego (residuum jest w tym przypadku wektor $\mathbf{b} - A\boldsymbol{x}$) lub kryterium przyrostowego (opartego na badaniu wartości wyrażenia

$\|x_{k+1} - x_k\|$). Podstawową operacją, wykonywaną w każdej iteracji, jest mnożenie pewnego wektora przez macierz układu A, lub inną macierz, skonstruowaną na podstawie A. Dzięki takiemu ograniczeniu można korzystać z bardzo oszczędnych reprezentacji macierzy, które są w istocie wykazami (tablicami lub listami) miejsc, w których są niezerowe współczynniki. Koszt mnożenia wektora przez macierz jest wtedy proporcjonalny do liczby tych współczynników.

Metody iteracji prostej

Metody iteracji prostej polegają na tym, że na podstawie macierzy A i wektora prawej strony $\mathbf{b}$ konstruuje się pewną macierz B i wektor $\mathbf{t}$, przyjmuje początkowe przybliżenie rozwiązania, x_0, i w kolejnych iteracjach oblicza

$$x_{k+1} = Bx_k + \mathbf{t}.$$

Macierz B i wektor $\mathbf{t}$ muszą być tak dobrane, aby zachodziła równość $\boldsymbol{\alpha} = B\boldsymbol{\alpha} + \mathbf{t}$, tj. aby rozwiązanie było punktem stałym funkcji $\boldsymbol{\varphi}(x) = Bx + \mathbf{t}$. Ponadto, aby ciąg wektorów x_k był zbieżny do $\boldsymbol{\alpha}$ dla dowolnego punktu startowego x_0, funkcja $\boldsymbol{\varphi}$ musi być przekształceniem zwężającym, a zatem pewna (dowolna) norma indukowana macierzy B musi być mniejsza od 1. Okazuje się, że warunkiem koniecznym i dostatecznym istnienia takiej normy, która dla macierzy B przyjmuje wartość mniejszą od 1, jest nierówność $\rho(B) < 1$, gdzie $\rho(B)$ jest to promień spektralny, tj. największa wartość bezwzględna wartości własnej[10] macierzy B. Aby zbieżność była szybka, $\rho(B)$ musi być jak najmniejsze, ale możliwość skonstruowania macierzy B o małym promieniu spektralnym zależy od macierzy A (i w szczególności od jej wskaźnika uwarunkowania).

Metoda Jacobiego: macierz A przedstawiamy w postaci sumy $A = L + D + U$, gdzie macierz L powstaje z A przez zastąpienie zerami współczynników na i nad diagonalą, macierz D jest diagonalna, a macierz U ma zerowe współczynniki na i pod diagonalą. Układ $Ax = \mathbf{b}$ przepisujemy w postaci

$$(L + D + U)x = \mathbf{b}, \quad \text{czyli} \quad Dx = \mathbf{b} - (L + U)x,$$

skąd otrzymujemy metodę:

$$x_{k+1} = D^{-1}\big(\mathbf{b} - (L + U)x_k\big).$$

[10] Wartość własna macierzy kwadratowej B jest to liczba λ, taka że $Bx = \lambda x$ dla pewnego niezerowego wektora x, który jest nazywany wektorem własnym macierzy B. Każda liczbowa macierz $n \times n$ ma co najmniej jedną i co najwyżej n różnych wartości własnych, które są liczbami rzeczywistymi lub zespolonymi. Więcej informacji o algebraicznym zagadnieniu własnym jest w rozdziale szóstym skryptu.

W metodzie Jacobiego mamy zatem $B_J = -D^{-1}(L + U)$ oraz $\mathbf{t}_J = D^{-1}\mathbf{b}$.

W obliczeniach nie wyznaczamy macierzy D^{-1}, choć to jest łatwe; zamiast tego rozwiązujemy układ równań z macierzą diagonalną D (podobnie postępujemy w innych metodach). Warunek $\rho(B_J) < 1$ jest spełniony dla wielu macierzy A; łatwym do sprawdzenia przypadkiem jest macierz diagonalnie dominująca, tj. taka że $|a_{ii}| > \sum_{j\neq i} |a_{ij}|$ dla każdego i.

Metoda Gaussa-Seidela: jak poprzednio, bierzemy $A = L + D + U$, po czym piszemy układ

$$(L + D)\mathbf{x} = \mathbf{b} - U\mathbf{x},$$

skąd otrzymujemy układ równań do rozwiązania w każdej iteracji:

$$(L + D)\mathbf{x}_{k+1} = \mathbf{b} - U\mathbf{x}_k.$$

W metodzie Gaussa-Seidela mamy zatem $B_{GS} = -(L + D)^{-1}U$ i $\mathbf{t}_{GS} = (L + D)^{-1}\mathbf{b}$. Podobnie jak w metodzie Jacobiego, jeśli macierz A jest diagonalnie dominująca, to ciąg $(\mathbf{x}_k)_{k\in\mathbb{N}}$ jest zbieżny do $\boldsymbol{\alpha}$ dla każdego $\mathbf{x}_0 \in \mathbb{R}^n$, przy czym zbieżność metody Gaussa-Seidela jest zwykle szybsza.

Metoda Richardsona: Wprowadzamy parametr τ i piszemy układ równoważny układowi $A\mathbf{x} = \mathbf{b}$:

$$x + \tau A\mathbf{x} = \mathbf{x} + \tau\mathbf{b},$$

skąd mamy

$$\mathbf{x} = (I - \tau A)\mathbf{x} + \tau\mathbf{b} = \mathbf{x} + \tau(\mathbf{b} - A\mathbf{x}).$$

Metoda Richardsona polega na obliczaniu wektorów

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \tau(\mathbf{b} - A\mathbf{x}_k).$$

W tej metodzie mamy $B_R = I - \tau A$, $\mathbf{t}_R = \tau\mathbf{b}$. Podstawowym problemem jest dobranie parametru τ tak, aby ciąg $(\mathbf{x}_k)_k$ był zbieżny i aby ta zbieżność była jak najszybsza. Jeszcze do tego wrócimy.

Metoda sprzężonych gradientów

Metoda sprzężonych gradientów (ang. *conjugate gradient method*, w skrócie *CG*) służy do rozwiązywania układu n równań liniowych $A\mathbf{x} = \mathbf{b}$ z macierzą A

symetryczną i dodatnio określoną. Co ciekawe, z punktu widzenia algebry metoda ta jest metodą bezpośrednią, mianowicie startując z dowolnego punktu $\boldsymbol{x}_0$, wytwarza skończony ciąg $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ dla pewnego $m \leqslant n$; gdyby nie było błędów zaokrągleń, to ostatni element tego ciągu byłby rozwiązaniem $\boldsymbol{\alpha}$. Błędy zaokrągleń dla układów z macierzami wielkimi i źle uwarunkowanymi (jakie często występują w zastosowaniach) likwidują tę własność — wytworzony ciąg punktów początkowo zbliża się do rozwiązania, a następnie oddala. Dlatego zwykle metodę CG wykorzystuje się jak metodę iteracyjną, która wytwarza dostatecznie dokładne przybliżenie $\boldsymbol{x}_k$ rozwiązania dla pewnego k znacznie mniejszego niż n.

Podstawowym krokiem metody CG jest minimalizacja wielomianu kwadratowego wzdłuż pewnej prostej. Funkcja określona wzorem

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{x}^T\boldsymbol{b}$$

jest wielomianem drugiego stopnia zmiennych $x_1, \ldots, x_n$. Ponieważ macierz A jest dodatnio określona, funkcja f ma minimum; jej gradient jest równy

$$\nabla f(\boldsymbol{x}) = A\boldsymbol{x} - \boldsymbol{b},$$

a więc wektor residuum, $\boldsymbol{r} = \boldsymbol{b} - A\boldsymbol{x} = -\nabla f(\boldsymbol{x})$, określa kierunek najszybszego spadku funkcji f w punkcie $\boldsymbol{x}$. Poszukiwane rozwiązanie $\boldsymbol{\alpha}$ jest właśnie tym punktem przestrzeni $\mathbb{R}^n$, w którym funkcja f przyjmuje wartość minimalną. Niech $\boldsymbol{x}_k$ oznacza bieżące przybliżenie rozwiązania, zaś $\boldsymbol{v}_k$ oznacza pewien niezerowy wektor. Zbiór $L = \{\boldsymbol{x} = \boldsymbol{x}_k + t\boldsymbol{v}_k \colon t \in \mathbb{R}\}$ jest prostą w $\mathbb{R}^n$. Podstawiając jako argument funkcji f wyrażenie $\boldsymbol{x}_k + t\boldsymbol{v}_k$, otrzymujemy wielomian jednej zmiennej,

$$g(t) = f(\boldsymbol{x}_k + t\boldsymbol{v}_k) = \frac{1}{2}\big(\boldsymbol{v}_k^T A\boldsymbol{v}_k t^2 - 2\boldsymbol{v}_k^T(\boldsymbol{b} - A\boldsymbol{x}_k)t + \boldsymbol{x}_k^T(A\boldsymbol{x}_k - 2\boldsymbol{b})\big).$$

Trójmian kwadratowy $g(t) = at^2 - 2bt + c$ ze współczynnikiem $a > 0$ przyjmuje wartość minimalną dla $t = b/a$. Zatem, obliczając

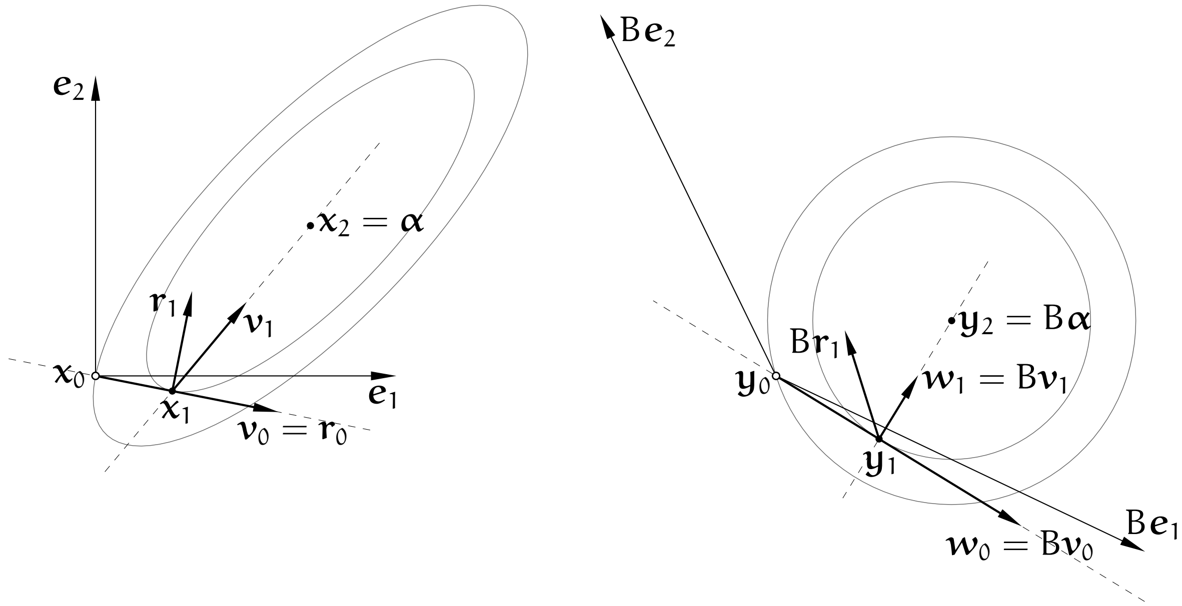
$$\begin{aligned} \boldsymbol{r}_k &= \boldsymbol{b} - A\boldsymbol{x}_k, \\ t_k &= \frac{\boldsymbol{v}_k^T\boldsymbol{r}_k}{\boldsymbol{v}_k^T A\boldsymbol{v}_k}, \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_k + t_k\boldsymbol{v}_k, \end{aligned}$$

otrzymamy punkt $\boldsymbol{x}_{k+1}$ prostej L, w którym wartość funkcji f jest najmniejsza. Możemy zauważyć (lub przynajmniej sprawdzić), że następny wektor residuum, $\boldsymbol{r}_{k+1} = \boldsymbol{b} - A\boldsymbol{x}_{k+1} = \boldsymbol{r}_k - t_k A\boldsymbol{v}_k$, spełnia warunek $\boldsymbol{v}_k^T\boldsymbol{r}_{k+1} = 0$.

Istnieje rodzina metod iteracyjnych, w których powyższe obliczenie łączy się z pewną strategią wyboru wektorów $\boldsymbol{v}_k$, wyznaczających kierunki odpowiednich prostych w kolejnych iteracjach; oczywiście, istotne są tylko *kierunki* tych wektorów. W metodzie CG, która należy do tej rodziny, wektory $\boldsymbol{v}_k$ mają kierunki sprzężone względem macierzy A, przez co rozumiemy, że spełniają równości $\boldsymbol{v}_i^T A\boldsymbol{v}_k = 0$ dla $i \neq k$. Przyjrzyjmy się tej własności i temu, co z niej wynika. Dla dowolnej symetrycznej i dodatnio określonej macierzy A istnieje (jednoznacznie określona) macierz B symetryczna i dodatnio określona[11], taka że $A = B^2$. Jeśli oznaczymy $\boldsymbol{w}_i = B\boldsymbol{v}_i$ dla każdego i, to dla $i \neq k$ mamy $\langle \boldsymbol{w}_k, \boldsymbol{w}_i\rangle = \boldsymbol{w}_i^T\boldsymbol{w}_k = 0$. Po podstawieniu $\boldsymbol{x} = B^{-1}\boldsymbol{y}$ jako argumentu funkcji f, otrzymujemy wielomian

$$h(\boldsymbol{y}) = f(B^{-1}\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^T B^{-T} A B^{-1}\boldsymbol{y} - \boldsymbol{y}^T B^{-T}\boldsymbol{b} = \frac{1}{2}\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T B^{-T}\boldsymbol{b},$$

którego część kwadratowa jest równa $\frac{1}{2}(y_1^2 + \cdots + y_n^2)$. Poszukiwanie minimum funkcji $f(\boldsymbol{x})$ wzdłuż prostych o kierunkach $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_k$, zaczynając od punktu $\boldsymbol{x}_0$, jest równoważne minimalizacji funkcji $h(\boldsymbol{y})$ wzdłuż prostych o wzajemnie prostopadłych kierunkach $\boldsymbol{w}_0, \ldots, \boldsymbol{w}_k$, zaczynając od punktu $\boldsymbol{y}_0 = B\boldsymbol{x}_0$.



Obliczenie dla pewnego układu dwóch równań przedstawia rysunek. Wektor $\boldsymbol{v}_0 = \boldsymbol{r}_0$ jest prostopadły do przechodzącej przez $\boldsymbol{x}_0$ warstwicy funkcji f, która jest elipsą o środku $\boldsymbol{\alpha}$. Punkt $\boldsymbol{x}_1$ leży na kolejnej warstwicy; residuum w tym punkcie, $\boldsymbol{r}_1$, ma kierunek najszybszego spadku funkcji f, prostopadły do $\boldsymbol{v}_0$. Obrazy $\boldsymbol{w}_0$ i $\boldsymbol{w}_1$ wektorów $\boldsymbol{v}_0$ i $\boldsymbol{v}_1$ w przekształceniu liniowym określonym przez macierz B są prostopadłe do siebie; obrazami warstwic funkcji f, tj. warstwicami funkcji h, są

[11] Istnieje macierz ortogonalna X i macierz diagonalna Λ z dodatnimi współczynnikami $\lambda_1, \ldots, \lambda_n$ na diagonali, takie że $A = X\Lambda X^T$; liczby $\lambda_1, \ldots, \lambda_n$ są wartościami własnymi macierzy A, a jej wektorami własnymi są kolumny macierzy X. Macierz B jest równa XMX^T, gdzie M jest macierzą diagonalną, z liczbami $\sqrt{\lambda_1}, \ldots, \sqrt{\lambda_n}$ na diagonali.

okręgi. Punkt $\boldsymbol{y}_1$ i wektor $\boldsymbol{w}_1 = B\boldsymbol{v}_1$ wyznaczają prostą — symetralną cięciwy warstwicy, na której leży punkt $\boldsymbol{y}_0$ (i cięciw wielu innych warstwic). Środek okręgów, tj. punkt $B\boldsymbol{\alpha}$, w którym funkcja h ma minimalną wartość w $\mathbb{R}^2$, leży na tej prostej, zatem wektor $\boldsymbol{v}_1 = B^{-1}\boldsymbol{w}_1$ wyznacza kierunek prostej łączącej punkt $\boldsymbol{x}_1$ i rozwiązanie $\boldsymbol{\alpha}$.

W ogólności do znalezienia w $\mathbb{R}^n$ minimum funkcji kwadratowej h, której warstwice są sferami, wystarczy wykonać co najwyżej n kroków minimalizacji wzdłuż prostych wzajemnie prostopadłych.

W metodzie CG przyjmujemy $\boldsymbol{v}_0 = \boldsymbol{r}_0 = \boldsymbol{b} - A\boldsymbol{x}_0$ oraz

$$\boldsymbol{v}_{k+1} = \boldsymbol{r}_{k+1} + s_k\boldsymbol{v}_k, \text{ gdzie } s_k = -\frac{\boldsymbol{v}_k^T A\boldsymbol{r}_{k+1}}{\boldsymbol{v}_k^T A\boldsymbol{v}_k}, \text{ dla } k \geqslant 0, \text{ takiego że } \boldsymbol{v}_k \neq \boldsymbol{0}.$$

Twierdzenie. *Otrzymane w ten sposób wektory $\boldsymbol{v}_k$ mają kierunki sprzężone względem macierzy A i metoda znajduje rozwiązanie po co najwyżej n iteracjach.*

Szkic dowodu: Zauważamy, że dla każdego $j \leqslant k+1$ wektor $\boldsymbol{v}_j$ jest kombinacją liniową wektorów $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_j$, a wektor $\boldsymbol{r}_j$ jest kombinacją liniową wektorów $\boldsymbol{v}_0, \ldots, \boldsymbol{v}_j$. Niech K_j oznacza podprzestrzeń przestrzeni $\mathbb{R}^n$ rozpiętą przez te wektory[12]. Przed obliczeniem wektora $\boldsymbol{v}_{k+1}$ były wyznaczone wektory $\boldsymbol{r}_0, \ldots, \boldsymbol{r}_{k+1}$ (oraz $\boldsymbol{v}_0, \ldots \boldsymbol{v}_k$), które określają przestrzenie $K_0 \subset \cdots \subset K_{k+1}$. Przyjmujemy założenie indukcyjne, że dla każdego wektora $\boldsymbol{u} \in K_i$, gdzie $0 \leqslant i < j \leqslant k$, jest $\boldsymbol{u}^T A\boldsymbol{v}_j = \boldsymbol{u}^T\boldsymbol{r}_j = 0$. Równoważnie, dla wszystkich par (i,j) takich że $0 \leqslant i < j \leqslant k$ są spełnione równości $\boldsymbol{v}_i^T A\boldsymbol{v}_j = \boldsymbol{r}_i^T A\boldsymbol{v}_j = \boldsymbol{v}_i^T\boldsymbol{r}_j = \boldsymbol{r}_i^T\boldsymbol{r}_j = 0$.

Wiemy, że residuum w punkcie $\boldsymbol{x}_{k+1}$ spełnia warunek $\boldsymbol{v}_k^T\boldsymbol{r}_{k+1} = 0$. Mamy też

$$\boldsymbol{v}_k^T A\boldsymbol{v}_{k+1} = \boldsymbol{v}_k^T A(\boldsymbol{r}_{k+1} + s_k\boldsymbol{v}_k) = \boldsymbol{v}_k^T A\boldsymbol{r}_{k+1} - \frac{\boldsymbol{v}_k^T A\boldsymbol{r}_{k+1}}{\boldsymbol{v}_k^T A\boldsymbol{v}_k}\boldsymbol{v}_k^T A\boldsymbol{v}_k = 0.$$

Niech $i < k$. Jeśli $\boldsymbol{w} \in K_i$, to

$$\boldsymbol{w}^T\boldsymbol{r}_{k+1} = \boldsymbol{w}^T(\boldsymbol{r}_k - t_k A\boldsymbol{v}_k) = \boldsymbol{w}^T\boldsymbol{r}_k - t_k\boldsymbol{w}^T A\boldsymbol{v}_k = 0.$$

Jeśli zatem $\boldsymbol{u} \in K_k$, to $\boldsymbol{u}^T\boldsymbol{r}_{k+1} = 0$, bo wektor $\boldsymbol{u}$ jest kombinacją liniową wektora $\boldsymbol{v}_k$ i pewnego wektora $\boldsymbol{w} \in K_{k-1}$. Ponieważ $\boldsymbol{v}_{k+1} = \boldsymbol{r}_{k+1} + s_k\boldsymbol{v}_k$ oraz $t_i A\boldsymbol{v}_i = \boldsymbol{r}_i - \boldsymbol{r}_{i+1} \in K_{i+1} \subset K_k$, mamy również

$$\boldsymbol{v}_i^T A\boldsymbol{v}_{k+1} = \boldsymbol{v}_i^T A(\boldsymbol{r}_{k+1} + s_k\boldsymbol{v}_k) = \boldsymbol{v}_i^T A\boldsymbol{r}_{k+1} + s_k\underbrace{\boldsymbol{v}_i^T A\boldsymbol{v}_k}_{=0} = \frac{1}{t_i}(\boldsymbol{r}_i - \boldsymbol{r}_{i+1})^T\boldsymbol{r}_{k+1} = 0.$$

[12] Przestrzenie $K_j = \mathrm{lin}\{\boldsymbol{r}_0, \ldots, \boldsymbol{r}_j\} = \mathrm{lin}\{\boldsymbol{v}_0, \ldots, \boldsymbol{v}_j\}$ są nazywane przestrzeniami Kryłowa.

Z założenia indukcyjnego i wykazanych wyżej równości wynika, że
$\boldsymbol{u}^T A\boldsymbol{v}_{k+1} = \boldsymbol{u}^T\boldsymbol{r}_{k+1} = 0$ dla każdego wektora $\boldsymbol{u} \in K_k$.

Z przeprowadzonego wyżej rachunku wynika, że wektory $\boldsymbol{r}_0, \boldsymbol{r}_1, \ldots$ są do siebie nawzajem prostopadłe. Ale ciąg ortogonalny w $\mathbb{R}^n$ może składać się z co najwyżej n niezerowych wektorów, zatem residuum w którymś punkcie $\boldsymbol{x}_k$, otrzymanym po co najwyżej n iteracjach metody CG, musi być zerowe. □

Dzięki temu, że $\boldsymbol{v}_k^T\boldsymbol{r}_k = \boldsymbol{r}_k^T\boldsymbol{r}_k$ (dlaczego tak jest?), wyrażenia opisujące t_k i s_k można przekształcić tak, aby zmniejszyć koszty ich obliczania:

$$\begin{aligned} t_k &= \frac{\boldsymbol{v}_k^T\boldsymbol{r}_k}{\boldsymbol{v}_k^T A\boldsymbol{v}_k} = \frac{\boldsymbol{r}_k^T\boldsymbol{r}_k}{\boldsymbol{v}_k^T A\boldsymbol{v}_k}, \\ s_k &= -\frac{\boldsymbol{v}_k^T A\boldsymbol{r}_{k+1}}{\boldsymbol{v}_k^T A\boldsymbol{v}_k} = -\frac{1}{t_k}\frac{(\boldsymbol{r}_k - \boldsymbol{r}_{k+1})^T\boldsymbol{r}_{k+1}}{\boldsymbol{v}_k^T A\boldsymbol{v}_k} = \frac{\boldsymbol{v}_k^T A\boldsymbol{v}_k}{\boldsymbol{r}_k^T\boldsymbol{r}_k}\frac{\boldsymbol{r}_{k+1}^T\boldsymbol{r}_{k+1}}{\boldsymbol{v}_k^T A\boldsymbol{v}_k} = \frac{\boldsymbol{r}_{k+1}^T\boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^T\boldsymbol{r}_k}. \end{aligned}$$

Teoretycznie (a dokładniej, na podstawie takiej teorii, która nie przewiduje błędów zaokrągleń, czyli algebry) obliczenia należy zakończyć po otrzymaniu $\boldsymbol{r}_k = \boldsymbol{0}$. W implementacji korzystającej z arytmetyki zmiennopozycyjnej obliczenia są przerywane, gdy wektor $\boldsymbol{v}_k$ lub $\boldsymbol{r}_{k+1}$ ma dostatecznie małą normę drugą.

Metodę CG realizuje następujący podprogram:

```
r = b − Ax;   /* r = r0 */
v = r;        /* v = v0 */
c = rᵀr;
for ( k = 0; k < n; k++ ) {
  if ( vᵀv < δ² ) return;
  z = Av;
  t = c/(vᵀz);  /* t = tk */
  x = x + tv;   /* x = xk+1 */
  r = r − tz;   /* r = rk+1 */
  d = rᵀr;
  if ( d < ε² ) return;
  v = r + (d/c)v;  /* v = vk+1 */
  c = d;
}
```

Tablica $\boldsymbol{x}$ początkowo zawiera współrzędne wektora $\boldsymbol{x}_0$. Obliczone przybliżenie rozwiązania jest końcową zawartością tej tablicy; oprócz niej podprogram używa jeszcze trzech tablic o długości n. Mnożenie wektora przez macierz A może być realizowane przez podprogram podany jako parametr i będący „czarną skrzynką” dla implementacji metody. Parametry δ i ε określają kryteria stopu.

Poprawianie uwarunkowania

Rozważmy metodę Richardsona. Jeśli macierz A jest symetryczna i dodatnio określona, to jej wartości własne są rzeczywiste i dodatnie, zawarte w przedziale $[\lambda_{\min}, \lambda_{\max}]$. Macierz B_R o najmniejszym promieniu spektralnym otrzymamy, przyjmując

$$\tau = \tau_{\mathrm{opt}} = \frac{2}{\lambda_{\min} + \lambda_{\max}}.$$

Często mamy pewne informacje na temat wartości własnych macierzy A, co umożliwia dobranie optymalnego lub prawie optymalnego parametru. Przyjrzyjmy się jednak szybkości zbieżności. Promień spektralny macierzy $B_R = I - \tau_{\mathrm{opt}} A$ (który dla macierzy symetrycznej jest równy jej normie drugiej indukowanej) jest równy

$$\rho(B_R) = 1 - \tau_{\mathrm{opt}}\lambda_{\min} = 1 - \frac{2\lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} = \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}}.$$

Dla macierzy symetrycznej i dodatnio określonej (założyliśmy, że taka jest macierz A) jest $\operatorname{cond}_2 A = \lambda_{\max}/\lambda_{\min}$. Korzystając z tej formuły, otrzymujemy

$$\rho(B_R) = \frac{\operatorname{cond}_2 A - 1}{\operatorname{cond}_2 A + 1}.$$

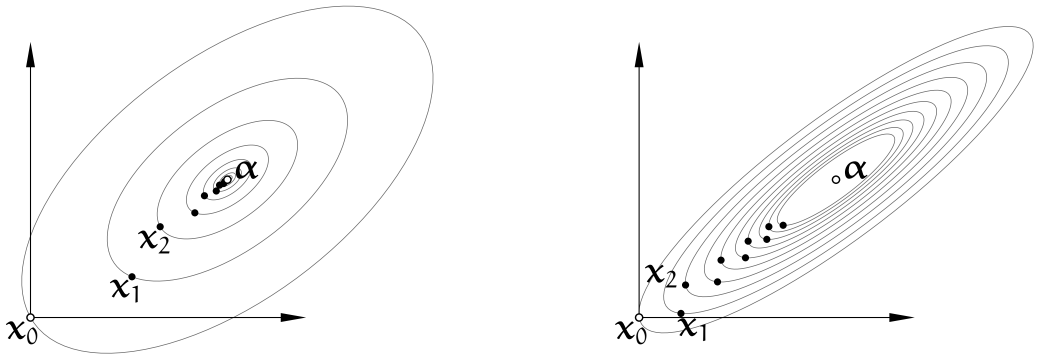
Tak więc, nawet jeśli wybierzemy optymalnie parametr τ, jeśli macierz A ma wielki wskaźnik uwarunkowania, zbieżność konstruowanego przez metodę Richardsona ciągu $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ do rozwiązania jest bardzo wolna. Podobne spostrzeżenie dotyczy także innych metod iteracyjnych. W szczególności dla metody CG błąd $\boldsymbol{\varepsilon}_k = \boldsymbol{x}_k - \boldsymbol{\alpha}$ ma oszacowanie (z macierzą $B = B^T$, taką że $B^2 = A$)

$$\|B\boldsymbol{\varepsilon}_k\|_2 \leqslant 2\left(\frac{\sqrt{\operatorname{cond}_2 A} - 1}{\sqrt{\operatorname{cond}_2 A} + 1}\right)^k \|B\boldsymbol{\varepsilon}_0\|_2.$$

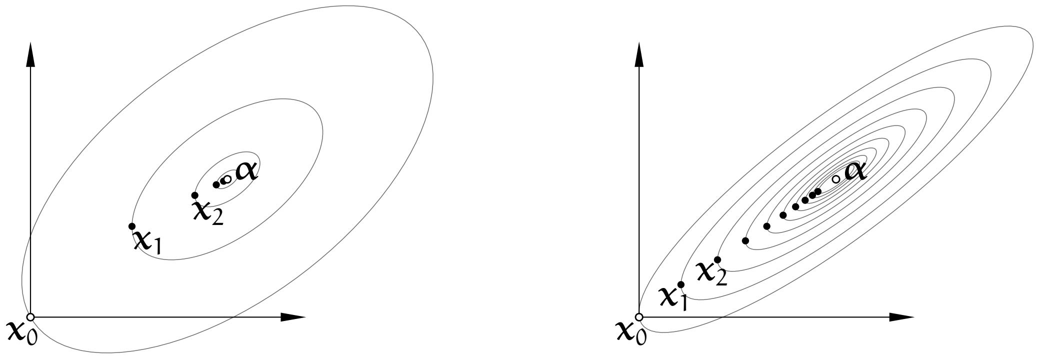
<u>Zbieżność metod iteracyjnych można przyspieszyć</u>, zastępując układ dany układem równoważnym, którego macierz ma mniejszy wskaźnik uwarunkowania. Cel ten można osiągnąć za pomocą macierzy S o następujących własnościach: układy równań z macierzą S są łatwe do rozwiązania, i zachodzi nierówność $\operatorname{cond}(S^{-1}A) \ll \operatorname{cond} A$; aby tak było, macierz S musi w jakiś sposób przybliżać macierz A. Metodę iteracyjną stosujemy do układu $S^{-1}A\boldsymbol{x} = S^{-1}\boldsymbol{b}$, przy czym

<u>nigdy</u> nie wyznaczamy macierzy $S^{-1}A$; zamiast tego, mając obliczyć wektor $\boldsymbol{u} = S^{-1}A\boldsymbol{v}$, obliczamy $\boldsymbol{w} = A\boldsymbol{v}$ i rozwiązujemy układ $S\boldsymbol{u} = \boldsymbol{w}$.

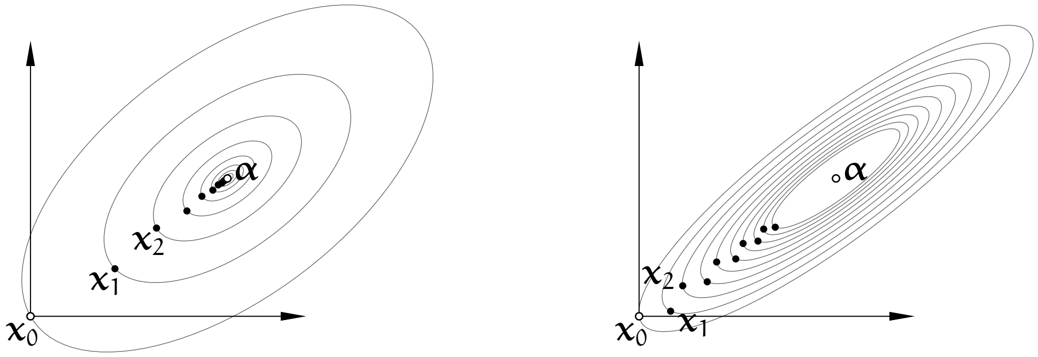
metoda Jacobiego:



metoda Gaussa-Seidela:



metoda Richardsona:



Rysunek przedstawia eksperyment, w którym były rozwiązywane dwa układy równań o tym samym rozwiązaniu, których macierze symetryczne i dodatnio określone mają wskaźniki uwarunkowania odpowiednio 4 i 16. Pokazane są ciągi przybliżeń rozwiązania wygenerowane przez opisane wcześniej metody (dla metody Richardsona przyjęty został optymalny parametr τ) i warstwice wielomianu kwadratowego f rozważanego w opisie metody CG, przechodzące przez kolejne punkty $\boldsymbol{x}_k$.

Nawet jeśli macierze A i S są symetryczne, macierz $S^{-1}A$ na ogół nie jest taka. Niektóre metody, na przykład CG, wymagają, aby macierz układu była symetryczna. To wymaganie można spełnić.

Metoda PCG

Pokażemy metodę sprzężonych gradientów z poprawianiem uwarunkowania (ang. *preconditioned conjugate gradient method*, *PCG*). Niech C oznacza macierz, taką że $\operatorname{cond}(C^{-T}AC^{-1}) \ll \operatorname{cond} A$. Dany układ $A\boldsymbol{x} = \boldsymbol{b}$ z macierzą symetryczną i dodatnio określoną A zastępujemy układem równań

$$C^{-T}AC^{-1}\hat{\boldsymbol{x}} = C^{-T}\boldsymbol{b}. \tag{*}$$

Macierz $C^{-T}AC^{-1}$ jest symetryczna i zachodzi równość $\boldsymbol{x} = C^{-1}\hat{\boldsymbol{x}}$. W obliczeniach chcemy używać oryginalnych danych, tj. macierzy A i wektora $\boldsymbol{b}$, ale chcemy, aby zbieżność otrzymanego ciągu wektorów $\boldsymbol{x}_0, \boldsymbol{x}_1, \ldots$ była określona przez uwarunkowanie układu (*).

Wprowadzimy kilka oznaczeń: symbole $\boldsymbol{x}_k$ i $\boldsymbol{r}_k$ oznaczają kolejne przybliżenie rozwiązania układu $A\boldsymbol{x} = \boldsymbol{b}$ i wektor residuum, zaś $\hat{\boldsymbol{x}}_k = C\boldsymbol{x}_k$ i $\hat{\boldsymbol{r}}_k$ to odpowiednio kolejne przybliżenie rozwiązania i residuum układu (*). Wektor $\boldsymbol{v}_k$ wyznacza kierunek prostej, na której leży punkt $\boldsymbol{x}_k$ i kolejny punkt $\boldsymbol{x}_{k+1}$, zaś symbolem $\hat{\boldsymbol{v}}_k$ oznaczymy wektor kierunkowy prostej łączącej punkty $\hat{\boldsymbol{x}}_k$ i $\hat{\boldsymbol{x}}_{k+1}$; mamy przy tym $\hat{\boldsymbol{v}}_k = C\boldsymbol{v}_k$. Dla każdego k jest

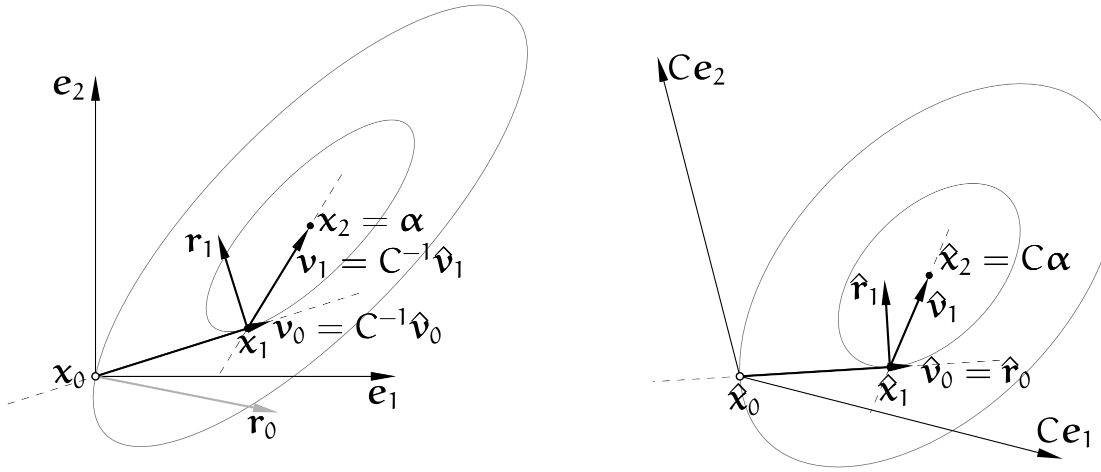
$$\hat{\boldsymbol{r}}_k = C^{-T}\boldsymbol{b} - C^{-T}AC^{-1}\hat{\boldsymbol{x}}_k = C^{-T}(\boldsymbol{b} - A\boldsymbol{x}_k) = C^{-T}\boldsymbol{r}_k, \qquad \text{czyli} \qquad \boldsymbol{r}_k = C^T\hat{\boldsymbol{r}}_k.$$

Oznaczmy $S = C^TC$. W k-tym kroku obliczamy (porównaj to z opisem metody CG)

$$t_k = \frac{\hat{\boldsymbol{r}}_k^T\hat{\boldsymbol{r}}_k}{\hat{\boldsymbol{v}}_k^TC^{-T}AC^{-1}\hat{\boldsymbol{v}}_k} = \frac{(C^{-T}\boldsymbol{r}_k)^TC^{-T}\boldsymbol{r}_k}{\boldsymbol{v}_k^TA\boldsymbol{v}_k} = \frac{\boldsymbol{r}_k^TS^{-1}\boldsymbol{r}_k}{\boldsymbol{v}_k^TA\boldsymbol{v}_k}.$$

Następnie moglibyśmy obliczyć $\hat{\boldsymbol{x}}_{k+1} = \hat{\boldsymbol{x}}_k + t_k\hat{\boldsymbol{v}}_k$ itd., ale zamiast tego obliczamy bezpośrednio

$$\begin{aligned}
\boldsymbol{x}_{k+1} &= C^{-1}\hat{\boldsymbol{x}}_{k+1} = C^{-1}\hat{\boldsymbol{x}}_k + t_kC^{-1}\hat{\boldsymbol{v}}_k = \boldsymbol{x}_k + t_k\boldsymbol{v}_k,\\
\boldsymbol{r}_{k+1} &= C^T\hat{\boldsymbol{r}}_{k+1} = C^T\hat{\boldsymbol{r}}_k - C^Tt_kC^{-T}AC^{-1}\hat{\boldsymbol{v}}_k = \boldsymbol{r}_k - t_kA\boldsymbol{v}_k,\\
s_k &= \frac{\hat{\boldsymbol{r}}_{k+1}^T\hat{\boldsymbol{r}}_{k+1}}{\hat{\boldsymbol{r}}_k^T\hat{\boldsymbol{r}}_k} = \frac{(C^{-T}\boldsymbol{r}_{k+1})^TC^{-T}\boldsymbol{r}_{k+1}}{(C^{-T}\boldsymbol{r}_k)^TC^{-T}\boldsymbol{r}_k} = \frac{\boldsymbol{r}_{k+1}^TS^{-1}\boldsymbol{r}_{k+1}}{\boldsymbol{r}_k^TS^{-1}\boldsymbol{r}_k},\\
\boldsymbol{v}_{k+1} &= C^{-1}\hat{\boldsymbol{v}}_{k+1} = C^{-1}\hat{\boldsymbol{r}}_{k+1} + s_kC^{-1}\hat{\boldsymbol{v}}_k = S^{-1}\boldsymbol{r}_{k+1} + s_k\boldsymbol{v}_k.
\end{aligned}$$



Rysunek przedstawia przykład ilustrujący działanie metody PCG. Macierz C, użyta w tym przykładzie, jest symetryczna, dodatnio określona i taka, że $A = C^4$, dzięki czemu $\mathrm{cond}_2(C^{-T}AC^{-1}) = \sqrt{\mathrm{cond}_2 A}$. Oczywiście, niezależnie od uwarunkowania, dla układu dwóch równań metoda znajduje rozwiązanie $\boldsymbol{\alpha}$ (z dokładnością do błędów zaokrągleń) już po drugiej iteracji. Ale porównując ten rysunek z rysunkiem na s. 4.24 (w obu przypadkach rozwiązywany jest ten sam układ równań) widzimy, że przybliżenie otrzymane już w pierwszej iteracji, $\boldsymbol{x}_1$, leży bliżej rozwiązania niż w przypadku, gdy metodę stosujemy do oryginalnego układu równań $A\boldsymbol{x} = \mathbf{b}$. Mówiąc niezbyt ściśle, im lepiej uwarunkowany jest układ równań, tym lepiej wektory kierunkowe prostych, wzdłuż których minimalizujemy funkcję kwadratową, „celują” w punkt $\boldsymbol{\alpha}$. To ma ogromne znaczenie praktyczne, bo dla *wielkich* układów równań liniowych metodę CG stosujemy jako metodę iteracyjną i chcemy wykonać *dużo mniej* iteracji niż n.

Nigdzie w końcowych wzorach, wyprowadzonych wyżej, nie występuje macierz C. Zamiast niej mamy macierz symetryczną S, a dokładniej S^{-1}. Macierz S powinna przybliżać macierz A, ale układ równań z macierzą S powinien być znacznie łatwiejszy do rozwiązania. W różnych zastosowaniach bywa tak, że nie znamy macierzy S w postaci jawnej (tzn. w postaci tablicy współczynników), ale dysponujemy macierzą S^{-1}, a dokładniej jakimś szybkim algorytmem, który dla dowolnego wektora $\mathbf{r}$ produkuje wektor $S^{-1}\mathbf{r}$. Dwa parametry procedury, która realizuje metodę PCG, powinny być wskaźnikami do procedur, z których pierwsza ma za zadanie obliczenie wektora $A\boldsymbol{v}$, a zadaniem drugiej jest obliczenie wektora $S^{-1}\mathbf{r}$, dla wektorów $\boldsymbol{v}$ i $\mathbf{r}$ przekazywanych jako parametry. Kod podprogramu realizującego algorytm PCG ma postać:

$\mathbf{r} = \mathbf{b} - A\boldsymbol{x}$; /* $\mathbf{r} = \mathbf{r}_0$ */
$\boldsymbol{z} = S^{-1}\mathbf{r}$; /* *rozwiązywanie układu* $S\boldsymbol{z} = \mathbf{r}$ */
$\boldsymbol{v} = \boldsymbol{z}$; /* $\boldsymbol{v} = \boldsymbol{v}_0$ */
$c = \boldsymbol{z}^T\mathbf{r}$;

```
for ( k = 0; k < n; k++ ) {
  if ( vᵀv < δ² ) return;
  z = Av;
  t = c/(vᵀz); /* t = t_k */
  x = x + tv; /* x = x_{k+1} */
  r = r − tz; /* r = r_{k+1} */
  z = S⁻¹r; /* rozwiązywanie układu Sz = r */
  d = zᵀr;
  if ( d < ε² ) {
    if ( rᵀr < ε² ) /* kończymy, jeśli residuum */
      return;        /* oryginalnego układu jest dostatecznie małe */
  }
  v = z + (d/c)v; /* v = v_{k+1} */
  c = d;
}
```

Zastępowanie układu równań układem o macierzy lepiej uwarunkowanej ma angielską nazwę *preconditioning*, a używana do tego macierz S to tzw. *preconditioner*; terminy te nie mają powszechnie przyjętych polskich odpowiedników. Metody znajdowania odpowiednich macierzy są silnie związane ze specyfiką zadania i daleko wykraczają poza ten wykład. Niemniej, warto wiedzieć, że metody iteracyjne z poprawianiem uwarunkowania są w zasadzie jedynymi skutecznymi metodami rozwiązywania *naprawdę wielkich* układów równań liniowych (które w praktyce często są *naprawdę źle* uwarunkowane).

Zadania i problemy

1. Znajdź wyrażenie opisujące wskaźnik uwarunkowania zadania obliczania iloczynu danej macierzy nieosobliwej A i wektora x.

2. Udowodnij, że jeśli macierz kwadratowa nieosobliwa A jest iloczynem macierzy kwadratowych B i C, to (dla wskaźników uwarunkowania określonych za pomocą dowolnej normy indukowanej) zachodzi nierówność $\operatorname{cond} A \leqslant \operatorname{cond} B \operatorname{cond} C$.

3. Rozwiąż układy równań liniowych

$$\begin{bmatrix} 1 & 1 \\ 1 & 0.98 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \begin{bmatrix} 1 & 1 \\ 1 & 0.99 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

   Oblicz wskaźnik uwarunkowania macierzy pierwszego układu w normie pierwszej.

4. Niech

$$A = \begin{bmatrix} 3 & 2 & 1 \\ 30 & 22 & 11 \\ 0 & 20 & 110 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 19 \\ 90 \end{bmatrix}.$$

   Dokonując odpowiedniego częściowego wyboru elementu głównego, znajdź macierze permutacji P, trójkątną dolną L i trójkątną górną U, takie że $LU = PA$. Korzystając z tych macierzy, rozwiąż układ równań liniowych $Ax = \mathbf{b}$.

5. Udowodnij, że jeśli macierz A $n \times n$ jest symetryczna (i dodatnio określona) i współczynnik $a_{11} \neq 0$, to po wykonaniu pierwszego kroku eliminacji Gaussa blok $n-1 \times n-1$ w prawym dolnym rogu przekształconej macierzy jest symetryczny (i dodatnio określony).

6. Udowodnij, że jeśli macierz A jest diagonalnie dominująca (tj. $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$ dla każdego i), to własność tę mają wszystkie macierze wytworzone w kolejnych krokach eliminacji Gaussa bez wyboru elementu głównego.

7. Macierz rzadką, tj. taką, której tylko niewielka część współczynników różni się od zera, można reprezentować na wiele sposobów, w ogólności zależnych od rozmieszczenia niezerowych współczynników.

   Na przykład współczynniki macierzy trójdiagonalnej $n \times n$ najprościej jest przechowywać w trzech tablicach o długości n, ze współczynnikami odpowiednio z dolnej kodiagonali (`a[i]`$=a_{i,i-1}$), diagonali (`b[i]`$=a_{i,i}$) i górnej kodiagonali (`c[i]`$=a_{i,i+1}$). Jeśli taka macierz została rozłożona za pomocą eliminacji Gaussa na czynniki trójkątne, to ich współczynniki można przechowywać w tych samych tablicach. Jeśli przestawianie wierszy jest potrzebne, to do przechowywania czynników rozkładu jest potrzebna jeszcze jedna tablica, na współczynniki $u_{i,i+2}$ macierzy U (i tablicę liczb całkowitych, do reprezentowania wykonanej permutacji wierszy).

Napisz procedury eliminacji Gaussa bez wyboru elementu głównego dla macierzy trójdiagonalnej i cyklicznej trójdiagonalnej.

8. Macierz $n \times n$

$$A = \begin{bmatrix} H & \mathbf{d} \\ \mathbf{d}^T & c \end{bmatrix}$$

ma blok $H = I - 2\boldsymbol{w}\boldsymbol{w}^T$, gdzie $\boldsymbol{w} \in \mathbb{R}^{n-1}$ jest wektorem jednostkowym. Jaki warunek muszą spełniać wektory $\boldsymbol{w}$ i $\mathbf{d}$ oraz liczba c, aby macierz A była nieosobliwa? Zaproponuj algorytm rozwiązywania układów równań liniowych z taką macierzą, przy założeniu, że dane są wektory $\boldsymbol{w}$ i $\mathbf{d}$, liczba c i wektor prawej strony.

9. Odbicie symetryczne w przestrzeni $\mathbb{C}^n$ jest reprezentowane przez macierz określoną wzorem

$$H = I - \boldsymbol{v}\frac{1}{\boldsymbol{v}^H\boldsymbol{v}}\boldsymbol{v}^H,$$

w którym $\boldsymbol{v} \in \mathbb{C}^n \setminus \{0\}$ jest wektorem normalnym hiperpłaszczyzny odbicia. Sprawdź, że to przekształcenie jest inwolucją i izometrią w przestrzeni $\mathbb{C}^n$.

Niech $\boldsymbol{a}$ będzie niezerowym wektorem, którego obraz w odbiciu (Householdera) ma mieć kiernek wektora $\boldsymbol{e}_1 = [1, 0, \ldots, 0]^T$. Podaj konstrukcję, która zapewni otrzymanie najdłuższego wektora normalnego hiperpłaszczyzny odbicia w $\mathbb{C}^n$.

Wskazówka: Znak liczby rzeczywistej to $+1$, 0 lub -1. Znak liczby zespolonej różnej od zera może być dowolną liczbą z, taką że $|z| = 1$.

10. Macierz o strukturze blokowej

$$\begin{bmatrix} A & B \\ B^T & 0 \end{bmatrix},$$

jeśli blok A jest symetryczny i dodatnio określony, a kolumny macierzy B są liniowo niezależne, jest nieosobliwa. Zaproponuj korzystający z tych informacji algorytm rozwiązywania układu równań liniowych z taką macierzą. Oblicz jego koszt, w zależności od wymiarów bloku B.

Wskazówka: Metodą Choleskiego (zobacz wykład i następne zadanie) znajdź macierz trójkątną dolną L, taką że $A = LL^T$. Następnie oblicz macierz $C = L^{-1}B$ (rozwiązując odpowiedni układ równań) i użyj odbić Householdera do znalezienia macierzy ortogonalnej Q i trójkątnej górnej R, takich że $C = QR$ oraz $C^TC = R^TR$.

11. Metodą Choleskiego rozłóż na czynniki trójkątne macierz

$$A = \begin{bmatrix} 4 & 2 & 0 & 2 \\ 2 & 2 & 0 & 1 \\ 0 & 0 & 4 & -6 \\ 2 & 1 & -6 & 12 \end{bmatrix}$$

Korzystając z tego rozkładu rozwiąż układ równań $Ax = \mathbf{b}$ z wektorem prawej strony $\mathbf{b} = [4, 1, -6, 13]^T$.

12. Wielką macierz rzadką o dowolnym (np. całkowicie nieregularnym) rozmieszczeniu niezerowych współczynników można reprezentować za pomocą wykazu tych współczynników. Wykaz jest tablicą, której każdy element jest trójką liczb; pierwsze dwie to indeksy współczynnika (tj. numery wiersza i kolumny), a trzecia liczba to ten współczynnik.

    Napisz podprogram, który oblicza iloczyn Ax macierzy rzadkiej reprezentowanej przez opisany wyżej wykaz i danego wektora x. Nie zakładaj żadnego szczególnego uporządkowania wykazu.

13. (Lab.) Niech $\mathbf{p}_1 = [0, 0]^T$, $\mathbf{p}_2 = [3.141, 2.718]^T$, $\mathbf{q}_1 = [-1, 1]^T$, $\mathbf{q}_2 = [-0.961, 1.033]^T$. Korzystając z pakietu Octave, znajdź punkt wspólny dwóch prostych przechodzących odpowiednio przez punkty $\mathbf{p}_1$, $\mathbf{p}_2$ i $\mathbf{q}_1$, $\mathbf{q}_2$. Powtórz eksperyment, zamieniając punkt $\mathbf{q}_2$ na $[-0.961, 1.034]^T$. Oblicz normy 1, 2 i ∞ oraz wskaźniki uwarunkowania w tych normach macierzy układu równań liniowych, który należy rozwiązać, aby znaleźć punkt przecięcia prostych.

14. Przypuśćmy, że w roli macierzy C, użytej do poprawiania uwarunkowania układu równań $Ax = \mathbf{b}$ na potrzeby metody sprzężonych gradientów, występuje macierz symetryczna i dodatnio określona B taka że $A = B^2$ (macierz B rozpatrywaliśmy w opisie metody CG). Jak będą przebiegać obliczenia w metodzie CG?

    Jeśli taka macierz B jest skuteczna, to dlaczego jej się nie używa?

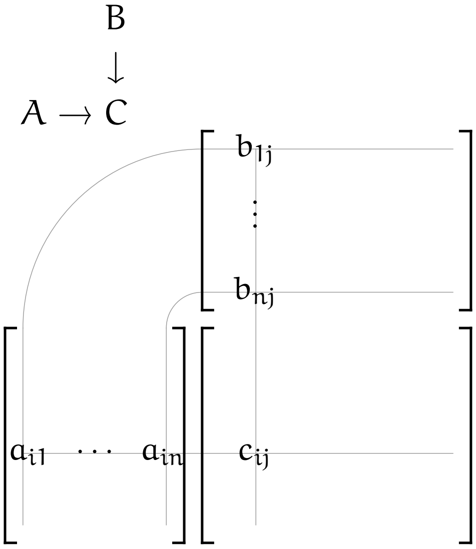
15. Niech macierze A i S będą symetryczne i dodatnio określone, przy czym istnieje macierz C, taka że $C^T C = S$ i $\operatorname{cond}(C^{-T} A C^{-1}) \ll \operatorname{cond} A$. Mając możliwość szybkiego rozwiązywania układów równań z macierzą S (np. podprogram, który to robi), napisz wzory będące podstawą metody Richardsona rozwiązywania układu $Ax = \mathbf{b}$ z poprawianiem uwarunkowania przy użyciu macierzy S (metoda ma wykonywać iteracje dla układu z symetryczną macierzą $C^{-T} A C^{-1}$).

Przypomnienia

1. <u>Mnożenie macierzy — schemat Falka</u>.
   Niech $C = AB$. Znany z algebry liniowej wzór

   $$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj},$$

   definiujący możenie macierzy, ma znakomitą interpretację graficzną, pokazaną na rysunku obok. Za pomocą takiego przedstawienia można łatwo udowodnić wiele własności mnożenia macierzy, a także zobaczyć, które współczynniki iloczynu zależą od poszczególnych współczynników czynników i w szczególności, jeśli czynniki te mają pewną strukturę (tj. określone rozmieszczenie współczynników zerowych), to można stąd wyciągnąć wnioski na temat struktury iloczynu (a stąd już droga do efektywnych algorytmów mnożenia macierzy, które mają wiele zerowych współczynników).



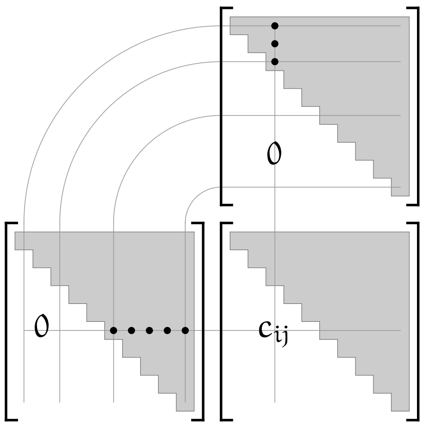
2. Schemat Falka można zastosować nie tylko do pojedynczych współczynników, ale także do *bloków* macierzy, np. tak, jak na rysunku obok. W podziale blokowym w każdym wierszu wszystkie bloki mają tyle samo wierszy, w każdej kolumnie wszystkie bloki mają tyle samo kolumn i aby mnożenie było wykonalne, liczby kolumn bloków macierzy A (pierwszego czynnika) muszą być równe liczbom wierszy odpowiednich bloków macierzy B (drugiego czynnika).

$$\begin{array}{cc} & \begin{bmatrix} B_{11} & B_{12} \end{bmatrix} \\ \begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix} & \begin{bmatrix} A_{11}B_{11} & A_{11}B_{12} \\ A_{21}B_{11} & A_{21}B_{12} \end{bmatrix} \end{array}$$

3. Na schemacie Falka łatwo jest pokazać, że iloczyn macierzy trójkątnych górnych jest macierzą trójkątną górną. Podobnie to wygląda dla macierzy trójkątnych dolnych.

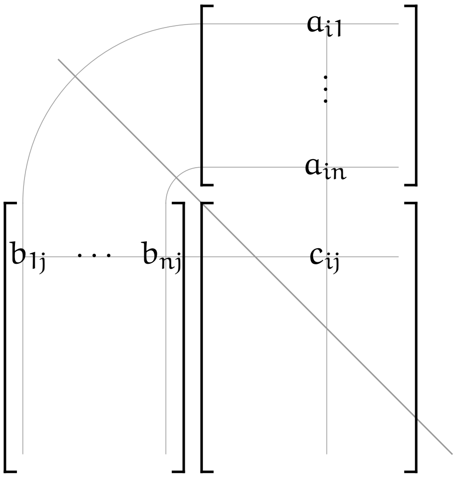
   Co można powiedzieć mnożeniu macierzy blokowo-trójkątnych?

   Narysuj odpowiedni schemat Falka, aby pokazać, że iloczyn dwóch macierzy trójdiagonalnych jest macierzą pięciodiagonalną. Uogólnij to stwierdzenie na przypadek mnożenia macierzy wstęgowych.



4. Ze schematu Falka natychmiast widać, że $(AB)^T = B^T A^T$ — o ile mnożenie współczynników jest przemienne (co, rzecz jasna, ma miejsce dla macierzy liczbowych, ale nie tylko takie macierze występują przy różnych okazjach).

$$\begin{array}{ccc} & A^T & \\ & \downarrow & \\ B^T & \to & (AB)^T \end{array}$$



Na co trzeba zwrócić uwagę, jeśli rozpatrujemy transpozycję iloczynu macierzy podzielonych na bloki?

Korzystając ze schematu Falka uzasadnij stwierdzenie, że dla każdej macierzy kwadratowej nieosobliwej A zachodzi równość $(A^T)^{-1} = (A^{-1})^T$ (a zatem można pisać w skrócie A^{-T}).

5. Struktura blokowa iloczynu macierzy przedstawionego na schemacie obok uzasadnia stwierdzenie, że otrzymana w eliminacji Gaussa macierz $L = L_1 \dots L_{n-1}$ ma pod diagonalą współczynniki $l_{ik} = a_{ik}^{(k-1)}/a_{kk}^{(k-1)}$; polecam przeprowadzenie dowodu indukcyjnego tego faktu.

$$\begin{array}{cc} & \begin{bmatrix} I & 0 \\ 0 & D \end{bmatrix} \\ \begin{bmatrix} B & 0 \\ C & I \end{bmatrix} & \begin{bmatrix} B & 0 \\ C & D \end{bmatrix} \end{array}$$

6. Eliminacja Gaussa z częściowym wyborem elementu głównego daje w wyniku macierz

$$U = \tilde{L}_{n-1}^{-1} T_{n-1} \dots \tilde{L}_2^{-1} T_2 \tilde{L}_1^{-1} T_1 A.$$

W powyższym wzorze symbole $T_1, \dots, T_{n-1}$ oznaczają macierze transpozycji; macierz T_k reprezentuje przestawienie wierszy k i l dla pewnego $l \geqslant k$ (jeśli $k = l$ to T_k jest macierzą jednostkową). Niech $P_k = T_{n-1} \dots T_k$. Macierz P_k opisuje pewną permutację ostatnich $n+1-k$ wierszy, ma zatem strukturę blokową

$$P_k = \begin{bmatrix} I & & \\ & 1 & \\ & & \tilde{P}_k \end{bmatrix},$$

w której blok $\tilde{P}_k \in \mathbb{R}^{n+1-k \times n+1-k}$ jest pewną macierzą permutacji. Jak wiemy, dowolna macierz permutacji jest ortogonalna, tj. jej transpozycja jest jej odwrotnością. Macierz $\tilde{L}_k^{-1}$ ma postać

$$\tilde{L}_{k-1}^{-1} = \begin{bmatrix} I & & \\ & 1 & \\ & -\tilde{l}_{k-1} & I \end{bmatrix}$$

z blokiem $-\tilde{l}_{k-1} \in \mathbb{R}^{n+1-k}$. Rozważmy iloczyn

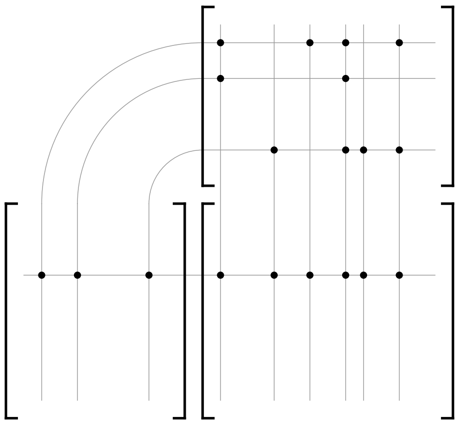
$$P_k \tilde{L}_{k-1}^{-1} T_{k-1} = P_k \tilde{L}_{k-1}^{-1} P_k^T P_k T_{k-1} = P_k \tilde{L}_{k-1}^{-1} P_k^T P_{k-1}.$$

Ze schematu Falka dla pierwszych trzech czynników ostatniego wyrażenia otrzymujemy

$$\begin{array}{ccc} & \begin{bmatrix} I & & \\ & 1 & \\ & -\tilde{l}_{k-1} & I \end{bmatrix} & \begin{bmatrix} I & & \\ & 1 & \\ & & \tilde{P}_k^T \end{bmatrix} \\ \begin{bmatrix} I & & \\ & 1 & \\ & & \tilde{P}_k \end{bmatrix} & \begin{bmatrix} I & & \\ & 1 & \\ & -\tilde{P}_k\tilde{l}_{k-1} & \tilde{P}_k \end{bmatrix} & \begin{bmatrix} I & & \\ & 1 & \\ & -\tilde{P}_k\tilde{l}_{k-1} & I \end{bmatrix} \end{array}$$

Zatem, $P_k\tilde{L}_{k-1}^{-1}T_k = L_{k-1}^{-1}P_{k-1}$ i macierz L_{k-1}^{-1} jest trójkątna dolna z odpowiednio przestawionymi współczynnikami w k—pierwszej kolumnie pod diagonalą. Stąd za pomocą indukcji łatwo już pokazać, że $U = L_{n-1}^{-1}\dots L_1^{-1}P_1A = L^{-1}PA$; macierz permutacji $P = P_1$ opisuje złożenie wszystkich wykonanych przestawień wierszy. Wynik eliminacji jest taki sam, jak gdyby wiersze macierzy A zostały poprzestawiane przed przystąpieniem do eliminacji.

7. Schemat Falka ułatwia opracowanie algorytmu mnożenia macierzy rzadkich, reprezentowanych za pomocą wykazów niezerowych współczynników (zobacz zadanie 12). W tym algorytmie mnożenie i dodawanie zmiennopozycyjne jest wykonywane tylko wtedy, gdy *oba* mnożone współczynniki macierzy A i B są niezerowe.



Aby znaleźć iloczyn macierzy A i B, wykazy ich niezerowych współczynników (zobacz zadanie 12) porządkujemy tak, aby współczynniki w kolejnych wierszach znalazły się obok siebie; znakomicie nadaje się do tego algorytm sortowania kubełkowego. Dla każdej macierzy tworzymy tablicę dodatkową, która dla każdego wiersza macierzy zawiera indeks do pierwszego elementu reprezentującego współczynnik z tego wiersza w uporządkowanym wykazie.

Następnie obliczamy niezerowe współczynniki iloczynu w jego kolejnych wierszach. Dla i-tego wiersza macierzy A przeglądamy elementy jej wykazu. Przetwarzając element reprezentujący współczynnik a_{ik} w k-tej kolumnie, dla wszystkich elementów wykazu macierzy B, reprezentujących niezerowe współczynniki w k-tym wierszu, tworzymy trójki (p, q, j), gdzie p jest indeksem (w wykazie A) współczynnika a_{ik}, q jest indeksem (w wykazie B) niezerowego współczynnika b_{kj}, zaś j jest indeksem kolumny tego współczynnika.

Po posortowaniu tablicy takich trójek względem indeksów j trójki reprezentujące iloczyny $a_{ik}b_{kj}$, które trzeba zsumować, występują obok siebie. Wykonanie działań zmiennopozycyjnych i utworzenie elementów wykazu niezerowych współczynników w i-tym wierszu iloczynu jest już łatwe.

Liniowe zadania najmniejszych kwadratów

Rozważamy układ równań liniowych $Ax = \mathbf{b}$ z macierzą $A \in \mathbb{R}^{m \times n}$ i wektorem $\mathbf{b} \in \mathbb{R}^m$. Układ ten może (choć nie musi) być sprzeczny. Liniowe zadanie najmniejszych kwadratów (LZNK) polega na znalezieniu wektora $x^\star$, takiego że norma druga wektora residuum, $\mathbf{b} - Ax^\star$, jest najmniejsza. Jeśli układ jest niesprzeczny, to rozwiązanie LZNK jest zwykłym rozwiązaniem tego układu.

Twierdzenie. *LZNK ma rozwiązanie; jest nim taki wektor $x^\star$, że wektor residuum jest prostopadły (w sensie iloczynu skalarnego $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{v}^T \mathbf{u}$) do przestrzeni liniowej (podprzestrzeni $\mathbb{R}^m$) rozpiętej przez kolumny $\mathbf{a}_1, \ldots, \mathbf{a}_n$ macierzy A.*

Dowód. Rozważmy wektor $\mathbf{y}^\star$, który jest rzutem prostopadłym wektora $\mathbf{b}$ na tę podprzestrzeń. Zatem istnieje wektor $x^\star$, taki że $\mathbf{y}^\star = Ax^\star$ i wektor $\mathbf{b} - \mathbf{y}^\star = \mathbf{b} - Ax^\star$ (czyli residuum) jest prostopadły do tej podprzestrzeni. Jeśli weźmiemy dowolny wektor $x \in \mathbb{R}^n$ i obliczymy $\mathbf{y} = Ax$, to wektor $\mathbf{y} - \mathbf{y}^\star = A(x - x^\star)$ jest prostopadły do wektora $\mathbf{b} - \mathbf{y}^\star$. Ale wtedy, na podstawie twierdzenia Pitagorasa, mamy

$$\|\mathbf{b} - Ax\|_2^2 = \|\mathbf{b} - \mathbf{y}\|_2^2 = \|\mathbf{b} - \mathbf{y}^\star\|_2^2 + \|\mathbf{y}^\star - \mathbf{y}\|_2^2 \geqslant \|\mathbf{b} - \mathbf{y}^\star\|_2^2 = \|\mathbf{b} - Ax^\star\|_2^2.$$

Dla $\mathbf{y} \neq \mathbf{y}^\star$ powyższa nierówność jest ostra. $\square$

LZNK ma rozwiązanie jednoznaczne wtedy i tylko wtedy, gdy kolumny macierzy A są liniowo niezależne (co jest możliwe tylko dla $m \geqslant n$). Zadania z takimi macierzami to tzw. regularne liniowe zadania najmniejszych kwadratów (RLZNK).

Uwaga. Wektor $\mathbf{y}^\star$ w tym przypadku jest równy $A(A^T A)^{-1} A^T \mathbf{b}$; możemy sprawdzić, że $A^T A$ jest macierzą symetryczną i dodatnio określoną. Zatem określony tym wzorem wektor $\mathbf{y}^\star$ istnieje, jest on kombinacją liniową kolumn macierzy A, a ponadto

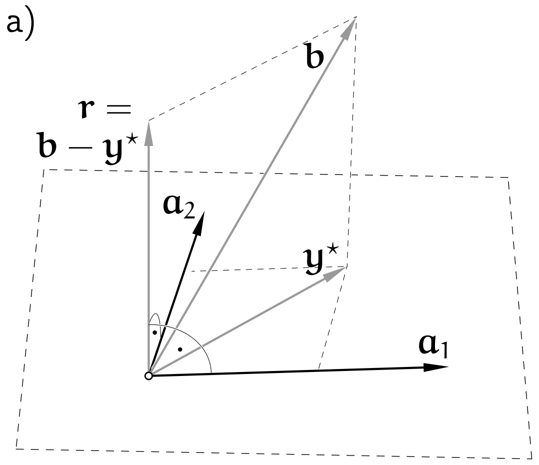
$$A^T(\mathbf{b} - \mathbf{y}^\star) = A^T \mathbf{b} - A^T A (A^T A)^{-1} A^T \mathbf{b} = 0.$$

Wektor $\mathbf{b} - \mathbf{y}^\star$ jest więc prostopadły do wszystkich kolumn macierzy A.

Jeśli macierz ma kolumny liniowo zależne, to zadanie (nieregularne, NLZNK) ma wiele rozwiązań, ich zbiór jest warstwą przestrzeni $\mathbb{R}^n$ o wymiarze $n - r$ (gdzie r oznacza rząd macierzy A).

W tym przypadku można napisać wzór $\mathbf{y}^\star = B(B^T B)^{-1} B^T \mathbf{b}$, w którym symbol B

oznacza macierz powstałą z A przez wyrzucenie kolumn tak, aby pozostawić $k = \operatorname{rank} A$ kolumn liniowo niezależnych.



Ilustracje liniowych zadań najmniejszych kwadratów mamy na rysunku. Rysunek a) przedstawia zadanie regularne dla układu z macierzą 3×2. Kolumny macierzy $A = [\mathbf{a}_1, \mathbf{a}_2]$ rozpinają dwuwymiarową podprzestrzeń przestrzeni $\mathbb{R}^3$, nie zawierającą wektora prawej strony $\mathbf{b}$. Ponieważ kolumny te są liniowo niezależne, rzut prostopadły $\mathbf{y}^\star$ wektora $\mathbf{b}$ na tę podprzestrzeń jest ich kombinacją liniową o jednoznacznie określonych współczynnikach — współrzędnych wektora $\mathbf{x}^\star$, który jest jedynym rozwiązaniem tego zadania.

Na rysunku b) jest pokazane zadanie nieregularne, z macierzą 3×3 o liniowo zależnych kolumnach. Kolumny te rozpinają przestrzeń dwuwymiarową, której elementem wektor $\mathbf{b}$ nie jest. Jego rzut prostopadły $\mathbf{y}^\star$ na tę podprzestrzeń jest jednoznacznie określony, ale można go wyrazić jako kombinację liniową kolumn $\mathbf{a}_1$, $\mathbf{a}_2$, $\mathbf{a}_3$ na nieskończenie wiele sposobów i właśnie tyle rozwiązań ma zadanie.

Regularne LZNK

Prostopadłość dowolnego wektora w $\mathbb{R}^m$ do podprzestrzeni jest równoważna prostopadłości tego wektora do wszystkich elementów dowolnej bazy tej podprzestrzeni. Zatem, mając układ równań $A\mathbf{x} = \mathbf{b}$, możemy pomnożyć skalarnie residuum przez kolumny macierzy A i przyrównać do zera:

$$\langle \mathbf{b} - A\mathbf{x}, \mathbf{a}_i \rangle = 0, \quad i = 1, \ldots, n.$$

Można to zapisać w postaci macierzowej, po prostych przekształceniach otrzymując tzw. <u>układ równań normalnych</u>:

$$A^T A\mathbf{x} = A^T\mathbf{b}.$$

Jeśli kolumny $\mathbf{a}_1, \ldots, \mathbf{a}_n$ są liniowo niezależne, to ich zbiór jest bazą odpowiedniej podprzestrzeni; wtedy macierz symetryczna $M = A^T A$ jest dodatnio określona (skąd wynika, że nieosobliwa) i układ ma jednoznaczne rozwiązanie — rozwiązanie RLZNK (jeśli kolumny są liniowo zależne, to układ równań normalnych jest niesprzeczny, ale ma nieskończenie wiele rozwiązań, którymi są wszystkie rozwiązania NLZNK).

Algorytm równań normalnych jest najprostszą i najtańszą metodą numeryczną rozwiązywania RLZNK. Polega on na obliczeniu macierzy $M = A^T A$ i wektora $\mathbf{d} = A^T \mathbf{b}$, a następnie rozwiązaniu układu równań $M\mathbf{x} = \mathbf{d}$. Ponieważ macierz M jest symetryczna, obliczenie jej współczynników może być wykonane kosztem $mn(n+1)/2$ działań (mnożeń i dodawań zmiennopozycyjnych). Układ równań z macierzą M może być rozwiązany metodą Choleskiego.

Większą dokładność rozwiązania można osiągnąć, korzystając z rozkładu ortogonalno-trójkątnego macierzy A. Dla ustalonej macierzy $A \in \mathbb{R}^{m\times n}$ istnieje macierz ortogonalna $Q \in \mathbb{R}^{m\times m}$ i macierz $R \in \mathbb{R}^{m\times n}$, której współczynniki poniżej diagonali są zerowe, takie że $A = QR$, przy czym jeśli macierz A ma liniowo niezależne kolumny, to macierz R również (zatem ma niezerowe współczynniki diagonalne). Dla $m \geqslant n$ pierwsze n kolumn macierzy Q i wierszy macierzy R są określone z dokładnością do zwrotów.

Macierze Q i R podzielimy na bloki, odpowiednio

$$Q = [Q_1, Q_2], \qquad R = \begin{bmatrix} R_1 \\ R_2 \end{bmatrix},$$

takie że $Q_1 \in \mathbb{R}^{m\times n}$ i $R_1 \in \mathbb{R}^{n\times n}$. Ponieważ blok R_2 jest zerowy, mamy $A = Q_1 R_1$. Podstawmy to do układu równań normalnych:

$$R_1^T Q_1^T Q_1 R_1 \mathbf{x} = R_1^T Q_1^T \mathbf{b}.$$

Macierz $Q_1^T Q_1$ jest macierzą jednostkową $n \times n$, a ponieważ macierz R_1 jest nieosobliwa, mamy układ równoważny układowi równań normalnych:

$$R_1 \mathbf{x} = Q_1^T \mathbf{b},$$

z nieosobliwą macierzą trójkątną górną R_1.

Jeśli dany układ równań, $A\mathbf{x} = \mathbf{b}$, dla którego stawiamy LZNK, pomnożymy stronami przez Q^T, to możemy w nim wyróżnić dwa podukłady:

$$\begin{cases} R_1 \mathbf{x} = Q_1^T \mathbf{b}, \\ \;\;0\mathbf{x} = Q_2^T \mathbf{b}. \end{cases}$$

Układ dany jest niesprzeczny wtedy i tylko wtedy, gdy wektor $Q_2^T\mathbf{b} = 0$. Co więcej, ponieważ pierwszy podukład ma rozwiązanie jednoznaczne (jest nim rozwiązanie LZNK), a macierz drugiego podukładu jest zerowa, długość wektora $Q_2^T\mathbf{b}$ jest najmniejszą osiągalną normą residuum, $\mathbf{b} - A\boldsymbol{x}^\star$.

Istnieje wiele metod rozkładania macierzy A na czynniki Q i R albo Q_1 i R_1. Jedną z nich jest zastosowanie odbić Householdera. Za pomocą n odbić, konstruowanych tak samo, jak w zastosowaniu do układu równań liniowych z nieosobliwą macierzą $n \times n$, macierz A przekształcamy na macierz R. Macierz ortogonalną Q reprezentujemy za pomocą wektorów normalnych hiperpłaszczyzn kolejnych odbić (które możemy przechowywać w tablicy początkowo zawierającej współczynniki macierzy A); mamy

$$Q^T = H_n H_{n-1} \dots H_1, \quad Q = H_1 \dots H_{n-1} H_n,$$

gdzie $H_i = I - \gamma_i \boldsymbol{v}_i \boldsymbol{v}_i^T$. Macierzy Q *nie wyznaczamy* w postaci jawnej; zapamiętujemy tylko wektory $\boldsymbol{v}_i$ i liczby γ_i.

Algorytm rozwiązywania RLZNK za pomocą odbić składa się z następujących kroków:

1. Znajdź rozkład macierzy A, tj. reprezentację macierzy Q w postaci wektorów odbić i macierz R.
2. Oblicz wektor $\boldsymbol{y} = Q^T\mathbf{b} = H_n \dots H_1\mathbf{b}$.
3. Wybierz pierwsze n wierszy macierzy R i wektora $\boldsymbol{y}$, tj. macierz $R_1 = Q_1^T A$ i wektor $\boldsymbol{y}_1 = Q_1^T\mathbf{b}$, i rozwiąż układ $R_1\boldsymbol{x} = \boldsymbol{y}_1$.

Rozkładu macierzy A na czynniki Q_1 i R_1 możemy dokonać za pomocą ortonormalizacji Grama-Schmidta; są różne numeryczne implementacje tej procedury. W tak zwanym algorytmie modyfikowanym (MGS), dającym (w implementacji używającej arytmetyki zmiennopozycyjnej) dokładniejszy wynik niż algorytm klasyczny z podręcznika algebry liniowej, konstruujemy macierze $A^{(0)} = A, \dots, A^{(n)} = Q_1$. Jeśli kolumny macierzy $A^{(k)}$ oznaczymy $\boldsymbol{a}_1^{(k)}, \dots, \boldsymbol{a}_n^{(k)}$, to obliczamy

<u>for</u> ($k = 1$; $k \leqslant n$; k++) {
  $r_{kk} = \sqrt{\boldsymbol{a}_k^{(k-1)T}\boldsymbol{a}_k^{(k-1)}}$;
  $\boldsymbol{a}_k^{(k)} = \frac{1}{r_{kk}}\boldsymbol{a}_k^{(k-1)}$;
  <u>for</u> ($i = k+1$; $i \leqslant n$; i++) {
    $r_{ki} = \boldsymbol{a}_k^{(k)T}\boldsymbol{a}_i^{(k-1)}$;
    $\boldsymbol{a}_i^{(k)} = \boldsymbol{a}_i^{(k-1)} - r_{ki}\boldsymbol{a}_k^{(k)}$;
  }
}

Wynikiem obliczenia są kolumny $\mathbf{q}_i = \mathbf{a}_i^{(n)}$ macierzy Q_1 i współczynniki r_{ki} na i powyżej diagonali macierzy R_1.

Do rozwiązania RLZNK za pomocą ortonormalizacji służy następujący algorytm:

1. Za pomocą ortonormalizacji Grama-Schmidta znajdź macierze Q_1 i R_1.
2. Oblicz wektor $\mathbf{y}_1 = Q_1^T\mathbf{b}$.
3. Rozwiąż układ równań $R_1\mathbf{x} = \mathbf{y}_1$.

Przyczyna, dla której algorytmy korzystające z rozkładu ortogonalno-trójkątnego dają dokładniejsze wyniki niż algorytm równań normalnych jest taka, że wyjściowe zadanie jest zwykle znacznie lepiej uwarunkowane niż układ równań normalnych. Dlatego błędy zaokrągleń popełnione podczas obliczania macierzy M i jej rozkładania na czynniki trójkątne przenoszą się na wynik ze znacznie większym czynnikiem. Tymczasem uwarunkowanie układu równań $R_1\mathbf{x} = Q_1^T\mathbf{b}$ (w normie drugiej) jest takie samo, jak uwarunkowanie zadania wyjściowego.

Dualne LZNK

Inny rodzaj liniowego zadania najmniejszych kwadratów możemy postawić, gdy dany układ równań, $A\mathbf{x} = \mathbf{b}$, jest niesprzeczny i nieokreślony.
W dualnym liniowym zadaniu najmniejszych kwadratów (DLZNK) celem jest wybranie jednego elementu ze zbioru rozwiązań układu $A\mathbf{x} = \mathbf{b}$. Należy wybrać rozwiązanie $\mathbf{x}^\star$ najkrótsze (o najmniejszej normie drugiej), lub takie, aby dla ustalonego wektora $\hat{\mathbf{x}} \in \mathbb{R}^n$ wektor $\mathbf{x}^\star - \hat{\mathbf{x}}$ był najkrótszy; pierwsza sytuacja jest szczególnym przypadkiem drugiej.

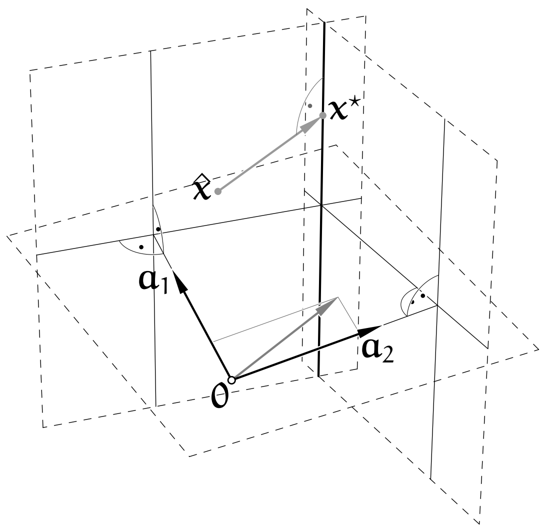
Twierdzenie. *DLZNK ma rozwiązanie. Niech $A^T = [\mathbf{a}_1, \dots, \mathbf{a}_m]$, tj. niech wektory $\mathbf{a}_1, \dots, \mathbf{a}_m \in \mathbb{R}^n$ będą transponowanymi wierszami macierzy A. Rozwiązaniem DLZNK jest taki wektor $\mathbf{x}^\star$, że $A\mathbf{x}^\star = \mathbf{b}$ i różnica $\mathbf{x}^\star - \hat{\mathbf{x}}$ jest kombinacją liniową wektorów $\mathbf{a}_1, \dots, \mathbf{a}_m$.*

Dowód. Zbiór rozwiązań równania liniowego $\mathbf{a}_i^T\mathbf{x} = b_i$ jest warstwą równoległą do podprzestrzeni o wymiarze $n-1$ prostopadłej do wektora $\mathbf{a}_i$. Zbiór rozwiązań całego układu równań $A\mathbf{x} = \mathbf{b}$ jest przecięciem tych warstw, i jest to warstwa przestrzeni $\mathbb{R}^n$ równoległa do podprzestrzeni, do której należą wszystkie wektory prostopadłe do wektorów $\mathbf{a}_1, \dots, \mathbf{a}_m$. Niech $\mathbf{x}^\star$ oznacza rzut prostopadły wektora $\hat{\mathbf{x}}$ na tę warstwę; jest on oczywiście rozwiązaniem układu. Jeśli zatem wektor $\mathbf{x}$ jest dowolnym rozwiązaniem układu $A\mathbf{x} = \mathbf{b}$, to wektor $\mathbf{x} - \mathbf{x}^\star$ jest prostopadły do wektorów $\mathbf{a}_1, \dots, \mathbf{a}_m$, a więc także do ich kombinacji liniowej

$x^\star - \hat{x}$, i z twierdzenia Pitagorasa mamy

$$\|x - \hat{x}\|_2^2 = \|x - x^\star\|_2^2 + \|x^\star - \hat{x}\|_2^2 \geqslant \|x^\star - \hat{x}\|_2^2.$$

Jeśli $x \neq x^\star$, to nierówność jest ostra. □



Ilustrację DLZNK dla układu dwóch równań z trzema niewiadomymi mamy na rysunku. Zbiór rozwiązań układu jest prostą prostopadłą do płaszczyzny rozpiętej przez wektory a_1, a_2, tj. przecięciem dwóch płaszczyzn prostopadłych do tych wektorów. Wektor $x^\star - \hat{x}$ jest prostopadły do tej prostej.

Jeśli więc wektor x jest rozwiązaniem DLZNK, to wektor $x - \hat{x}$ musi być kombinacją liniową transponowanych wierszy macierzy A, a zatem istnieje wektor $y \in \mathbb{R}^m$, taki że $A^T y = x - \hat{x}$. Jeśli strony tej równości pomnożymy przez macierz A, to otrzymamy

$$AA^T y = Ax - A\hat{x}.$$

Po podstawieniu $Ax = b$ mamy stąd układ równań z niewiadomym wektorem y

$$AA^T y = b - A\hat{x},$$

zwany dualnym układem równań normalnych. Macierz AA^T jest symetryczna i jeśli wiersze macierzy A są liniowo niezależne, to jest dodatnio określona. Aby tak było, musi być $n \geqslant m$. Jeśli wiersze macierzy A są liniowo zależne, to nie mamy gwarancji, że układ równań $Ax = b$ jest niesprzeczny, i mamy do czynienia z zadaniem nieregularnym.

Algorytm dualnych równań normalnych polega na obliczaniu macierzy $M = AA^T$ i wektora $d = b - A\hat{x}$, a następnie rozwiązaniu układu $My = d$ (do czego można użyć metody Choleskiego) i obliczeniu rozwiązania $x = \hat{x} + A^T y$. Jeśli ma być znalezione rozwiązanie o najmniejszej normie drugiej, to $\hat{x} = 0$; można wtedy pominąć niektóre obliczenia.

Większą dokładność można uzyskać, korzystając z rozkładu trójkątno-ortogonalnego macierzy A. Istnieje macierz $L \in \mathbb{R}^{m\times n}$, która ma zera za współczynnikiem diagonalnym w każdym wierszu, i macierz ortogonalna $Q \in \mathbb{R}^{n\times n}$, takie że $A = LQ^T$; macierze te można otrzymać, stosując do macierzy A^T (kolumnowo regularnej) te same algorytmy wyznaczania rozkładu ortogonalno-trójkątnego, których użycie do rozwiązania RLZNK było opisane wcześniej. Otrzymujemy macierze $L = [L_1, L_2]$ i $Q = [Q_1, Q_2]$, w których blok $L_1 \in \mathbb{R}^{m\times m}$ jest nieosobliwą macierzą trójkątną dolną, blok L_2 jest zerowy, i macierze L_1 i Q_1 są określone z dokładnością do zwrotów kolumn. Zachodzi równość $A = L_1 Q_1^T$.

Po podstawieniu czynników rozkładu do dualnego układu równań normalnych mamy

$$L_1 Q_1^T Q_1 L_1^T \boldsymbol{y} = \boldsymbol{b} - L_1 Q_1^T \hat{\boldsymbol{x}},$$

a ponieważ $Q_1^T Q_1 = I$ i macierz L_1 jest nieosobliwa, mamy układ równoważny

$$L_1^T \boldsymbol{y} = L_1^{-1}\boldsymbol{b} - Q_1^T \hat{\boldsymbol{x}}.$$

Rozwiązując powyższy układ równań, można by obliczyć wektor $\boldsymbol{y}$, a następnie obliczyć $\boldsymbol{x} = \hat{\boldsymbol{x}} + A^T\boldsymbol{y}$, ale ponieważ poza tym wektor $\boldsymbol{y}$ *nie jest do niczego potrzebny*, lepszym rozwiązaniem po znalezieniu czynników rozkładu macierzy A jest użycie *tylko* tych czynników. Oznaczając $\boldsymbol{w} = L_1^{-1}\boldsymbol{b} - Q_1^T\hat{\boldsymbol{x}}$ i podstawiając $\boldsymbol{y} = L_1^{-T}\boldsymbol{w}$, otrzymamy $A^T\boldsymbol{y} = Q_1 L_1^T L_1^{-T}\boldsymbol{w} = Q_1\boldsymbol{w}$. Stąd otrzymujemy algorytm rozwiązywania DLZNK:

1. Za pomocą ortonormalizacji Grama-Schmidta znajdź macierze trójkątną dolną L_1 i kolumnowo-ortogonalną Q_1, takie że $A = L_1 Q_1^T$.
2. Rozwiąż układ równań liniowych $L_1\boldsymbol{z} = \boldsymbol{b}$ i oblicz wektor $\boldsymbol{w} = \boldsymbol{z} - Q_1^T\hat{\boldsymbol{x}}$.
3. Oblicz $\boldsymbol{x} = \hat{\boldsymbol{x}} + Q_1\boldsymbol{w}$.

Powyższy algorytm można zrealizować również za pomocą odbić Householdera, bez jawnego wyznaczania macierzy Q_1. Inny algorytm rozwiązywania DLZNK korzystający z odbić możemy otrzymać w taki sposób: Niech $\boldsymbol{s} = Q^T\boldsymbol{x}$ i $\hat{\boldsymbol{s}} = Q^T\hat{\boldsymbol{x}}$. Podstawiając nowe wyrażenie do układu $LQ^T\boldsymbol{x} = \boldsymbol{b}$, otrzymujemy układ równań $L\boldsymbol{s} = \boldsymbol{b}$, który możemy przedstawić w postaci $L_1\boldsymbol{s}_1 + L_2\boldsymbol{s}_2 = \boldsymbol{b}$. Ponieważ blok L_2 jest zerowy, wektor $\boldsymbol{s}_1$ musi być rozwiązaniem układu równań $L_1\boldsymbol{s}_1 = \boldsymbol{b}$, zaś wektor $\boldsymbol{s}_2$ trzeba wybrać tak, aby wektor $\boldsymbol{x} - \hat{\boldsymbol{x}} = Q(\boldsymbol{s} - \hat{\boldsymbol{s}})$ miał najmniejszą normę drugą. Ale jest ona równa normie drugiej wektora $s - \hat{s}$. Zatem, jeśli wektor $\hat{s}$ podzielimy (w tym samym miejscu co s) na bloki $\hat{s}_1 = Q_1^T\hat{x}$ i $\hat{s}_2 = Q_2^T\hat{x}$, to aby zminimalizować normę drugą wektora $s - \hat{s}$, musimy przyjąć $s_2 = \hat{s}_2$. Mamy stąd taki algorytm:

1. Znajdź macierz trójkątną dolną L i wektory odbić reprezentujące macierz Q, takie że $A = LQ^T$. Wybierz blok L_1 macierzy L.
2. Oblicz $\hat{s} = Q^T\hat{x}$, stosując odpowiednie odbicia, i podziel na bloki $\hat{s}_1 \in \mathbb{R}^m$, $\hat{s}_2 \in \mathbb{R}^{n-m}$.
3. Rozwiąż układ $L_1 s_1 = b$ i złóż wektor s z bloków s_1 i $s_2 = \hat{s}_2$.
4. Oblicz $x = Qs$, stosując te same odbicia w odwrotnej kolejności.

Rozkład SVD

Do badania i rozwiązywania nieregularnych liniowych zadań najmniejszych kwadratów będzie nam potrzebne twierdzenie o istnieniu pewnego rozkładu macierzy prostokątnych.

Twierdzenie. *Dla dowolnej macierzy $A \in \mathbb{R}^{m\times n}$ istnieją macierze ortogonalne $U \in \mathbb{R}^{m\times m}$ i $V \in \mathbb{R}^{n\times n}$ oraz macierz diagonalna $\Sigma \in \mathbb{R}^{m\times n}$ o nieujemnych współczynnikach, takie że $A = U\Sigma V^T$. Macierz Σ jest określona z dokładnością do uporządkowania współczynników na diagonali, zwanych wartościami szczególnymi macierzy A (ang. singular values). Liczba r dodatnich wartości szczególnych jest rzędem macierzy A.*

Dowód. Macierz A^TA o wymiarach $n \times n$ jest symetryczna i nieujemnie określona. Jej wartości własne są zatem liczbami nieujemnymi i istnieje baza ortogonalna przestrzeni $\mathbb{R}^n$ złożona z wektorów własnych tej macierzy. Zatem istnieje macierz ortogonalna V i macierz diagonalna Λ, takie że

$$V^TA^TAV = \Lambda.$$

Kolumny $v_1, \dots, v_n$ macierzy V są wektorami własnymi macierzy A^TA przynależnymi do wartości własnych $\lambda_1, \dots, \lambda_n$, będących kolejnymi współczynnikami na diagonali macierzy Λ. Dla wygody i uszanowania tradycji uporządkujemy je tak, aby kolejne współczynniki na diagonali macierzy Λ tworzyły ciąg nierosnący. Liczba r dodatnich wartości własnych jest rzędem macierzy Λ, ale też jest rzędem macierzy AV, która ma r początkowych kolumn niezerowych i której każde dwie kolumny, co łatwo zauważyć, są wzajemnie prostopadłe. Ponieważ macierz V jest nieosobliwa, liczba r jest też rzędem macierzy A.

Mamy zatem $\lambda_1 \geqslant \cdots \geqslant \lambda_r > \lambda_{r+1} = \cdots = \lambda_n = 0$; liczba r jest rzędem macierzy A. Dla liczb $\sigma_i = \sqrt{\lambda_i}$, gdzie $i = 1, \ldots, \min\{m, n\}$, określamy macierz diagonalną Σ o wymiarach $m \times n$, której to są współczynniki na diagonali. Niech

$$\boldsymbol{u}_i = \frac{1}{\sigma_i} A\boldsymbol{v}_i, \quad \text{dla } i = 1, \ldots, r.$$

Wtedy $\boldsymbol{u}_i^T \boldsymbol{u}_j = \frac{1}{\sigma_i \sigma_j} \boldsymbol{v}_i^T A^T A \boldsymbol{v}_j$, a stąd wektory $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r$ są jednostkowe i wzajemnie prostopadłe. Dołączając do nich pewne wektory $\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_m$, możemy otrzymać macierz ortogonalną $U = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m]$ (jeśli $r < m$, to można to zrobić na dwa lub nieskończenie wiele sposobów).

Pozostaje sprawdzić, że $A = U\Sigma V^T$, czyli równoważnie $\Sigma = U^T A V$. Obliczmy współczynniki tego ostatniego iloczynu. Dla $i \leqslant r$ jest

$$s_{ii} = \boldsymbol{u}_i^T A \boldsymbol{v}_i = \frac{1}{\sigma_i} \boldsymbol{v}_i^T A^T A \boldsymbol{v}_i = \frac{\lambda_i}{\sigma_i} = \sigma_i.$$

Dla $i \leqslant r$, $j \neq i$ dostajemy

$$s_{ij} = \boldsymbol{u}_i^T A \boldsymbol{v}_j = \frac{1}{\sigma_i} \boldsymbol{v}_i^T A^T A \boldsymbol{v}_j = \frac{\lambda_j}{\sigma_i} \boldsymbol{v}_i^T \boldsymbol{v}_j = 0.$$

Dla $j \leqslant r < i$ otrzymamy

$$s_{ij} = \boldsymbol{u}_i^T A \boldsymbol{v}_j = \boldsymbol{u}_i^T \sigma_j \boldsymbol{u}_j = 0,$$

i wreszcie dla $j > r$ okazuje się, że

$$s_{ij} = \boldsymbol{u}_i^T A \boldsymbol{v}_j = \boldsymbol{u}_i^T 0 = 0. \quad \square$$

Podobne twierdzenie można udowodnić dla macierzy zespolonych; wszędzie w twierdzeniu i w dowodzie transpozycję zastępuje hermitowskie sprzężenie (zamiast macierzy symetrycznych w twierdzeniu występują macierze hermitowskie, zaś odpowiednikiem macierzy ortogonalnych są macierze unitarne, tj. spełniające warunek $U^H U = I$).

Jak widać, macierze U i V *nie są* określone jednoznacznie — nawet dla macierzy kwadratowych o wszystkich wartościach szczególnych jednokrotnych możemy zmieniać zwroty kolumn macierzy U (razem ze zwrotami kolumn macierzy V o tych samych numerach). Jeśli macierz jest prostokątna lub ma wartość szczególną o krotności większej niż 1 (co jest równoważne istnieniu wartości własnej macierzy $A^T A$ o krotności większej niż 1), to macierze U i V można wybierać na nieskończenie wiele sposobów.

Znalezienie opisanego w twierdzeniu rozkładu względem wartości szczególnych (ang. *singular value decomposition*, *SVD*) jest równoważne z rozwiązaniem algebraicznego zagadnienia własnego dla macierzy $A^T A$ i jest to zadanie dosyć trudne (i kosztowne) do rozwiązania numerycznego. Niemniej, rozkład SVD jest bardzo użyteczny, o czym niżej.

Nieregularne LZNK

Jeśli rząd r macierzy A jest mniejszy zarówno od liczby kolumn n, jak i od liczby wierszy m, to liniowe zadanie najmniejszych kwadratów dla układu $Ax = \mathbf{b}$ jest nieregularne. Zbiór rozwiązań takiego zadania jest nieskończony; jest on warstwą $n-r$-wymiarową (przestrzeni $\mathbb{R}^n$), której elementami są takie wektory x, że wektor $y^\star = Ax$ jest rzutem prostopadłym wektora $\mathbf{b}$ na podprzestrzeń rozpiętą przez kolumny macierzy A (tj. residuum, $\mathbf{b} - y^\star$, jest wektorem prostopadłym do tej podprzestrzeni). Dokładnie jeden element tej warstwy ma najmniejszą normę drugą; co więcej, dla dowolnego wektora $\hat{x} \in \mathbb{R}^n$ istnieje dokładnie jeden element $x^\star$ tej warstwy, taki że norma druga różnicy $x^\star - \hat{x}$ jest najmniejsza. Rozwiązanie NLZNK zwykle polega na znalezieniu tego wektora $x^\star$. Rozumowanie podobne do przeprowadzonego wcześniej dla DLZNK uzasadnia stwierdzenie, że wektor $x^\star - \hat{x}$ jest kombinacją liniową transponowanych wierszy macierzy A.

NLZNK *są trudne* do numerycznego rozwiązania. Jest tak dlatego, że rozwiązanie zadania zależy od danych w sposób *paskudnie nieciągły*. NLZNK jest szczególnie trudne, jeśli nie znamy rzędu macierzy A i dopiero mamy go na podstawie obliczeń numerycznych ustalić. Najodporniejsze numeryczne algorytmy rozwiązywania NLZNK korzystają z rozkładu względem wartości szczególnych. Zobaczmy, jak rozkład ten się tu stosuje.

NLZNK dla układu równań $Ax = \mathbf{b}$ i wektora $\hat{x}$ można zastąpić zadaniem równoważnym dla układu równań $\Sigma y = \mathbf{d}$, gdzie $\mathbf{d} = U^T\mathbf{b}$, i wektora $\hat{y} = V^T\hat{x}$ (po rozwiązaniu tego zadania możemy obliczyć $x = Vy$). Załóżmy dla uproszczenia, że $\hat{x} = 0$, czyli $\hat{y} = 0$. Wtedy rozwiązaniem NLZNK dla układu $\Sigma y = \mathbf{d}$, czyli

$$
\begin{bmatrix}
\sigma_1 & & & & \\
& \ddots & & & \\
& & \sigma_r & & \\
& & & 0 & \\
& & & & \ddots \\
& & & & \\
\end{bmatrix}
\begin{bmatrix} y_1 \\ \vdots \\ y_r \\ y_{r+1} \\ \vdots \end{bmatrix}
=
\begin{bmatrix} d_1 \\ \vdots \\ d_r \\ d_{r+1} \\ \vdots \\ d_m \end{bmatrix}
$$

jest wektor o współrzędnych

$$y_i = \begin{cases} d_i/\sigma_i & \text{dla } i \leqslant r, \\ 0 & \text{dla } i > r. \end{cases}$$

Mamy stąd wyjaśnienie trudności zadania: niewielkie zaburzenie macierzy A może spowodować pewne niegroźne zmiany[13] macierzy U i V, oraz zaburzenie macierzy Σ: jeśli dowolna zerowa wartość szczególna zmieni się na niezerową (czyli skutkiem zaburzenia będzie zwiększenie rzędu macierzy A) i $d_i \neq 0$, to trzeba będzie przyjąć $y_i = d_i/\sigma_i$, zamiast zera, dla pewnego $i > r$. Tak więc, *im mniej* zaburzymy macierz A (w sposób zmieniający σ_i), *tym większa* będzie zmiana wyniku.

Rozkład SVD, a dokładniej, jego przybliżenie, tj. macierze ortogonalne $\tilde{U}$ i $\tilde{V}$ oraz macierz diagonalna $\tilde{\Sigma}$, takie że $\tilde{U}\tilde{\Sigma}\tilde{V}^T \approx A$, może być znaleziony za pomocą algorytmu Goluba, który jest blisko związany z tzw. algorytmem QR rozwiązywania algebraicznego zagadnienia własnego; będzie o tym mowa dalej. Macierze $\tilde{U}$ i $\tilde{V}$ są otrzymywane w postaci sfaktoryzowanej, tzn. w postaci ciągów macierzy odbić Householdera i obrotów Givensa (macierze odbić są reprezentowane przez wektory normalne hiperpłaszczyzn odbić, macierze obrotów przez pojedyncze parametry, które są sinusami lub odwrotnościami kosinusów kątów obrotu — zobacz ćwiczenie 11). Można otrzymać te macierze w postaci jawnej, ale lepszym rozwiązaniem jest zachowanie postaci sfaktoryzowanej i korzystanie tylko z niej, aby obliczyć wektory $\tilde{\mathbf{d}} = \tilde{U}^T\mathbf{b}$ i $\hat{\mathbf{y}} = \tilde{V}^T\hat{\mathbf{x}}$, a po rozwiązaniu zadania dla układu $\tilde{\Sigma}\tilde{\mathbf{y}} = \tilde{\mathbf{d}}$ obliczyć $\tilde{\mathbf{x}} = \tilde{V}\tilde{\mathbf{y}}$.

Jeśli znamy rząd r macierzy A, to po znalezieniu rozkładu SVD możemy zamienić na zera obliczone numerycznie wartości szczególne $\tilde{\sigma}_i$ dla $i > r$ — obliczone wartości niezerowe są skutkiem błędów zaokrągleń i aproksymacji popełnionych podczas rozkładania. Jeśli rzędu nie znamy, to możemy przyjąć pewien próg (zależny od oszacowania błędów) i zamienić na zera znalezione wartości szczególne mniejsze od tego progu; to postępowanie nazywa się regularyzacją dyskretną. Inne podejście to tzw. regularyzacja ciągła — do *wszystkich* wartości szczególnych dodajemy pewną liczbę $s > 0$, otrzymując zadanie z macierzą pełnego rzędu, tj. RLZNK, jeśli $m > n$, układ równań z macierzą kwadratową nieosobliwą, jeśli $n = m$, albo DLZNK, jeśli $m < n$. Wybór metody regularyzacji zależy od zastosowania.

[13] Jeśli macierz A ma wartość szczególną o krotności większej niż 1, to czynniki U i V rozkładu SVD zależą od współczynników w sposób nieciągły, ale zaburzenie nie zmieniające rzędu macierzy ma niewielki wpływ na rozwiązanie LZNK; w tym sensie zmiany tych czynników są niegroźne.

Zadania i problemy

1. Niech

$$A = \begin{bmatrix} 1 & 1 & 1 \\ \varepsilon & 0 & 0 \\ 0 & \varepsilon & 0 \\ 0 & 0 & \varepsilon \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 3 \\ \varepsilon \\ \varepsilon \\ \varepsilon \end{bmatrix}.$$

Jaki będzie skutek użycia arytmetyki pojedynczej precyzji do rozwiązania LZNK dla układu $Ax = \mathbf{b}$ za pomocą algorytmu równań normalnych, jeśli $|\varepsilon| < 10^{-4}$?

2. Tabela zawiera wyniki pomiarów (z błędami) wartości pewnej funkcji:

x	-2	-1	1	2
$f(x)$	-1	5	2	14

Przy założeniu, że funkcja ma postać $f(x) = a_0 + a_1(x^2 - 4)$, znajdź liczby a_0 i a_1 najlepiej pasujące do wyników tych pomiarów. Postaw i rozwiąż w tym celu LZNK, używając algorytmu równań normalnych oraz odbić Householdera.

3. Pseudoodwrotność macierzy $A \in \mathbb{R}^{m\times n}$ jest to macierz $A^+ \in \mathbb{R}^{n\times m}$, taka że macierz AA^+ jest macierzą rzutu ortogonalnego przestrzeni $\mathbb{R}^m$ na podprzestrzeń rozpiętą przez kolumny macierzy A, zaś macierz A^+A jest macierzą rzutu ortogonalnego na podprzestrzeń rozpiętą przez transponowane wiersze macierzy A.

Udowodnij, że

a) jeśli macierz A jest kwadratowa nieosobliwa, to $A^+ = A^{-1}$,

b) jeśli macierz A jest kolumnowo regularna, to $A^+ = (A^TA)^{-1}A^T$,

c) jeśli macierz A jest wierszowo regularna, to $A^+ = A^T(AA^T)^{-1}$.

d) w każdym przypadku (czyli zarówno dla macierzy kwadratowej nieosobliwej, jak i macierzy prostokątnej niepełnego rzędu), czynniki rozkładu SVD macierzy $A = U\Sigma V^T$ określają pseudoodwrotność $A^+ = V\Sigma^+U^T$, przy czym Σ^+ jest macierzą diagonalną $n \times m$ ze współczynnikami $1/\sigma_1, \dots, 1/\sigma_r, 0, \dots, 0$ na diagonali (r oznacza rząd macierzy A).

Zbadaj związek pseudoodwrotności z liniowymi zadaniami najmniejszych kwadratów.
Wskazówka: znajdź wzory opisujące rozwiązania odpowiednich układów równań normalnych.

Podstaw iloczyn $U\Sigma V^T$ w miejsce A we wzorach na pseudoodwrotność dla przypadków a), b), c) i sprawdź, że w każdym przypadku powstaje wzór d).

4. Sprawdź dla czterech przypadków podanych wyżej, że pseudoodwrotność macierzy A spełnia tzw. warunki Penrose'a:

$$1\colon A^+A = (A^+A)^T, \quad 2\colon AA^+ = (AA^+)^T, \quad 3\colon AA^+A = A, \quad 4\colon A^+AA^+ = A^+.$$

Dla dowolnej macierzy A istnieje dokładnie jedna macierz A^+ spełniająca te warunki. Przypuśćmy, że dwie macierze, B i C, spełniają te warunki. Wtedy

$$\begin{aligned} B &\underset{4}{=} BAB \underset{3}{=} B(ACA)B \underset{3}{=} B(ACA)C(ACA)B = (BA)(CA)C(AC)(AB) \\ &\underset{1,2}{=} (BA)^T(CA)^TC(AC)^T(AB)^T = A^TB^TA^TC^TCC^TA^TB^TA^T \\ &= (ABA)^TC^TCC^T(ABA)^T \underset{3}{=} A^TC^TCC^TA^T = (CA)^TC(AC)^T \underset{1,2}{=} CACAC \\ &\underset{3}{=} CAC \underset{4}{=} C. \end{aligned}$$

5. Wykaż, że $(A^+)^T = (A^T)^+$ (co uprawnia do używania oznaczenia A^{+T}).
6. LZNK z więzami. Dany układ równań liniowych jest podzielony na dwa podukłady, z macierzami $B \in \mathbb{R}^{m\times n}$ i $C \in \mathbb{R}^{k\times n}$ (takimi, że $k < n < m + k$):

$$\begin{cases} B\mathbf{x} = \mathbf{d}, \\ C\mathbf{x} = \mathbf{e}, \end{cases}$$

przy czym macierz $A = \left[\begin{smallmatrix} B \\ C \end{smallmatrix}\right]$ jest kolumnowo regularna, a macierz C jest wierszowo regularna. Zadanie polega na znalezieniu wektora $\mathbf{x}^\star$, takiego że $C\mathbf{x}^\star = \mathbf{e}$ oraz norma druga wektora residuum pierwszego podukładu jest najmniejsza.

Algorytm zamiany zmiennych: Znajdujemy rozkład macierzy C na czynniki L i Q^T, odpowiednio trójkątny dolny i ortogonalny. Dokonujemy zamiany zmiennych, wprowadzając nowy wektor niewiadomy, $\mathbf{y} = Q^T\mathbf{x}$. Macierz L dzielimy na bloki, $L = [L_1, L_2]$, z których pierwszy jest nieosobliwą macierzą trójkątną dolną, a drugi jest macierzą zerową. Macierz Q dzielimy na bloki Q_1 i Q_2. W ten sposób drugi podukład zastępujemy układem równoważnym $L_1\mathbf{y}_1 = \mathbf{e}$, z którego wyznaczymy $\mathbf{y}_1$. Po zamianie zmiennych w pierwszym podukładzie otrzymujemy układ równań liniowych

$$BQ\mathbf{y} = \mathbf{d}, \quad \text{czyli} \quad BQ_1\mathbf{y}_1 + BQ_2\mathbf{y}_2 = \mathbf{d}, \quad \text{czyli} \quad BQ_2\mathbf{y}_2 = \mathbf{d} - BQ_1\mathbf{y}_1.$$

w ogólności sprzeczny. Dla tego układu rozwiązujemy regularne liniowe zadanie najmniejszych kwadratów, po czym na podstawie otrzymanych wektorów $\mathbf{y}_1$ i $\mathbf{y}_2$ możemy obliczyć $\mathbf{x} = Q\mathbf{y}$.

Algorytm z mnożnikami Lagrange'a: Jeśli macierz B jest kolumnowo regularna, to symetryczna macierz $M = B^TB$ jest dodatnio określona. Wtedy kwadrat normy drugiej residuum pierwszego podukładu jest wielomianem drugiego stopnia współrzędnych wektora x, który w każdej warstwie przestrzeni $\mathbb{R}^n$ ma jednoznacznie określone minimum:

$$f(x) = \|d - Bx\|_2^2 = x^T M x - 2x^T B^T d + d^T d.$$

Aby znaleźć minimum funkcji f w zbiorze rozwiązań drugiego podukładu, wystarczy rozwiązać układ równań

$$\begin{bmatrix} M & C^T \\ C & 0 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} B^T d \\ e \end{bmatrix}.$$

Po rozwiązaniu tego układu blok y odrzucamy; jego współrzędne są tzw. mnożnikami Lagrange'a dla postawionego tu zadania minimalizacji z więzami. Podany wyżej blokowy układ równań możemy rozwiązać w podobny sposób, jak układ w zadaniu 7 na s. 4.33, ale zamiast obliczenia macierzy M i zastosowania do niej metody Choleskiego, lepiej jest dokonać rozkładu ortogonalno-trójkątnego macierzy B.

Opracuj szczegóły obu algorytmów (w szczególności określ wymiary i sposoby reprezentowania wszystkich macierzy otrzymanych w obliczeniach).

Uzasadnij stwierdzenie, że macierz BQ_2 przetwarzana w pierwszym algorytmie jest kolumnowo-regularna.

Udowodnij, że drugi algorytm daje rozwiązanie zadania.

7. Opisz, jak zrealizować pierwszy podany na wykładzie algorytm rozwiązywania DLZNK korzystający z rozkładu trójkątno-ortogonalnego, za pomocą odbić Householdera (bez jawnego wyznaczania macierzy Q_1).

8. Głównym elementem metody śledzenia promieni (*ray tracing*), używanej do tworzenia obrazów obiektów trójwymiarowych, jest znajdowanie przecięć półprostych (tzw. promieni) z powierzchniami tych obiektów. Często składają się one z trójkątów. Zadanie polega na takim wstępnym przetworzeniu danych opisujących trójkąt, aby znalezienie przecięcia trwało jak najkrócej (to się robi dla bardzo dużej liczby promieni) i aby dane zajmowały jak najmniej miejsca (zazwyczaj powierzchnie obiektów składają się z wielu trójkątów).

   Dla trójkąta T o niewspółliniowych wierzchołkach p_0, p_1, $p_2 \in \mathbb{R}^3$ możemy obliczyć wektory $v_1 = p_1 - p_0$ i $v_2 = p_2 - p_0$ oraz wektor normalny płaszczyzny trójkąta, $n = v_1 \wedge v_2$. Płaszczyzna ta składa się z punktów p spełniających równanie $n^T(p - p_0) = 0$. Promień o początku w punkcie q i kierunku wektora w ma przedstawienie parametryczne $p = q + tw$, $t > 0$. Podstawiając je do równania płaszczyzny trójkąta, dostajemy równanie

$$n^T(q + tw - p_0) = 0, \quad \text{skąd możemy obliczyć} \quad t = \frac{n^T(p_0 - q)}{n^T w},$$

a następnie punkt $\mathbf{p}$ (jeśli $\mathbf{n}^T\mathbf{w} = 0$, to promień jest równoległy do płaszczyzny trójkąta; nawet jeśli leży w tej płaszczyźnie, to uznajemy, że nie ma przecięcia).

Punkt $\mathbf{p}$ można opisać wzorem $\mathbf{p} = \mathbf{p}_0 + u_1\mathbf{v}_1 + u_2\mathbf{v}_2$, który opisuje niesprzeczny układ trzech równań liniowych z dwiema niewiadomymi, u_1, u_2. Punkt $\mathbf{p}$ należy do trójkąta T wtedy i tylko wtedy, gdy $u_1 \geqslant 0$, $u_2 \geqslant 0$ i $u_1 + u_2 \leqslant 1$. Stawiamy LZNK dla układu $A\mathbf{x} = \mathbf{b}$, gdzie

$$A = [\mathbf{v}_1, \mathbf{v}_2], \quad \mathbf{x} = [u_1, u_2]^T, \quad \mathbf{b} = \mathbf{p} - \mathbf{p}_0 = \mathbf{q} + t\mathbf{w} - \mathbf{p}_0.$$

Z uwagi na mały rozmiar zadania i *niewielką dokładność* wymaganą w grafice, najwygodniej jest wyznaczyć jawnie macierz A^+. W tym celu rozwiązujemy układ równań z niewiadomą macierzą A^+ (zobacz zad. 3)

$$(A^TA)A^+ = A^T.$$

Mając dany promień, możemy obliczyć wektor $\mathbf{b}$ i otrzymać u_1, u_2 przez obliczenie iloczynu $A^+\mathbf{b}$. Ostatecznie, dla trójkąta T wystarczy zapamiętać tylko wierzchołek $\mathbf{p}_0$ i obliczone w preprocesingu wektor $\mathbf{n}$ i macierz A^+ — razem 12 liczb.

9. Znajdź najkrótsze (o najmniejszej normie drugiej) rozwiązanie układu równań liniowych $A\mathbf{x} = \mathbf{b}$, w którym

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 0 & 3 & 3 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 2 \\ 12 \end{bmatrix},$$

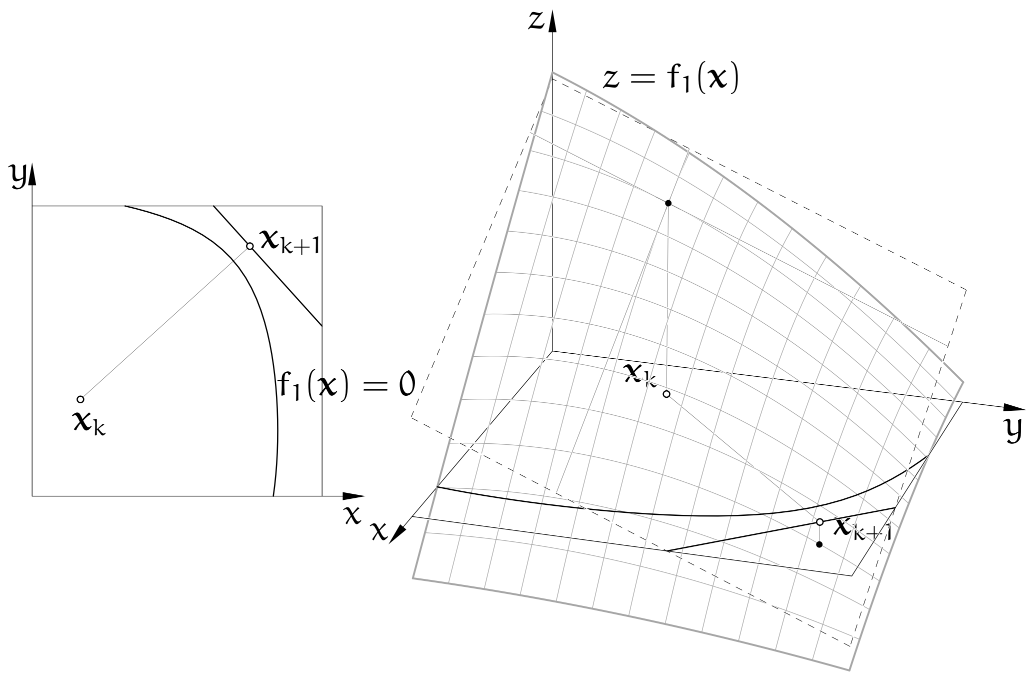
dwoma sposobami: układając i rozwiązując dualny układ równań normalnych i wyznaczając (metodą odbić Householdera) rozkład trójkątno-ortogonalny macierzy A.

10. Metoda Newtona z pseudoodwrotnością służy do numerycznego rozwiązywania układów równań nieliniowych, w których liczba równań, m, jest mniejsza niż liczba niewiadomych, n. Układ równań liniowych $J_k\boldsymbol{\delta} = -\mathbf{f}_k$ jest rozwiązywany jako DLZNK, w celu wyznaczenia najkrótszego spełniającego ten układ wektora $\boldsymbol{\delta}$, po czym przyjmuje się $\mathbf{x}_{k+1} = \mathbf{x}_k + \boldsymbol{\delta}$. W ten sposób, jeśli funkcja $\mathbf{f}$ spełnia odpowiednie warunki, powstaje ciąg $(\mathbf{x}_k)_{k\in\mathbb{N}}$ zbieżny do pewnego rozwiązania $\boldsymbol{\alpha}$ (z nieskończonego zbioru rozwiązań), położonego w pobliżu punktu startowego $\mathbf{x}_0$. Przykład (jedno równanie z dwiema niewiadomymi) na obrazku.

Wykonaj dwa kroki metody Newtona z pseudoodwrotnością dla układu równań

$$\begin{cases} x^2 + y^2 - z^2 = 3 \\ \quad x + y + z = 1 \end{cases}$$

przyjmując $\mathbf{x}_0 = [0, 0, 0]^T$. W rachunkach „ręcznych” układaj i rozwiązuj dualny układ równań normalnych.



<u>Uwaga</u>: Otrzymany ciąg punktów $(\boldsymbol{x}_k)_k$ nie musi zbiegać do rozwiązania równania $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ leżącego najbliżej punktu startowego $\boldsymbol{x}_0$. Jeśli zależy nam na takim rozwiązaniu (np. w zastosowaniu w robotyce), to w każdym kroku metody Newtona z pseudoodwrotnością zamiast *najkrótszego* rozwiązania układu $J_k\boldsymbol{\delta} = -\boldsymbol{f}_k$ możemy wyznaczyć takie rozwiązanie $\boldsymbol{\delta}^*$, aby wektor $\boldsymbol{x}_{k+1} - \boldsymbol{x}_0 = \boldsymbol{x}_k + \boldsymbol{\delta}^* - \boldsymbol{x}_0$ był najkrótszy. W tym celu, rozwiązując DLZNK, należy przyjąć $\hat{\boldsymbol{x}} = \boldsymbol{x}_0 - \boldsymbol{x}_k$ (tj. wektor $\boldsymbol{\delta}^* - \hat{\boldsymbol{x}}$ ma być kombinacją liniową transponowanych wierszy macierzy J_k, zobacz twierdzenie na wykładzie). Jeśli otrzymany w ten sposób ciąg $(\boldsymbol{x}_k)_k$ jest zbieżny, to jego granica *nie musi* być rozwiązaniem równania $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ najbliższym $\boldsymbol{x}_0$, ale jest tzw. punktem krytycznym odległości od $\boldsymbol{x}_0$ w zbiorze rozwiązań tego równania.

11. Niech $c, s \in \mathbb{R}$ i $c^2 + s^2 = 1$. Dla $i \neq j$ macierz G_{ij}, która ma współczynniki $g_{ii} = g_{jj} = c$, $g_{ji} = -g_{ij} = s$, a pozostałe współczynniki takie, jak macierz jednostkowa, jest <u>macierzą obrotu</u> w płaszczyźnie $\mathrm{lin}\{\boldsymbol{e}_i, \boldsymbol{e}_j\}$; istnieje liczba ϕ, taka że $c = \cos\phi$ oraz $s = \sin\phi$ — jest to kąt obrotu. Macierz G_{ij} jest ortogonalna. Iloczyn $G_{ij}A$ macierzy obrotu i dowolnej macierzy A ma takie same wiersze co A, poza i-tym i j-tym. Iloczyn AG_{ij} ma takie same kolumny poza i-tą i j-tą.

    Niech $A \in \mathbb{R}^{n,n}$ będzie macierzą daną i niech $B = G_{ij}A$. Kąt obrotu można dobrać tak, aby współczynnik b_{ij} macierzy B był równy 0; tak skonstruowane przekształcenie nazywa się <u>obrotem Givensa</u>.

    Liczby c i s należy obliczyć za pomocą współczynników a_{jj} oraz a_{ij}, położonych na przecięciu j-tej kolumny macierzy A z wierszami j oraz i. Przypuśćmy, że $a_{ij} \neq 0$; w przeciwnym razie wystarczy wziąć $\phi = 0$, tj. $G_{ij} = I$. Jeśli $a_{ij} \neq 0$, to przyjmujemy $\phi = \operatorname{arc\,tg} a_{ij}/a_{jj}$, skąd mamy

$$c = \frac{a_{jj}}{\sqrt{a_{jj}^2 + a_{ij}^2}}, \qquad s = -\frac{a_{ij}}{\sqrt{a_{jj}^2 + a_{ij}^2}}.$$

Nie ma zatem potrzeby obliczania wartości funkcji trygonometrycznych ani cyklometrycznych, we wzorach jest tylko pierwiastek kwadratowy. Obrót możemy reprezentować za pomocą jednej liczby ξ, określonej według algorytmu Stewarta:

```
if ( |a_jj| > |a_ij| ) { /* ξ = sin φ */
  if ( a_ij ≠ 0 ) { d = a_ij/a_jj; r = √(1 + d²); ξ = d/r; }
  else ξ = 0;
}
else { /* ξ = 1/cos φ */
  if ( a_ij ≠ 0 ) { d = a_jj/a_ij; r = √(1 + d²); ξ = r/d; }
  else ξ = 1;
}
```

Dokonując rozkładu macierzy na czynniki ortogonalny i trójkątny, można liczbę ξ zapamiętać na miejscu współczynnika a_{ij}. Na podstawie ξ możemy obliczyć c i s:

```
if ( |ξ| < 1 ) { c = √(1 − ξ²); s = −ξ; }
else if ( |ξ|==1 ) { c = 1; s = 0; }
else { c = 1/ξ; s = −√(1 − c²); }
```

12. Udowodnij, że jeśli macierz $A \in \mathbb{R}^{m\times n}$ jest kolumnowo-regularna i jest iloczynem macierzy $Q_1 \in \mathbb{R}^{m\times n}$ i $R_1 \in \mathbb{R}^{n\times n}$, takich że macierz Q_1 ma kolumny wzajemnie prostopadłe i o długości 1, zaś macierz R_1 jest trójkątna górna, to rozkład ten jest jednoznaczny z dokładnością do zwrotów kolumn macierzy Q_1 i wierszy macierzy R_1.

    Z powyższego stwierdzenia wynika wniosek, że wszystkie metody znajdowania rozkładu ortogonalno-trójkątnego (Grama-Schmidta, Householdera, Givensa) dają identyczne wyniki z dokładnością do błędów zaokrągleń i zwrotów kolumn Q_1 i wierszy R_1 (ale, z wyjątkiem metody Grama-Schmidta, macierzy Q_1 *nie wyznaczamy* w jawnej postaci, jeśli tylko nie musimy!).

    Zauważ, że jeśli macierz A jest kwadratowa $n \times n$ i ma rząd $n-1$ lub n, to jej rozkład na macierz ortogonalną Q i trójkątną górną R też jest jednoznaczny z dokładnością do zwrotów kolumn Q i wierszy R.

13. Niech A oznacza rzeczywistą macierz kwadratową $n \times n$. Wszystkie czynniki rozkładu SVD macierzy A mają te same wymiary i w szczególności macierz diagonalna Σ jest symetryczna. Możemy napisać

$$A = U\Sigma V^T = UV^TV\Sigma V^T = QS,$$

gdzie $Q = UV^T$ jest macierzą ortogonalną, zaś macierz $S = V\Sigma V^T$ jest symetryczna i nieujemnie określona. Z twierdzenia o rozkładzie SVD wynika zatem istnienie rozkładu macierzy kwadratowej na opisane tu czynniki Q i S; nazywa się to

<u>rozkładem biegunowym</u>[14]. Jego geometryczna interpretacja jest taka: każde przekształcenie liniowe $\mathbb{R}^n \to \mathbb{R}^n$ jest złożeniem skalowania (w ogólności nierównomiernego) wzdłuż pewnych osi wzajemnie prostopadłych (o kierunkach kolumn macierzy V) i izometrii.

Macierz Q, będącą ortogonalnym czynnikiem rozkładu biegunowego macierzy A, moglibyśmy otrzymać, „usuwając" z rozkładu SVD czynnik Σ. *Jeśli* macierz A jest nieosobliwa, to macierz Q można otrzymać jako granicę nieskończonego ciągu macierzy $A_0 = A, A_1, A_2, \ldots$, takiego że

$$A_{k+1} = \frac{1}{2}(A_k + A_k^{-T}).$$

Jak to działa? Przyjmijmy założenie indukcyjne, że $A_k = U\Sigma_k V^T$, tzn. macierze U i V w rozkładzie SVD macierzy A_k są takie same dla każdego k. Podstawiając iloczyn $U\Sigma_k V^T$ w miejsce A_k, widzimy, że

$$A_{k+1} = \frac{1}{2}(U\Sigma_k V^T + U\Sigma_k^{-T}V^T) = U\frac{1}{2}(\Sigma_k + \Sigma_k^{-T})V^T = U\Sigma_{k+1}V^T.$$

Współczynnikami diagonalnymi macierzy Σ_{k+1} (czyli wartościami szczególnymi macierzy A_{k+1}) są liczby $\sigma_i^{(k+1)} = (\sigma_i^{(k)} + 1/\sigma_i^{(k)})/2$. Dla każdego $\sigma_i^{(0)} = \sigma_i > 0$ ciąg $(\sigma_i^{(k)})$ zbiega do liczby 1, ponieważ jest to ciąg, który otrzymamy, próbując rozwiązać równanie $x^2 - 1 = 0$ przy użyciu metody Newtona. Zauważywszy to, możemy stwierdzić, że wykładnik zbieżności takiego ciągu jest równy 2.

Przedstawiony wyżej <u>algorytm Highama</u> jest jednym z nielicznych algorytmów numerycznej algebry liniowej, w którym zachodzi konieczność *jawnego wyznaczania* odwrotności macierzy kwadratowych.

[14]Rozkład biegunowy istnieje też dla macierzy zespolonych; w tym przypadku jeden czynnik jest macierzą unitarną, a drugi — hermitowską.

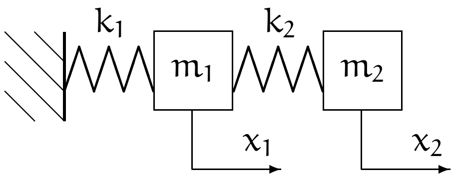
Algebraiczne zagadnienie własne

Niech $A \in \mathbb{R}^{n \times n}$. Jeśli wektor $x \neq 0$ spełnia równanie $Ax = \lambda x$ dla pewnej liczby λ, to mówimy, że jest to wektor własny macierzy A, zaś liczba λ jest to wartość własna tej macierzy; parę (x, λ) nazywamy parą własną macierzy A.

Algebraiczne zagadnienie własne polega na znalezieniu, dla danej macierzy A, jej (wszystkich, kilku lub jednej) wartości własnych albo par własnych. Algebraiczne zagadnienia własne występują w różnych zastosowaniach, np. w mechanice, mają też związek z innymi zadaniami numerycznej algebry liniowej, np. rozwiązywaniem układów równań lub liniowych zadań najmniejszych kwadratów.

Przykład zastosowania

Rozważmy układ złożony z dwóch (może też być ich więcej) ciężarków połączonych ze sobą i z nieruchomym podłożem sprężynkami. Jeśli ciężarki potrącimy, to będą one drgać, przy czym drgania, które są skutkiem tylko początkowego wytrącenia z położenia równowagi, są nazywane drganiami własnymi układu.



Ciężarki mają masy odpowiednio m_1 i m_2; ich odchylenia od położeń równowagi oznaczymy symbolami x_1 i x_2. Każda ze sprężynek działa z siłą proporcjonalną do jej odkształcenia, przy czym współczynniki proporcjonalności oznaczymy odpowiednio k_1 i k_2.

Na pierwszy ciężarek działa siła $-k_1x_1 + k_2(x_2 - x_1) = -(k_1 + k_2)x_1 + k_2x_2$ (uwaga na zwrot: siła jest dodatnia jeśli ma ten sam zwrot co dodatnie przemieszczenie). Na drugi ciężarek działa siła $k_2(x_1 - x_2)$. Każda z tych sił jest równoważona przez bezwładność ciężarka proporcjonalną do jego masy, zatem ruch ciężarków jest opisany przez taki układ równań różniczkowych zwyczajnych:

$$\begin{cases} m_1\ddot{x}_1 = -(k_1 + k_2)x_1 + k_2x_2, \\ m_2\ddot{x}_2 = k_2x_1 - k_2x_2 \end{cases}$$

(każda kropka oznacza tu jednokrotne różniczkowanie względem czasu). Możemy to zapisać w postaci macierzowej:

$$\begin{bmatrix} -(k_1 + k_2) & k_2 \\ k_2 & -k_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} m_1\ddot{x}_1 \\ m_2\ddot{x}_2 \end{bmatrix}.$$

Można (dokonując zamiany zmiennych) przekształcić ten układ tak, aby utrzymać symetrię macierzy z prawej strony, ale to zaniedbamy; zamiast tego weźmy

$$\begin{bmatrix} -(k_1+k_2)/m_1 & k_2/m_1 \\ k_2/m_2 & -k_2/m_2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{bmatrix}.$$

Przypuśćmy, że $x_1(t) = a_1 \sin \omega t$ oraz $x_2(t) = a_2 \sin \omega t$. Wtedy $\ddot{x}_1(t) = -a_1\omega^2 \sin \omega t$ oraz $\ddot{x}_2(t) = -a_2\omega^2 \sin \omega t$. Po podstawieniu i podzieleniu przez $-\sin \omega t$ dostaniemy równanie

$$\begin{bmatrix} (k_1+k_2)/m_1 & -k_2/m_1 \\ -k_2/m_2 & k_2/m_2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \end{bmatrix} = \omega^2 \begin{bmatrix} a_1 \\ a_2 \end{bmatrix},$$

czyli algebraiczne zagadnienie własne z macierzą 2×2. Wektor własny występującej w nim macierzy opisuje amplitudy a_1 i a_2 drgań własnych, zaś odpowiadająca mu wartość własna jest kwadratem prędkości fazowej ω.
Można dowieść, że w zagadnieniach własnych utworzonych dla układów ciężarków podobnych do rozpatrywanego wyżej wszystkie wartości własne są rzeczywiste i dodatnie, a więc hipoteza, że drgania mogą być opisane za pomocą funkcji sinus, znajduje potwierdzenie.

Podstawowe własności

Równanie $A\boldsymbol{x} = \lambda\boldsymbol{x}$ można przepisać w postaci $(A - \lambda I)\boldsymbol{x} = \boldsymbol{0}$. Z tej postaci natychmiast wynika, że para $(\boldsymbol{x}, \lambda)$, w której wektor $\boldsymbol{x} \neq \boldsymbol{0}$, może spełniać to równanie (czyli być parą własną) wtedy i tylko wtedy, gdy macierz $A - \lambda I$ jest osobliwa. To oznacza, że jej wyznacznik jest zerowy. Wyrażenie $\det(A - \lambda I)$ jest wielomianem stopnia n zmiennej λ. Na podstawie zasadniczego twierdzenia algebry (Gauss, 1799 r.), równanie charakterystyczne $\det(A - \lambda I) = 0$ ma rozwiązanie, które jest liczbą rzeczywistą albo zespoloną. Tak więc każda macierz kwadratowa ma jakąś wartość własną. Zbiór (w ogólności zespolonych) wartości własnych dowolnej macierzy A nazywa się widmem tej macierzy; oznaczamy je symbolem $\mathrm{spect}(A)$. Promień spektralny, oznaczany symbolem $\rho(A)$, jest największą liczbą w zbiorze wartości bezwzględnych wartości własnych macierzy A.

Dla ustalonego λ układ równań $(A - \lambda I)\boldsymbol{x} = \boldsymbol{0}$ jest jednorodny; jeśli zatem $\lambda \in \mathbb{R}$ jest wartością własną macierzy A, to zbiór rozwiązań jest podprzestrzenią liniową przestrzeni $\mathbb{R}^n$. Jest to tzw. podprzestrzeń własna macierzy A przynależna do wartości własnej λ. Wymiar tej podprzestrzeni jest nazywany krotnością geometryczną wartości własnej λ. Z kolei, wielomian charakterystyczny można przedstawić w postaci

$$\det(A - \lambda I) = (\lambda_1 - \lambda) \cdot \ldots \cdot (\lambda_n - \lambda).$$

Liczby $\lambda_1, \ldots, \lambda_n$ to wartości własne, które mogą się powtarzać. Liczba wystąpień wartości własnej λ_i w tym rozkładzie jest zwana jej krotnością algebraiczną. Krotność algebraiczna dowolnej wartości własnej jest większa lub równa krotności geometrycznej tej wartości własnej.

O macierzach A i B, dla których istnieje nieosobliwa macierz C, taka że $B = C^{-1}AC$ mówimy, że to są macierze podobne. Podobieństwo macierzy jest oczywiście relacją równoważności. Można udowodnić, że jeśli macierze są podobne, to mają identyczne wartości własne, o identycznych krotnościach algebraicznych i geometrycznych.

Wektory własne przynależne do różnych wartości własnych dowolnej danej macierzy są liniowo niezależne. Jeśli krotność algebraiczna każdej wartości własnej macierzy A jest równa krotności geometrycznej, to suma baz wszystkich podprzestrzeni własnych składa się z n niezależnych liniowo wektorów własnych macierzy A. Ustawmy te wektory w macierz $X = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n]$; macierz ta jest nieosobliwa. Wtedy

$$AX = [A\boldsymbol{x}_1, \ldots, A\boldsymbol{x}_n] = [\lambda_1\boldsymbol{x}_1, \ldots, \lambda_n\boldsymbol{x}_n] = X\Lambda,$$

gdzie macierz Λ jest diagonalna; jej współczynniki diagonalne są wartościami własnymi macierzy A. Możemy napisać równości

$$X^{-1}AX = \Lambda \quad \text{i} \quad X\Lambda X^{-1} = A.$$

Taka macierz A jest zatem podobna do macierzy diagonalnej, mówimy też, że jest diagonalizowalna. Macierz nie jest diagonalizowalna, jeśli co najmniej jedna jej wartość własna ma krotność algebraiczną różną (większą) od geometrycznej.

Przykłady: Macierz

$$\begin{bmatrix} 3 & 4 \\ 4 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} 7 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}$$

jest diagonalizowalna. Macierz

$$\begin{bmatrix} 3 & -4 \\ 4 & 3 \end{bmatrix}$$

też jest diagonalizowalna, ale jej wartości własne są liczbami zespolonymi, $\lambda_1 = (3, -4)$, $\lambda_2 = (3, 4)$, zatem wektory własne — kolumny odpowiedniej macierzy X — mają co najmniej jedną współrzędną zespoloną. Natomiast macierz

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

nie jest diagonalizowalna; krotność algebraiczna wartości własnej 1 jest równa 2, a krotność geometryczna jest równa 1.

Twierdzenie Jordana. *Dla każdej rzeczywistej lub zespolonej macierzy kwadratowej* A *istnieje macierz nieosobliwa* X*, taka że macierz* $J = X^{-1}AX$ *jest blokowo-diagonalna, i jej bloki na diagonali mają postać*

$$J_k = \begin{bmatrix} \lambda_k & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_k \end{bmatrix},$$

zwaną klatką Jordana. Macierz J *jest określona z dokładnością do kolejności klatek. Jeśli macierz* A *jest rzeczywista i ma wszystkie wartości własne rzeczywiste, to istnieje rzeczywista macierz* X*, określająca podobieństwo macierzy* A *i* J*.*

Dowód pomijam. Macierz J to tak zwana postać kanoniczna Jordana macierzy A. Liczba λ_k jest wartością własną macierzy A, przy czym krotność algebraiczna tej wartości własnej jest sumą wymiarów klatek Jordana, w których występuje λ_k, a liczba tych klatek jest jej krotnością geometryczną.

Twierdzenie Schura. *(a) Dla każdej rzeczywistej lub zespolonej macierzy kwadratowej* A *istnieje macierz unitarna* U*, taka że macierz* $G = U^{-1}AU$ *jest trójkątna górna. Jeśli macierz* A *jest rzeczywista i ma wszystkie wartości własne rzeczywiste, to istnieje odpowiednia macierz ortogonalna* U*.*
(b) Jeśli rzeczywista macierz kwadratowa A *ma zespolone wartości własne, to istnieje macierz ortogonalna* Q*, taka że macierz* $G = Q^{-1}AQ$ *jest blokowo-trójkątna górna, przy czym bloki na diagonali mają wymiary* 1×1 *i* 2×2*; zespolone wartości własne macierzy* A *są wartościami własnymi tych bloków* 2×2*.*

Dowód (tylko punkt (a)). Z twierdzenia Jordana wynika istnienie nieosobliwej macierzy X, sprowadzającej macierz A do postaci kanonicznej. Macierz X możemy rozłożyć (np. metodą ortonormalizacji Grama-Schmidta) na czynniki unitarny U i trójkątny górny R; jest $X = UR$. Zatem,

$$A = XJX^{-1} = URJR^{-1}U^{-1} = UGU^{-1},$$

i macierz $G = RJR^{-1}$, będąca iloczynem macierzy trójkątnych górnych, jest trójkątna górna. □

Bez dowodów (zostawionych jako ćwiczenia) podaję kilka dalszych twierdzeń na temat algebraicznego zagadnienia własnego. Niech $A = [a_{ij}]_{i,j} \in \mathbb{C}^{n\times n}$.
Kołem Gerszgorina nazywa się zbiór liczb zespolonych z spełniających nierówność $|z - a_{ii}| \leqslant \sum_{j\neq i} |a_{ij}|$.

Twierdzenie Gerszgorina. *Każda wartość własna macierzy A leży w pewnym kole Gerszgorina.*

Niech $w(x) = a_k x^k + \cdots + a_1 x + a_0$ będzie dowolnym wielomianem. Możemy użyć macierzy A jako argumentu, tj. napisać

$$w(A) = a_k A^k + \cdots + a_1 A + a_0 I.$$

Twierdzenie. *Jeśli macierz A ma parę własną $(\boldsymbol{x}, \lambda)$, to macierz $w(A)$ ma parę własną $(\boldsymbol{x}, w(\lambda))$. Co więcej,* $\mathrm{spect}(w(A)) = w(\mathrm{spect}(A))$.

Twierdzenie Cayleya-Hamiltona. *Jeśli funkcja w jest wielomianem charakterystycznym macierzy A, to macierz $w(A)$ jest zerowa.*

Twierdzenie. *Jeśli nieosobliwa macierz A ma parę własną $(\boldsymbol{x}, \lambda)$, to macierz A^{-1} ma parę własną $(\boldsymbol{x}, 1/\lambda)$.*

Twierdzenie. *Jeśli funkcje w, v są wielomianami i $v(A)$ jest macierzą nieosobliwą, to $w(A)\big(v(A)\big)^{-1} = \big(v(A)\big)^{-1} w(A)$. Jeśli $(\boldsymbol{x}, \lambda)$ jest parą własną macierzy A, to macierz $w(A)\big(v(A)\big)^{-1}$ ma parę własną $(\boldsymbol{x}, w(\lambda)/v(\lambda))$.*

Oprócz wielomianów i funkcji wymiernych, których argument i wartości są macierzami, możemy określać funkcje za pomocą szeregów, np.

$$e^A = \frac{1}{0!}I + \frac{1}{1!}A + \frac{1}{2!}A^2 + \frac{1}{3!}A^3 + \cdots.$$

W ogólności, jeśli macierz A ma parę własną $(\boldsymbol{x}, \lambda)$ i szereg określający $f(A)$ jest zbieżny, to macierz $f(A)$ ma parę własną $(\boldsymbol{x}, f(\lambda))$, a ponadto $\mathrm{spect}(f(A)) = f(\mathrm{spect}(A))$.

<u>Twierdzenie</u>. *Dwie macierze diagonalizowalne, A i B, komutują, tj. $AB = BA$ wtedy i tylko wtedy, gdy można je sprowadzić do postaci diagonalnej przez to samo podobieństwo (tj. gdy istnieje macierz X, taka że obie macierze, $X^{-1}AX$ i $X^{-1}BX$, są diagonalne).*

<u>Twierdzenie</u>. *Macierz ortogonalna (unitarna) X sprowadzająca rzeczywistą (zespoloną) macierz kwadratową A do postaci diagonalnej przez podobieństwo istnieje wtedy i tylko wtedy, gdy macierz A jest symetryczna (hermitowska).*

Macierz symetryczna (albo zespolona hermitowska) jest zatem diagonalizowalna i ma rzeczywiste wartości własne, przy czym jeśli wszystkie wartości własne mają krotność 1, to *każde* podobieństwo przekształcające ją na macierz diagonalną jest opisane przez macierz ortogonalną (w przypadku zespolonym — unitarną). Zauważmy, że przedstawienie macierzy symetrycznej w postaci iloczynu $A = X^{-1}\Lambda X$ z macierzą ortogonalną X i macierzą diagonalną Λ jest (z dokładnością do zwrotów wierszy macierzy X) rozkładem SVD.

<u>Twierdzenie</u>. *Jeśli rzeczywista (zespolona) macierz A jest symetryczna (hermitowska), to jej norma druga indukowana jest jej promieniem spektralnym.*

Wnioskiem z tego i jednego z poprzednich twierdzeń (którego?) jest stwierdzenie, że wskaźnik uwarunkowania macierzy *symetrycznej* (*hermitowskiej*) A w normie drugiej indukowanej wyraża się przez jej wartości własne w taki sposób:

$$\operatorname{cond}_2 A = \frac{\max_i |\lambda_i|}{\min_i |\lambda_i|}.$$

W wielu zastosowaniach pojawia się potrzeba rozwiązania algebraicznego zagadnienia własnego z macierzą symetryczną — jest to przypadek prostszy do numerycznego rozwiązywania niż przypadek ogólny i głównie na nim się dalej skupimy. Przed przedstawieniem algorytmów zbadajmy uwarunkowanie numeryczne zadania.

<u>Twierdzenie Bauera-Fikego</u>. *Niech A oznacza macierz kwadratową, dla której istnieje nieosobliwa macierz X, taka że macierz $\Lambda = X^{-1}AX$ jest diagonalna, z wartościami własnymi $\lambda_1, \ldots, \lambda_n$ macierzy A na diagonali. Jeśli liczba μ jest wartością własną macierzy zaburzonej $A + \delta A$, oraz liczba i jest taka, że $|\mu - \lambda_i| = \min_j |\mu - \lambda_j|$, to*

$$|\mu - \lambda_i| \leqslant \operatorname{cond}_2 X \|\delta A\|_2.$$

<u>Dowód</u>. Niech (u, μ) będzie parą własną macierzy $A + \delta A$. Wektor u jest kombinacją liniową kolumn $x_1, \ldots, x_n$ macierzy X, tj. wektorów własnych macierzy A, zatem istnieje wektor $y = [y_1, \ldots, y_n]^T \neq 0$, taki że $u = Xy$. Możemy przyjąć wektor y jednostkowy, tj. $\|y\|_2 = 1$. Mamy zatem

$$X^{-1}(A + \delta A)Xy = (X^{-1}AX + X^{-1}\delta AX)y = \mu y,$$

skąd wynika, że

$$\Lambda y + X^{-1}\delta AXy = \mu y, \quad \text{czyli} \quad X^{-1}\delta AXy = (\mu I - \Lambda)y.$$

Oznaczmy $z = (\mu I - \Lambda)y$. Możemy oszacować

$$\|z\|_2 \leqslant \|X^{-1}\|_2\|\delta A\|_2\|X\|_2\|y\|_2 = \operatorname{cond}_2 X\|\delta A\|_2.$$

Z drugiej strony

$$\|z\|_2^2 = \sum_{j=1}^{n} |\mu - \lambda_j|^2 y_j^2 \geqslant |\mu - \lambda_i|^2 \sum_{j=1}^{n} y_j^2 = |\mu - \lambda_i|^2.$$

Teza wynika z tych dwóch nierówności natychmiast. □

Jeśli macierz A jest symetryczna (w przypadku zespolonym hermitowska), to za X możemy przyjąć macierz ortogonalną (unitarną) i wtedy $\operatorname{cond}_2 X = 1$. Stąd zadanie znajdowania wartości własnych macierzy symetrycznych o największych wartościach bezwzględnych jest dobrze uwarunkowane. Dokładniej, wskaźnik uwarunkowania zadania obliczania wartości własnej λ_j macierzy symetrycznej jest równy $\max_i |\lambda_i|/|\lambda_j|$. Wskaźnik ten dla wartości własnej o największej wartości własnej jest równy 1, zaś dla wartości własnej o najmniejszej wartości bezwzględnej jest równy $\operatorname{cond}_2 A$.

Dla macierzy niesymetrycznej diagonalizowalnej nad ciałem $\mathbb{R}$ *żadna* macierz X zbudowana z wektorów własnych nie jest ortogonalna i dlatego $\operatorname{cond}_2 X > 1$. Jeśli natomiast macierz A nie jest diagonalizowalna, to zmiany wartości własnych zależą od powodujących je zaburzeń macierzy A w sposób ciągły, ale nie lipschitzowski. Numeryczne obliczanie wartości własnych takich macierzy jest kłopotliwe.

<u>Twierdzenie Wielandta-Hoffmana</u>. *Jeśli macierze A i $A + \delta A$ są symetryczne i wektory $\boldsymbol{\lambda}$ i $\boldsymbol{\mu}$ są zbudowane odpowiednio z tak samo (np. nierosnąco) uporządkowanych wartości własnych tych macierzy, to zachodzi nierówność*

$$\|\boldsymbol{\mu} - \boldsymbol{\lambda}\|_2 \leqslant \|\delta A\|_F.$$

<u>Dowód</u> pominiemy.

Zadanie wyznaczania całego widma macierzy symetrycznej jest zatem dobrze uwarunkowane, choć jeśli pewne wartości własne mają bardzo małe wartości bezwzględne, to ich zaburzenia *względne* spowodowane dodaniem małego zaburzenia δA do macierzy A mogą być duże.

Uwarunkowanie zadania wyznaczania wektorów własnych zależy od odległości między wartościami własnymi, i jest tym gorsze, *im mniej* odpowiednie wartości własne się różnią. Zauważmy, że jeśli pewna wartość własna ma krotność geometryczną $k > 1$, to istnieje nieskończenie wiele baz odpowiedniej podprzestrzeni własnej, złożonych z wektorów jednostkowych. Macierz zaburzona może mieć zamiast tej wartości własnej k różnych wartości własnych (jednokrotnych) i dlatego w tym przypadku rozwiązanie zależy od zaburzenia w sposób nieciągły (jest to możliwe nawet, jeśli macierz A jest symetryczna).

Uwaga: Nie jest dobrym pomysłem obliczanie współczynników wielomianu charakterystycznego $\det(A - \lambda I)$, np. w bazie potęgowej, a następnie znajdowanie jego miejsc zerowych. Nawet jeśli zadanie wyjściowe jest dobrze uwarunkowane, zadanie znalezienia miejsc zerowych wielomianu na podstawie jego współczynników jest zwykle *bardzo źle* uwarunkowane. Natomiast istnieją użyteczne metody numeryczne znajdowania wartości własnych przez rozwiązanie równania charakterystycznego. Wartości wielomianu charakterystycznego oblicza się na podstawie argumentu i współczynników macierzy, przy czym zwykle jest to pewna macierz podobna do danej macierzy A, dla której wartość wielomianu charakterystycznego można obliczyć mniejszym kosztem.

Metoda potęgowa

Będziemy się zajmować macierzami rzeczywistymi, o rzeczywistych wartościach własnych. Przypuśćmy, że jedna z wartości własnych macierzy A dominuje, tj. jej wartość bezwzględna jest większa niż wartości bezwzględne wszystkich pozostałych wartości własnych, i przypuśćmy, że mamy wyznaczyć parę własną z właśnie tą wartością własną. Założymy, że dominująca wartość własna jest liczbą rzeczywistą (możemy, jeśli macierz A jest symetryczna) i chwilowo przyjmiemy, że jej krotność jest równa 1. Przyjmijmy taką numerację, aby dominująca wartość własna była oznaczona symbolem λ_1.

Wybieramy niezerowy wektor $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$, a następnie dla $k > 0$ określamy wektory $\boldsymbol{x}^{(k)}$, wzorem $\boldsymbol{x}^{(k)} = A\boldsymbol{x}^{(k-1)}$ (czyli $\boldsymbol{x}^{(k)} = A^k\boldsymbol{x}^{(0)}$). Jeśli macierz A jest diagonalizowalna, to istnieją liczby $c_1, \dots, c_n$, takie że

$$x^{(0)} = \sum_{i=1}^{n} c_i x_i,$$

gdzie x_i to wektory własne macierzy A. Wtedy mamy

$$x^{(k)} = \sum_{i=1}^{n} c_i \lambda_i^k x_i = \lambda_1^k \sum_{i=1}^{n} c_i \left(\frac{\lambda_i}{\lambda_1}\right)^k x_i.$$

Jeśli $|\lambda_i| < |\lambda_1|$, to dla $k \to \infty$ ciąg liczb $(\lambda_i/\lambda_1)^k$ dąży do zera. To oznacza, że jeśli $c_1 \neq 0$, to ciąg kierunków wektorów $x^{(k)}$ dąży do kierunku wektora własnego x_1, przynależnego do dominującej wartości własnej. Po wykonaniu dostatecznie wielu iteracji możemy w ten sposób znaleźć wektor bliski wektora własnego x_1.

Podane rozumowanie jest podstawą metody potęgowej rozwiązywania algebraicznego zagadnienia własnego, a dokładniej wyznaczania pary własnej (x_1, λ_1) z dominującą wartością własną. Jeśli krotność geometryczna tej wartości własnej jest większa niż 1, to kierunki otrzymanego ciągu wektorów zbiegają do kierunku *pewnego* wektora własnego związanego z dominującą wartością własną. Opisane postępowanie jest jednak niepraktyczne, ponieważ jeśli $|\lambda_1| \neq 1$, to długości wektorów $x^{(k)}$ maleją do zera lub rosną nieograniczenie. Dlatego należy stosować normalizację, tj. dzielić kolejne otrzymane wektory przez ich długości — pamiętamy, że istotne są tylko kierunki tych wektorów. Mamy stąd algorytm:

1. Przyjmij $z^{(0)} \neq 0$,
2. Dla $k = 1, 2, \dots$ obliczaj

$$y^{(k)} = Az^{(k-1)}, \qquad z^{(k)} = \frac{1}{\|y^{(k)}\|_2} y^{(k)}.$$

Jeśli pewien wektor z jest wektorem własnym macierzy A, to spełnia równanie $Az = \lambda z$. Możemy je potraktować jak układ n równań z jedną niewiadomą, którą jest wartość własna λ; macierz tego układu jest kolumnowa, jest nią wektor z. Dla takiego układu stawiamy RLZNK. Układ równań normalnych ma postać

$$z^T z \lambda = z^T A z,$$

aby go rozwiązać, obliczamy tzw. iloraz Rayleigha

$$\lambda = \frac{z^T A z}{z^T z}.$$

Jeśli wektor z *nie jest* wektorem własnym, to oczywiście układ $Az = \lambda z$ jest sprzeczny, ale jeśli wektor z jest przybliżeniem wektora własnego x_i, to iloraz Rayleigha jest przybliżeniem wartości własnej λ_i. Ale jeśli $\|z\|_2 = 1$, to mianownik ilorazu Rayleigha jest równy 1. Zatem, po obliczeniu wektora $z^{(k)}$ obliczamy liczbę $\rho_{k-1} = z^{(k-1)T}y^{(k)}$. Podczas gdy ciąg wektorów jednostkowych $z^{(k)}$ zbiega do wektora własnego, ciąg liczb ρ_k zbiega do dominującej wartości własnej λ_1.

Jeśli macierz A jest symetryczna, λ_2 jest drugą co do wartości bezwzględnej wartością własną, i symbolem t_k oznaczymy tangens najmniejszego kąta między wektorem $z^{(k)}$ i wektorem x_1 należącym do podprzestrzeni własnej przynależnej do wartości własnej λ_1, to można udowodnić, że

$$|t_k| \leqslant \left|\frac{\lambda_2}{\lambda_1}\right|^k |t_0|, \quad \text{oraz} \quad |\rho_k - \lambda_1| \leqslant 2\|A\||t_k|^2 = O\Big(\left|\frac{\lambda_2}{\lambda_1}\right|^{2k}\Big).$$

Szybkość zbieżności zależy więc od tego, „jak bardzo dominuje” wartość własna λ_1. Zbieżność nie ma miejsca, jeśli dwie różne wartości własne dominują, tj. $\lambda_2 = -\lambda_1$. W takim przypadku „prosta” metoda potęgowa nie wystarczy do rozwiązania zadania.

Jeśli liczba c_1 dla przyjętego wektora $z^{(0)}$ jest zerem, to teoretycznie ciąg $(z^{(k)})_{k\in\mathbb{N}}$ zbiega do wektora własnego związanego z którąś z pozostałych wartości własnych. Ale w obliczeniach numerycznych występują błędy zaokrągleń, których skutki w tym przypadku *mogą być dobroczynne*: zaburzenie spowodowane zaokrągleniem zwykle doprowadza do pojawienia się odpowiedniej składowej o kierunku wektora własnego związanego z wartością własną λ_1, po czym kolejne iteracje „wzmacniają” tę składową, jednocześnie „wygaszając” pozostałe.

Odwrotna metoda potęgowa

Jeśli liczba λ jest wartością własną macierzy A, to dla dowolnego $a \notin \operatorname{spect} A$ liczba $1/(\lambda - a)$ jest wartością własną macierzy $(A - aI)^{-1}$. Zauważmy, że jeśli liczba a jest najbliżej wartości własnej λ_i macierzy A (tj. $|\lambda_i - a| < |\lambda_j - a|$ dla każdego $j \neq i$), to wartość własna $1/(\lambda_i - a)$ macierzy $(A - aI)^{-1}$ dominuje; co więcej, im lepsze przybliżenie a wartości własnej λ_i wybierzemy, tym szybsza jest zbieżność metody potęgowej zastosowanej do macierzy $(A - aI)^{-1}$.

Otrzymana na podstawie powyższego spostrzeżenia odwrotna metoda potęgowa, zwana też metodą Wielandta, umożliwia obliczenie dowolnej wartości własnej macierzy A (a nie tylko dominującej), a poza tym umożliwia otrzymanie szybkiej zbieżności. Algorytm jest taki:

1. Przyjmij parametr a. Oblicz macierz $B = A - aI$ i rozłóż ją (np. na czynniki trójkątne, za pomocą eliminacji Gaussa).
2. Przyjmij $z^{(0)} \neq 0$,
3. Dla $k = 1, 2, \ldots$ obliczaj

$$y^{(k)} = B^{-1}z^{(k-1)}, \quad \text{rozwiązując układ równań} \quad By^{(k)} = z^{(k-1)},$$
$$z^{(k)} = \frac{1}{\|y^{(k)}\|_2} y^{(k)}.$$

Ciąg wektorów $(z^{(k)})_{k\in\mathbb{N}}$ dąży do wektora własnego macierzy B^{-1}, który jest także wektorem własnym macierzy A. Po obliczeniu ilorazu Rayleigha $\rho_{k-1} = z^{(k-1)T}y^{(k)}$ można obliczyć przybliżenie wartości własnej najbliższej liczbie a, $\lambda_i \approx 1/\rho_{k-1} + a$.

Jeśli macierz A jest pełna, to koszt jej rozłożenia w kroku pierwszym jest rzędu n^3, zaś koszt rozwiązywania układu równań w każdej iteracji jest rzędu n^2, czyli taki sam jak koszt jednej iteracji zwykłej metody potęgowej. Koszt jednej iteracji można zmniejszyć (do rzędu n, jeśli macierz A jest symetryczna), dokonując wstępnego przekształcenia macierzy, co będzie opisane dalej.

Przybliżenie a' wartości własnej otrzymane na podstawie ilorazu Rayleigha po wykonaniu pewnej liczby iteracji umożliwia (znaczne) przyspieszenie zbieżności, kosztem ponownego rozkładania na czynniki macierzy $A - a'I$. Macierz ta jest źle uwarunkowana (tym gorzej, im lepszym przybliżeniem wartości własnej macierzy A jest liczba a'), ale ponieważ prawa strona rozwiązywanego układu równań jest przybliżeniem wektora własnego przynależnego do dominującej wartości własnej macierzy $(A - a'I)^{-1}$, okazuje się, że skutki błędów zaokrągleń nie są groźne dla dokładności obliczeń.

Sprowadzanie macierzy symetrycznej do postaci trójdiagonalnej

Wprawdzie (dla macierzy $n \times n$, gdzie $n > 4$) na ogół *nie można* w skończenie wielu krokach skonstruować macierzy X, takiej że macierz $\Lambda = X^{-1}AX$ jest diagonalna, ale dla macierzy symetrycznej *można* skonstruować macierz ortogonalną U, taką że macierz $T = U^{-1}AU$ jest trójdiagonalna. Koszt tego obliczenia jest (dla macierzy pełnej) rzędu n^3, ale można je wykonać jednorazowo, a następnie rozwiązać zagadnienie własne dla macierzy T; ma ona te same wartości własne, co macierz A, jeśli zaś wektor y jest wektorem własnym macierzy T, to wektor $x = Uy$ jest wektorem własnym macierzy A. Zarówno koszt obliczania iloczynu $y^{(k)} = Tz^{(k-1)}$, jak i koszt rozwiązywania układu równań $(T - aI)y^{(k)} = z^{(k-1)}$, jest rzędu n. Wstępne przekształcenie macierzy do postaci trójdiagonalnej jest też wstępnym krokiem wielu innych algorytmów rozwiązywania algebraicznego zagadnienia własnego.

Opiszemy algorytm Ortegi-Householdera. Otrzymana w nim macierz U jest iloczynem $n-2$ macierzy odbić Householdera; jak zwykle, nie wyznaczamy jej w postaci jawnej, tylko zapamiętujemy odpowiedni ciąg wektorów normalnych hiperpłaszczyzn odbić. Obliczenie polega na skonstruowaniu ciągu macierzy symetrycznych, $A^{(0)} = A, A^{(1)}, \ldots, A^{(n-2)} = T$. Współczynniki macierzy $A^{(k)}$ spełniają warunek $a_{ij}^{(k)} = a_{ji}^{(k)} = 0$ dla $j \leqslant k$ oraz $i > j+1$. Ponadto, jeśli $i < k$ lub $j < k$, to $a_{ij}^{(k)} = a_{ij}^{(k-1)}$.

Idea przekształcenia jest przedstawiona na schemacie:

$$
\underbrace{\begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ \bullet & \bullet & \bullet & \bullet & \bullet \end{bmatrix}}_{A^{(0)}}
\longrightarrow
\underbrace{\begin{bmatrix} \bullet & \bullet & \bullet & \bullet & \bullet \\ \circ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \end{bmatrix}}_{H_1A^{(0)}}
\longrightarrow
\underbrace{\begin{bmatrix} \bullet & \circ & & & \\ \bullet & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \circ & \circ \end{bmatrix}}_{H_1A^{(0)}H_1 = A^{(1)}}
$$

$$
\underbrace{\begin{bmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \\ & \bullet & \bullet & \bullet & \bullet \end{bmatrix}}_{A^{(1)}}
\longrightarrow
\underbrace{\begin{bmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \bullet & \bullet & \bullet \\ & \circ & \circ & \circ & \circ \\ & & \circ & \circ & \circ \\ & & \circ & \circ & \circ \end{bmatrix}}_{H_2A^{(1)}}
\longrightarrow
\underbrace{\begin{bmatrix} \bullet & \bullet & & & \\ \bullet & \bullet & \circ & & \\ & \bullet & \circ & \circ & \circ \\ & & \circ & \circ & \circ \\ & & \circ & \circ & \circ \end{bmatrix}}_{H_2A^{(1)}H_2 = A^{(2)}}
$$

W podanych wyżej schematach symbol „$\bullet$” oznacza oryginalny lub niezmieniony współczynnik macierzy, zaś „$\circ$” oznacza współczynnik, który wskutek odbicia uległ zmianie. Puste miejsca oznaczają (wytworzone lub zachowane) zera.

Pierwsza współrzędna wektora v_1, określającego odbicie reprezentowane przez macierz $H_1 = I - \gamma_1 v_1 v_1^T$, jest równa 0. Dla takiego odbicia macierze $A^{(0)}$ i $H_1A^{(0)}$ mają taki sam pierwszy wiersz. Odbicie konstruujemy w taki sposób, aby w pierwszej kolumnie macierzy $H_1A^{(0)}$ w wierszach $3, \ldots, n$ otrzymać zera. Mnożenie przez macierz odbicia z prawej strony zachowuje pierwszą kolumnę macierzy $H_1A^{(0)}$, w tym jej zerowe współczynniki. Wykonane przekształcenie $A^{(0)} \to A^{(1)}$ jest podobieństwem macierzy, ponieważ macierz H_1 jest symetryczna i ortogonalna. Ponadto przekształcenie to zachowuje symetrię, a zatem w pierwszym wierszu macierzy $A^{(1)}$, w kolumnach $3, \ldots, n$ też mamy zera.

Wektor $\boldsymbol{v}_2$ ma dwie pierwsze współrzędne równe zero, czego konsekwencją jest zachowanie pierwszego wiersza i pierwszej kolumny macierzy $A^{(1)}$.

Teraz implementacja. W k-tym kroku mamy obliczyć macierz

$$\begin{aligned} A^{(k)} &= H_k A^{(k-1)} H_k = (I - \gamma_k \boldsymbol{v}_k \boldsymbol{v}_k^T) A^{(k-1)} (I - \gamma_k \boldsymbol{v}_k \boldsymbol{v}_k^T) \\ &= A^{(k-1)} - \gamma_k \boldsymbol{v}_k \boldsymbol{v}_k^T A^{(k-1)} - \gamma_k A^{(k-1)} \boldsymbol{v}_k \boldsymbol{v}_k^T + \gamma_k^2 \boldsymbol{v}_k \boldsymbol{v}_k^T A^{(k-1)} \boldsymbol{v}_k \boldsymbol{v}_k^T. \end{aligned}$$

Oznaczmy $\boldsymbol{w}_k = \gamma_k A^{(k-1)} \boldsymbol{v}_k$. Wtedy

$$A^{(k)} = A^{(k-1)} - \boldsymbol{v}_k \boldsymbol{w}_k^T - \boldsymbol{w}_k \boldsymbol{v}_k^T + \boldsymbol{v}_k (\gamma_k \boldsymbol{v}_k^T \boldsymbol{w}_k) \boldsymbol{v}_k^T.$$

Niech $\boldsymbol{p}_k = \boldsymbol{w}_k - \boldsymbol{v}_k (\boldsymbol{v}_k^T \boldsymbol{w}_k) \gamma_k / 2$. Możemy sprawdzić, że

$$A^{(k)} = A^{(k-1)} - (\boldsymbol{v}_k \boldsymbol{p}_k^T + \boldsymbol{p}_k \boldsymbol{v}_k^T).$$

Właśnie tego wzoru używamy w obliczeniach. Zauważmy, że wektory $\boldsymbol{w}_k$ i $\boldsymbol{p}_k$ obliczone w k-tym kroku mają $k-1$ początkowych współrzędnych równych 0. Dzięki symetrii można obliczać tylko współczynniki na i pod (albo na i nad) diagonalą, dla zmniejszenia kosztu.

Algorytm QR

Niech A będzie macierzą symetryczną i niech Z_{k-1} będzie dowolną macierzą nieosobliwą $n \times n$. Przypuśćmy (na chwilę), że $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n| > 0$. Kolumny macierzy $Y_k = AZ_{k-1}$, zgodnie ze spostrzeżeniami, na których opiera się metoda potęgowa, mają „kierunki bliższe" kierunku wektora własnego $\boldsymbol{x}_1$, przynależnego do dominującej wartości własnej, λ_1. Ale gdybyśmy układ wektorów $\boldsymbol{y}_1^{(k)}, \ldots, \boldsymbol{y}_n^{(k)}$, tj. kolumn macierzy Y_k poddali ortonormalizacji Grama-Schmidta, to otrzymalibyśmy układ wektorów $\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_n^{(k)}$, z których każdy ma „kierunek bliższy" kierunku wektora przynależnego do kolejnej wartości własnej. Jest tak dlatego, bo ortonormalizacja „likwiduje" składowe wektora $\boldsymbol{y}_i^{(k)}$ w kierunkach wektorów $\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_{i-1}^{(k)}$, które są przybliżeniami wektorów własnych $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{i-1}$ macierzy A. Stąd wynika przypuszczenie, że dla każdego $i \in \{1, \ldots, n\}$ ciąg wektorów $(\boldsymbol{z}_i^{(k)})_{k \in \mathbb{N}}$ dąży do wektora własnego $\boldsymbol{x}_i$ przynależnego do wartości własnej λ_i.

Macierz $Z_k = [\boldsymbol{z}_1^{(k)}, \ldots, \boldsymbol{z}_n^{(k)}]$ jest ortogonalna, a ponadto istnieje macierz trójkątna górna R_k, taka że $Y_k = Z_k R_k$. Niech Z_0 będzie dowolną macierzą ortogonalną (np. jednostkową). Oznaczmy

$$A_k \stackrel{\text{def}}{=} Z_k^T A Z_k$$

(czyli w szczególności $A_0 = Z_0^T A Z_0$, ponadto wszystkie macierze A_k są podobne do A i symetryczne). Wtedy dla $k > 0$

$$A_{k-1} = Z_{k-1}^T A Z_{k-1} = Z_{k-1}^T Y_k = Z_{k-1}^T Z_k R_k = Q_k R_k,$$

gdzie $Q_k \stackrel{\text{def}}{=} Z_{k-1}^T Z_k$. Stąd $Z_k = Z_{k-1} Q_k$, a przez indukcję mamy stąd

$$Z_k = Z_0 Q_1 \ldots Q_k.$$

Na tej podstawie

$$A_k = Q_k^T \ldots Q_1^T Z_0^T A Z_0 Q_1 \ldots Q_k = Q_k^T A_{k-1} Q_k = R_k Q_k.$$

Podany wyżej rachunek jest podstawą dla następującego algorytmu:

1. Przyjmij $A_0 = Z_0^T A Z_0$,
2. Dla $k = 1, 2, \ldots$

   znajdź macierze ortogonalną Q_k i trójkątną górną R_k,

   takie że $A_{k-1} = Q_k R_k$,

   oblicz $A_k = R_k Q_k$.

Jeśli ciąg macierzy $(Z_k)_{k \in \mathbb{N}}$ zbiega do macierzy X, której kolumny są wektorami własnymi macierzy A, to ciąg macierzy $(A_k)_{k \in \mathbb{N}}$ zbiega do macierzy diagonalnej Λ, której znalezienie jest równoznaczne z obliczeniem wszystkich wartości własnych. Zbieżność może jednak nie mieć miejsca, jeśli nie wszystkie nierówności w ciągu $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots \geqslant |\lambda_n|$ są ostre (z tego samego powodu, dla którego metoda potęgowa może nie być zbieżna — wystarczy, że dwie wartości własne mają tę samą wartość bezwzględną i przeciwne znaki).

Zanim zajmiemy się zbieżnością, dokonajmy pewnego spostrzeżenia, które ma wpływ na koszt algorytmu. Okazuje się, że jeśli macierz A_{k-1} jest trójdiagonalna, to macierz A_k też jest taka. Zobaczmy schemat.

$$\underbrace{\begin{bmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & \bullet & & \\ & & \bullet & \bullet & \bullet & \\ & & & \bullet & \bullet & \bullet \\ & & & & \bullet & \bullet \end{bmatrix}}_{A_{k-1}} \rightarrow \underbrace{\begin{bmatrix} \circ & \circ & \circ & & & \\ & \circ & \circ & \circ & & \\ & & \circ & \circ & \circ & \\ & & & \circ & \circ & \circ \\ & & & & \circ & \circ \\ & & & & & \circ \end{bmatrix}}_{R_k = Q_k^T A_{k-1}} \rightarrow \underbrace{\begin{bmatrix} \circ & \circ & & & & \\ \circ & \circ & \circ & & & \\ & \circ & \circ & \circ & & \\ & & \circ & \circ & \circ & \\ & & & \circ & \circ & \circ \\ & & & & \circ & \circ \end{bmatrix}}_{A_k = R_k Q_k}$$

Wiersz i-ty macierzy R_k jest kombinacją liniową wierszy $1, \ldots, i+1$ macierzy A_{k-1}, dlatego na przecięciu kolumn $i+3, \ldots, n$ z tym wierszem są zerowe współczynniki. Z drugiej strony, i-ta kolumna macierzy A_k jest kombinacją liniową kolumn $1, \ldots, i+1$ macierzy trójkątnej górnej R_k, w związku z czym musi mieć zerowe współczynniki poniżej wiersza $i+1$. A że macierz A_k jest symetryczna, musi być też trójdiagonalna.

Pierwszym etapem obliczeń jest przekształcenie danej macierzy do postaci trójdiagonalnej (przy użyciu algorytmu Ortegi-Householdera), co kosztuje $O(n^3)$ działań i jest równoważne przyjęciu, że macierz Z_0 rozważana wyżej jest iloczynem macierzy wykonanych przy tym odbić: $Z_0 = H_1 \ldots H_{n-2}$. Rozkładanie macierzy trójdiagonalnej na czynniki Q_k i R_k, a następnie obliczanie A_k jest wykonywane kosztem $O(n)$ działań. Zamiast ortonormalizacji Grama-Schmidta (która zawiedzie, jeśli macierz A_{k-1} jest osobliwa), lepiej jest tu użyć innej metody; zwykle korzysta się z obrotów Givensa (zobacz zadania do poprzedniego wykładu; ta metoda nie zawiedzie). Można by też użyć odbić Householdera, ale do rozkładania macierzy trójdiagonalnej są one mniej wygodne.

Aby osiągnąć zbieżność i sprawić, by była jak najszybsza, w kolejnych iteracjach dobiera się parametr a_k (tzw. <u>przesunięcie</u>) i znajduje czynniki rozkładu macierzy $A_{k-1} - a_k I = Q_k R_k$, a nastęnie oblicza się macierz $A_k = R_k Q_k + a_k I$. Zauważmy, że

$$A_k = Q_k^T (A_{k-1} - a_k I) Q_k + a_k I = Q_k^T A_{k-1} Q_k,$$

a więc dla dowolnego przesunięcia macierze A_{k-1} i A_k są podobne. Mamy też $A_k = Z_k^T A Z_k$ oraz $Q_k = Z_{k-1}^T Z_k$, tak samo jak w przypadku bez przesunięć.

Jak działają przesunięcia? Otóż, gdyby przesunięcie a_k było równe pewnej wartości własnej λ, to wszystkie kolumny iloczynu $Y_k = (A - a_k I) Z_{k-1}$ byłyby prostopadłe do wszystkich wektorów własnych x przynależnych do tej wartości własnej (oczywiście macierz Y_k byłaby osobliwa). Przypuśćmy, że krotność wartości własnej λ jest równa 1 i wektory $y_1, \ldots, y_{n-1}$ (początkowe kolumny Y_k) są liniowo niezależne; jeśli macierz jest trójdiagonalna i jej współczynniki kodiagonalne są różne od zera, to to przypuszczenie jest słuszne (zobacz ćwiczenia 15 i 16). Otrzymane z nich metodą Grama-Schmidta wektory $z_1, \ldots, z_{n-1}$ są prostopadłe do x. Macierz ortogonalna Z_k, której to są początkowe kolumny, ma kolumnę z_n do nich prostopadłą, ale to znaczy, że ta kolumna ma kierunek wektora x, czyli *jest jednostkowym wektorem własnym* przynależnym do wartości własnej λ macierzy A. Łatwo jest sprawdzić (ćwiczenie), że wtedy współczynnik macierzy $A_k = Z_k^T A Z_k$ na ostatnim miejscu diagonali byłby równy λ, a pozostałe współczynniki w ostatnim wierszu i kolumnie byłyby równe 0.

Gdyby zatem było $a_k = \lambda$, to *w jednym kroku* dostalibyśmy macierz A_k ze współczynnikiem $a_{nn}^{(k)} = \lambda$. Jeśli przesunięcie a_k jest tylko przybliżeniem λ, a dokładniej, są spełnione nierówności $|a_k - \lambda| < |a_k - \lambda_i|$ dla każdej wartości własnej $\lambda_i \neq \lambda$, to ciąg współczynników $(a_{nn}^{(k)})_{k\in\mathbb{N}}$ będzie zbieżny do λ, tym szybciej, im lepiej parametr przesunięcia przybliża tę wartość własną. Aby zbieżność była jeszcze szybsza, w każdej iteracji wybiera się nowe przesunięcie.

Istnieją różne sposoby wybierania przesunięcia; jego wartość powinna przybliżać pewną wartość własną macierzy A. Najprostszy (i skuteczny) wybór to $a_k = a_{nn}^{(k-1)}$. Inny sposób (tzw. przesunięcie Wilkinsona) polega na przyjęciu parametru a_k równego jednej z wartości własnych bloku 2×2 wybranego z dwóch ostatnich wierszy i kolumn macierzy A_{k-1} (w tym celu trzeba rozwiązać równanie kwadratowe).

Współczynniki diagonalne kolejnych macierzy A_k dążą (z różnymi szybkościami) do wartości własnych, zaś współczynniki kodiagonalne (tj. sąsiadujące z diagonalą) dążą do zera. Na podstawie twierdzenia Gerszgorina można oszacować błędy przybliżenia wartości własnych przez współczynniki diagonalne macierzy A_k (choć oszacowanie to nie uwzględnia skutków błędów zaokrągleń). Dla odpowiednio dobranych przesunięć najszybciej zbiegają współczynniki w ostatnim wierszu i kolumnie. Jeśli wartość bezwzględna pewnego współczynnika na kodiagonali jest dostatecznie mała, tj. na poziomie błędów zaokrągleń, to współczynnik ten zastępuje się zerem, ale wtedy powstaje macierz blokowo-diagonalna z trójdiagonalnymi blokami:

$$\begin{bmatrix} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & \circ & & \\ & & \circ & \bullet & \bullet & \\ & & & \bullet & \bullet & \circ \\ & & & & \circ & \bullet \end{bmatrix} \to \left[\begin{array}{ccc|cc|c} \bullet & \bullet & & & & \\ \bullet & \bullet & \bullet & & & \\ & \bullet & \bullet & & & \\ \hline & & & \bullet & \bullet & \\ & & & \bullet & \bullet & \\ \hline & & & & & \bullet \end{array}\right]$$

i obliczenia można kontynuować dla tych bloków niezależnie, dobierając niezależnie przesunięcia. Przejście od zadania postawionego dla całej macierzy do zadań w mniejszych blokach nazywa się deflacją. Algorytm QR ze wstępnym przekształceniem do postaci trójdiagonalnej, przesunięciami i rekurencyjną deflacją jest najefektywniejszym znanym algorytmem znajdowania wszystkich wartości własnych macierzy symetrycznej.

Jeśli oprócz wartości własnych należy też znaleźć wektory własne, to wykonane przekształcenia ortogonalne trzeba zastosować do kolumn macierzy jednostkowej, aby jawnie wyznaczyć macierz Z_k, która jest przybliżeniem macierzy X.

Zadania i problemy

1. W przykładzie zastosowania mamy zadanie $Kx = \omega^2 Mx$, z macierzą diagonalną M. Znajdź sposób przekształcenia równania $Kx = \lambda Mx$ z macierzą M symetryczną i dodatnio określoną do postaci $Ay = \lambda y$ z symetryczną macierzą A. Uogólnij to przekształcenie na przypadek, gdy M jest dowolną macierzą symetryczną i dodatnio określoną.
2. Udowodnij twierdzenie Gerszgorina.

   Wskazówka do dowodu: Wybierz wektor własny x, spełniający warunek $\|x\|_\infty = 1$, i zbadaj koło Gerszgorina o środku a_{ii}, takie że $|x_i| = 1$.

   Wnioski: 1. Ponieważ macierz A^T ma to samo widmo co A, wartości własne leżą w przecięciu sum kół Gerszgorina obu tych macierzy.
   2. Macierz diagonalnie dominująca jest nieosobliwa.
3. Zbadaj, jak zmieni się widmo macierzy dwudiagonalnej $n \times n$

$$\begin{bmatrix} a & 1 & & \\ & a & \ddots & \\ & & \ddots & 1 \\ & & & a \end{bmatrix}$$

   po zastąpieniu współczynnika $a_{n1} = 0$ przez $\varepsilon \neq 0$.
4. Udowodnij twierdzenia o związkach zagadnienia własnego macierzy A z zagadnieniami własnymi dla macierzy $f(A)$, gdzie f jest wielomianem, funkcją wymierną lub funkcją określoną przez szereg potęgowy.
5. Udowodnij tezę (b) twierdzenia Schura. Skorzystaj z faktu, że zespolone wartości własne macierzy rzeczywistej występują w parach sprzężonych.
6. Udowodnij twierdzenie Cayleya-Hamiltona.
7. Udowodnij, że jeśli wartości własne λ_i, λ_j macierzy $A = A^T \in \mathbb{R}^{n,n}$ są różne, to przynależne do nich wektory własne są prostopadłe.
8. Udowodnij, że jeśli macierz A jest symetryczna i dowolna kolumna z_i macierzy ortogonalnej Z jest wektorem własnym macierzy A, to macierz $Z^T A Z$ ma i-ty współczynnik na diagonali równy wartości własnej λ, do której przynależy wektor własny z_i, a poza tym zera w i-tym wierszu i kolumnie.

   Wskazówka. Wygodnie jest tu użyć schematu Falka.
9. Niech $\rho(A)$ oznacza promień spektralny macierzy A, tj. największą wartość bezwzględną wartości własnej tej macierzy. Udowodnij, że norma druga indukowana macierzy A wyraża się wzorem

$$\|A\|_2 = \sqrt{\rho(A^T A)}.$$

10. Jeśli macierz A jest symetryczna, to jest też $\|A\|_2 = \rho(a)$. Korzystając z tego wzoru i podanych wiadomości o zagadnieniu własnym, wykaż, że dla macierzy symetrycznej $\operatorname{cond}_2 A = |\lambda_{\max}|/|\lambda_{\min}|$ (tu $\lambda_{\max}$ i $\lambda_{\min}$ to odpowiednio wartości własne o największej i najmniejszej wartości bezwzględnej).
11. Udowodnij, powołując się na twierdzenie Gerszgorina, nierówność norm indukowanych

$$\|A\|_2 \leqslant \sqrt{\|A^T A\|_\infty}.$$

12. Udowodnij, że dowolna macierz A i iloczyn $U^T A V$, z macierzami ortogonalnymi U i V, mają taką samą normę drugą indukowaną i normę Frobeniusa. Znajdź wyrażenia opisujące te normy w zależności od wartości szczególnych macierzy A i udowodnij nierówności

$$\|A\|_2 \leqslant \|A\|_F \leqslant \sqrt{r}\|A\|_2,$$

gdzie r oznacza rząd macierzy A.
13. Udwodnij, że jeśli macierz symetryczna jest dodatnio określona, to wszystkie jej wartości własne są dodatnie.
14. Niech macierz $A = X\Lambda X^{-1}$ będzie symetryczna nieokreślona i niech X będzie macierzą ortogonalną, a Λ — diagonalną. Jak, znając macierze X i Λ, można skonstruować rozdład SVD macierzy A?
15. Udowodnij, że jeśli macierz T, symetryczna i trójdiagonalna, ma współczynniki kodiagonalne (tj. sąsiadujące z diagonalnymi) niezerowe, to każda jej wartość własna ma krotność 1.
16. Udowodnij, że jeśli macierz T jest trójdiagonalna $n \times n$ i ma niezerowe współczynniki kodiagonalne, zaś macierz Z jest nieosobliwa, to pierwsze $n-1$ kolumn iloczynu TZ tworzy układ liniowo niezależny nawet wtedy, gdy macierz T jest osobliwa.
17. Znajdź wektor $\mathbf{p}_k$, obliczany w algorytmie Ortegi–Householdera (sprowadzania macierzy do postaci trójdiagonalnej), jeśli macierz $A^{(k-1)}$ jest jednostkowa. Jaka będzie wtedy macierz $A^{(k)}$?
18. Dane są wektory $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, takie że $\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1$ oraz $\boldsymbol{v}_2^T\boldsymbol{v}_1 \neq 0$. Niech $A = H_1 + H_2$, gdzie $H_1 = I - 2\boldsymbol{v}_1\boldsymbol{v}_1^T$, $H_2 = I - 2\boldsymbol{v}_2\boldsymbol{v}_2^T$. Udowodnij, że układ równań liniowych $A\boldsymbol{x} = \mathbf{b}$, dla dowolnego $\mathbf{b} \in \mathbb{R}^n$ ma jednoznaczne rozwiązanie i podaj algorytm rozwiązujący ten układ kosztem rzędu n działań.

    Wskazówka. Wykaż, że wektory $\boldsymbol{w}_1 = \boldsymbol{v}_1 + \boldsymbol{v}_2$ i $\boldsymbol{w}_2 = \boldsymbol{v}_1 - \boldsymbol{v}_2$ są wektorami własnymi macierzy A i znajdź jej wszystkie wartości własne.
19. Dane są wektory $\boldsymbol{v}_1, \boldsymbol{v}_2 \in \mathbb{R}^n$, takie że $\|\boldsymbol{v}_1\|_2 = \|\boldsymbol{v}_2\|_2 = 1$ oraz $|\boldsymbol{v}_2^T\boldsymbol{v}_1| \neq \sqrt{2}/2$. Niech $A = H_1H_2 + H_2H_1$, gdzie $H_1 = I - 2\boldsymbol{v}_1\boldsymbol{v}_1^T$, $H_2 = I - 2\boldsymbol{v}_2\boldsymbol{v}_2^T$. Udowodnij, że układ

równań liniowych $Ax = \mathbf{b}$, dla dowolnego $\mathbf{b} \in \mathbb{R}^n$ ma jednoznaczne rozwiązanie i podaj algorytm rozwiązujący ten układ kosztem rzędu n działań.

Wskazówka. Wykaż, że wektory $\boldsymbol{v}_1$ i $\boldsymbol{v}_2$ są wektorami własnymi macierzy A i znajdź jej wszystkie wartości własne.

20. (Lab.) Niech

$$A = \begin{bmatrix} 2 & 2 & 0.5 \\ 2 & 2 & 0.25 \\ 0.5 & 0.25 & 1 \end{bmatrix}.$$

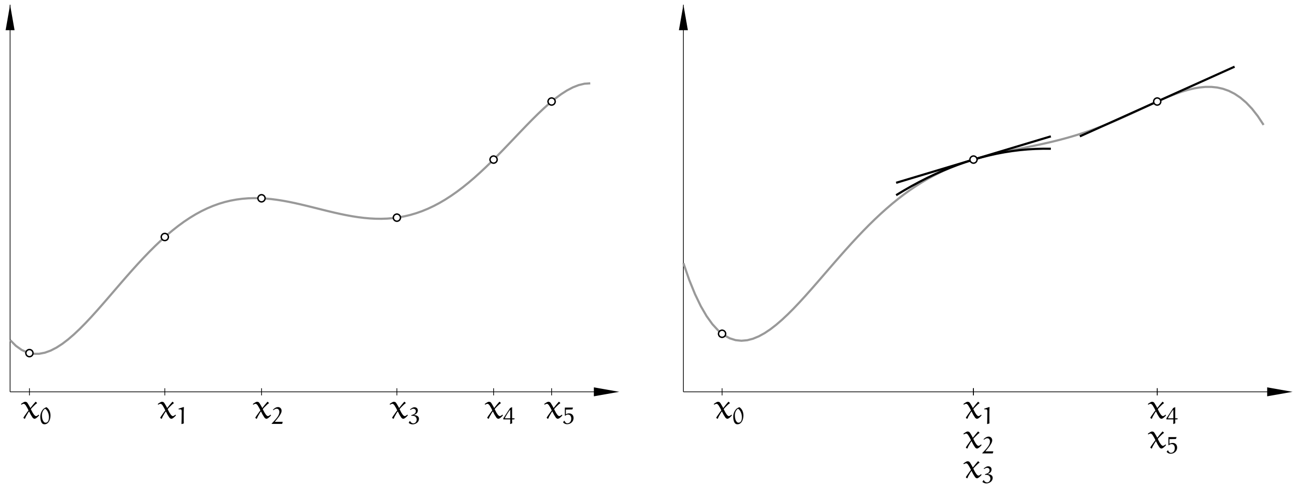
Korzystając z pakietu Octave, zastosuj metodę potęgową do znalezienia największej wartości własnej i odpowiedniego wektora własnego macierzy A. Wykonaj dwa eksperymenty: z wektorem $\boldsymbol{z}^{(0)} = [1, 1, 1]^T$ oraz $\boldsymbol{z}^{(0)} = [0.691678170958549, -0.702035294416588, 0]^T$. Obejrzyj wyniki 30 iteracji metody i wyjaśnij, co się tam dzieje.

Interpolacja wielomianowa

Zadania interpolacyjne Lagrange'a i Hermite'a

Niech $x_0, \ldots, x_n$ będą danymi liczbami, z których każde dwie są różne, i niech $y_0, \ldots, y_n$ będą liczbami dowolnymi. Zadanie interpolacyjne Lagrange'a polega na skonstruowaniu (tj. obliczeniu współczynników w pewnej ustalonej bazie) wielomianu $h(x)$ stopnia co najwyżej n, takiego że $h(x_i) = y_i$ dla $i = 0, \ldots, n$.

Wymaganie, aby liczby x_i, zwane węzłami interpolacyjnymi, były parami różne, jest oczywiste; nie można zadawać dwóch różnych wartości funkcji w tym samym punkcie, a dwukrotne podanie tej samej wartości mija się z celem. Ale możemy dopuścić, aby węzły powtarzały się, jeśli dla każdego dodatkowego „egzemplarza" węzła określimy inny warunek interpolacyjny. Jeśli warunek ten polega na podaniu wartości pochodnej kolejnego rzędu, to mamy ogólniejsze zadanie interpolacyjne Hermite'a: dla każdego węzła określamy jego krotność — jest to liczba jego wystąpień w danym ciągu węzłów. Dla węzła x_i o krotności $r > 1$ zadajemy wartość funkcji, $h(x_i)$, pochodnej, $h'(x_i)$, i pochodnych do rzędu $r-1$ włącznie.

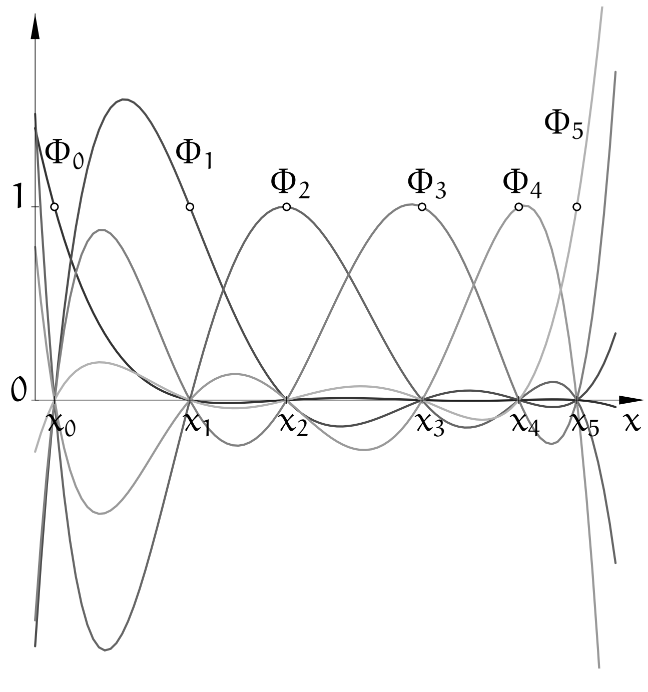


Zadanie interpolacyjne Hermite'a i jego przypadek szczególny — zadanie interpolacyjne Lagrange'a — ma jednoznaczne rozwiązanie. Jeśli poszukiwany wielomian przedstawimy jako kombinację liniową elementów dowolnej bazy przestrzeni $\mathbb{R}[x]_n$ (przestrzeni wielomianów stopnia co najwyżej n), to możemy warunki interpolacyjne zapisać w postaci układu równań liniowych, z niewiadomymi współczynnikami w wybranej bazie. Wymiar przestrzeni, czyli liczba niewiadomych, jest równy $n+1$, tj. taki sam jak liczba równań. Przypuśćmy, że warunki interpolacyjne są jednorodne, tj. wszystkie zadane wartości funkcji i pochodnych są równe 0. Wtedy rozwiązaniem układu jest wektor zerowy, który reprezentuje wielomian zerowy. Gdyby istniał niezerowy wielomian $h(x)$ stopnia co najwyżej n spełniający te same warunki interpolacyjne, to musiałby być podzielny przez wielomian $p_{n+1}(x) = (x - x_0) \cdot \ldots \cdot (x - x_n)$, ale to oznacza, że stopień wielomianu h musiałby być co najmniej $n + 1$. Jednoznaczność rozwiązania układu równań opisującego jednorodne warunki interpolacyjne oznacza, że macierz tego układu jest nieosobliwa, a więc dla dowolnej prawej strony (tj. dowolnych zadanych wartości funkcji i pochodnych) zadanie ma jednoznaczne rozwiązanie.

Rozwiązanie zadania interpolacyjnego Lagrange'a można przedstawić wzorem

$$h(x) = \sum_{i=0}^{n} y_i \Phi_i(x), \quad \text{gdzie} \quad \Phi_i(x) = \prod_{j \in \{0,\ldots,n\} \setminus \{i\}} \frac{x - x_j}{x_i - x_j}.$$

Dane liczby $y_0, \ldots, y_n$ są współczynnikami wielomianu h w bazie $\{\Phi_0, \ldots, \Phi_n\}$, ale wzór ten nie jest praktyczny w obliczeniach numerycznych (należy go raczej traktować jako dowód istnienia rozwiązania zadania, czasem przydaje się też w rachunkach symbolicznych i w rozważaniach teoretycznych).

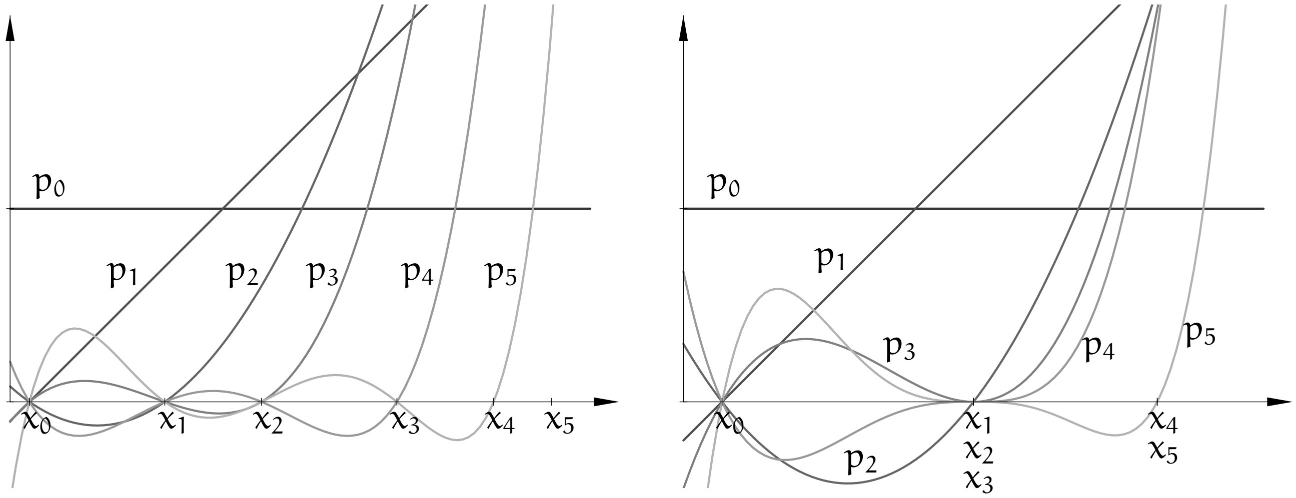


Wykresy wielomianów $\Phi_0, \ldots, \Phi_n$ dla przykładowego ciągu węzłów (z $n = 5$) są pokazane na rysunku. Warto zauważyć, że niektóre z tych wielomianów przyjmują między węzłami wartości bezwzględne sporo większe niż 1. Maksymalne wartości bezwzględne wielomianów bazowych Lagrange'a między węzłami zależą od liczby węzłów i od ich rozmieszczenia, i jeśli stopień jest duży, to mogą być bardzo duże. W konsekwencji rozwiązanie zadania interpolacji może przyjmować między węzłami interpolacyjnymi wartości leżące daleko poza przedziałem, w którym leżą dane wartości funkcji w tych węzłach.

Bazy Newtona

Niech $x_0, \dots, x_n$ będą liczbami danymi. Możemy określić wielomiany

$$\begin{aligned}
p_0(x) &= 1, \\
p_1(x) &= x - x_0, \\
p_2(x) &= (x - x_0)(x - x_1), \\
&\vdots \\
p_n(x) &= (x - x_0) \cdot \ldots \cdot (x - x_{n-1}), \\
p_{n+1}(x) &= (x - x_0) \cdot \ldots \cdot (x - x_{n-1})(x - x_n).
\end{aligned}$$



Zbiór wielomianów $\{p_0, \dots, p_k\}$ jest bazą przestrzeni $\mathbb{R}[x]_k$, której elementami są wszystkie wielomiany stopnia co najwyżej k. Ta tzw. baza Newtona, określona za pomocą danych węzłów, jest wygodniejsza od bazy potęgowej w zastosowaniu do zadań interpolacji wielomianowej[15]. W szczególności, mając współczynniki $b_0, \dots, b_n$ wielomianu stopnia co najwyżej n, możemy obliczyć wartość wielomianu $w(x) = \sum_{i=0}^{n} b_i p_i(x)$ za pomocą odpowiednio uogólnionego schematu Hornera:

$w = b_n$;
for ($i = n-1$; $i \geqslant 0$; i--)
  $w = w*(x - x_i)$ + b_i;

Aby rozwiązać zadanie interpolacyjne Lagrange'a, możemy dla wybranej bazy $\{f_0, \dots, f_n\}$ przestrzeni $\mathbb{R}[x]_n$ utworzyć macierz $A \in \mathbb{R}^{n+1 \times n+1}$, taką że jej współczynnik $a_{ij} = f_j(x_i)$ (numerujemy tu wiersze i kolumny od 0 do n). Rozwiązanie zadania sprowadza się do rozwiązania układu równań z tą macierzą.

[15]Baza potęgowa jest szczególnym przypadkiem bazy Newtona, dla $x_0 = \cdots = x_{k-1} = 0$.

Dla bazy potęgowej mamy układ równań z macierzą pełną

$$\begin{bmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \dots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix},$$

którego rozwiązaniem jest wektor współczynników wielomianu $h(x) = \sum_{k=0}^{n} a_k x^k$, zaś dla bazy Newtona określonej za pomocą węzłów interpolacyjnych mamy układ z macierzą trójkątną dolną:

$$\begin{bmatrix} 1 & 0 & \dots & 0 \\ 1 & p_1(x_1) & \ddots & \vdots \\ \vdots & \vdots & \ddots & 0 \\ 1 & p_1(x_n) & \dots & p_n(x_n) \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_n \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix}.$$

Możemy obliczyć współczynniki tej macierzy i rozwiązać układ kosztem tylko $\Theta(n^2)$ operacji (dalej poznamy inny algorytm obliczania współczynników wielomianu interpolacyjnego w bazie Newtona). W razie potrzeby, możemy następnie kosztem $\Theta(n^2)$ operacji przejść do bazy potęgowej, ale jeśli nie jest to konieczne, to nie warto tego robić.

Różnice dzielone

Niech f oznacza pewną funkcję określoną w przedziale $A \subset \mathbb{R} \to \mathbb{R}$. Dla ustalonych liczb $x_i \in A$ (węzłów interpolacyjnych) określamy <u>różnice dzielone rzędu 0</u>:

$$f[x_i] \stackrel{\text{def}}{=} f(x_i).$$

Zakładajac, że węzły są jednokrotne (czyli parami różne), możemy następnie określić dla $k > 0$ <u>różnice dzielone rzędu k</u> wzorem

$$f[x_i, \dots, x_{i+k}] \stackrel{\text{def}}{=} \frac{f[x_i, \dots, x_{i+k-1}] - f[x_{i+1}, \dots, x_{i+k}]}{x_i - x_{i+k}}. \qquad (*)$$

Różnicę dzieloną można postrzegać na dwa sposoby:

1. Dla ustalonej funkcji f jest to funkcja $k+1$ zmiennych. Funkcja ta jest symetryczna, tj. dowolne przestawienie jej argumentów (węzłów) nie zmienia jej wartości.
2. Dla ustalonych węzłów $x_i, \dots, x_{i+k}$ jest to kombinacja liniowa wartości funkcji f w tych węzłach, a zatem jest to funkcjonał liniowy w przestrzeni funkcji o ustalonej dziedzinie A, do której należą węzły.

Obie te własności są widoczne we wzorze podanym niżej.

Twierdzenie. *Jeśli wszystkie węzły są jednokrotne, to*

$$f[x_i,\dots,x_{i+k}] = \sum_{j=i}^{i+k} c_{ij}^{(k)} f(x_j), \quad \text{gdzie} \quad c_{ij}^{(k)} = \prod_{l\in\{i,\dots,i+k\}\setminus\{j\}} \frac{1}{x_j - x_l}. \tag{**}$$

Dowód. Dla $k = 0$ każdy współczynnik $c_{ij}^{(k)}$ (czyli jedyny współczynnik $c_{ii}^{(0)}$) jest neutralnym elementem mnożenia, $c_{ii}^{(0)} = 1$. Dla $k = 1$ możemy sprawdzić wzór (**) bezpośrednio. Przyjmijmy założenie indukcyjne, że wzory (**) i (*) są równoważne dla różnic dzielonych rzędu $0,\dots,k$ gdzie $k \geqslant 1$. Wtedy możemy różnicę dzieloną rzędu $k + 1$ obliczyć tak:

$$\begin{aligned} f[x_i,\dots,x_{i+k+1}] &= \frac{\sum_{j=i}^{i+k} c_{ij}^{(k)} f(x_j) - \sum_{j=i+1}^{i+k+1} c_{i+1,j}^{(k)} f(x_j)}{x_i - x_{i+k+1}} \\ &= \frac{c_{ii}^{(k)}}{x_i - x_{i+k+1}} f(x_i) + \sum_{j=i+1}^{i+k} \frac{c_{ij}^{(k)} - c_{i+1,j}^{(k)}}{x_i - x_{i+k+1}} f(x_j) - \frac{c_{i+1,i+k+1}^{(k)}}{x_i - x_{i+k+1}} f(x_{i+k+1}). \end{aligned}$$

Możemy sprawdzić, że

$$c_{ii}^{(k+1)} = \frac{c_{ii}^{(k)}}{x_i - x_{i+k+1}} = \prod_{l\in\{i+1,\dots,i+k+1\}} \frac{1}{x_i - x_l},$$

$$c_{ij}^{(k+1)} = \frac{c_{ij}^{(k)} - c_{i+1,j}^{(k)}}{x_i - x_{i+k+1}} =$$

$$\frac{1}{x_i - x_{i+k+1}} \left(\frac{x_j - x_{i+k+1}}{x_j - x_{i+k+1}} \prod_{l\in\{i,\dots,i+k\}\setminus\{j\}} \frac{1}{x_j - x_l} - \frac{x_j - x_i}{x_j - x_i} \prod_{l\in\{i+1,\dots,i+k+1\}\setminus\{j\}} \frac{1}{x_j - x_l} \right)$$

$$= \frac{x_j - x_{i+k+1} - (x_j - x_i)}{x_i - x_{i+k+1}} \prod_{l\in\{i,\dots,i+k+1\}\setminus\{j\}} \frac{1}{x_j - x_l} \qquad \text{dla } j \in \{i+1,\dots,i+k\},$$

$$c_{i,i+k+1}^{(k+1)} = \frac{-c_{i+1,i+k+1}^{(k)}}{x_i - x_{i+k+1}} = \prod_{l\in\{i,\dots,i+k\}} \frac{1}{x_{i+k+1} - x_l},$$

czyli w każdym przypadku otrzymaliśmy wzór (**) z k zastąpionym przez $k+1$. □

Twierdzenie. *Jeśli funkcja* f *ma w przedziale* A*, do którego należą węzły* $x_i,\dots,x_{i+k}$*, ciągłą pochodną rzędu* $k \geqslant 1$*, to*

$$f[x_i,\dots,x_{i+k}] = \int_{S_k}\!\!\dots\!\int f^{(k)}(t_0 x_i + t_1 x_{i+1} + \dots + t_k x_{i+k})\, dt_1 \dots dt_k, \tag{***}$$

gdzie $S_k = \{\,(t_0,\dots,t_k)\colon \sum_{j=0}^{k} t_j = 1, t_0,\dots,t_k \geqslant 0\,\}$.

<u>Dowód</u>. Sympleks S_1 jest odcinkiem, zatem dla $k = 1$ możemy bezpośrednio obliczyć

$$\int_0^1 f'\big((1-t_1)x_i + t_1 x_{i+1}\big)\,dt_1 = \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} = f[x_i, x_{i+1}].$$

Niech $k > 1$. Wprowadzamy zmienne pomocnicze

$$s_1 = 1 - t_1,\ s_2 = s_1 - t_2,\ \ldots,\ s_{k-1} = s_{k-2} - t_{k-1},\ s_k = s_{k-1} - t_k.$$

$$\begin{array}{ccccccccccccc} & t_0 & & t_k & & & \ldots & & t_2 & & t_1 & & \\ \hline 0 & & s_k & & s_{k-1} & & \ldots & & s_2 & & s_1 & & 1 \end{array}$$

Możemy zauważyć, że dla każdego j jest $s_j = 1 - \sum_{i=1}^{j} t_i = t_0 + \sum_{i=j+1}^{k} t_i$ (w szczególności $s_k = t_0$), oraz

$$\begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_k \end{bmatrix} = \begin{bmatrix} -1 & & & \\ 1 & -1 & & \\ & \ddots & \ddots & \\ & & 1 & -1 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \\ \vdots \\ s_k \end{bmatrix} + \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{$\oplus$}$$

a ponadto $(t_0, \ldots, t_k) \in S_k$ wtedy i tylko wtedy, gdy

$$0 \leqslant s_k \leqslant s_{k-1} \leqslant \cdots \leqslant s_2 \leqslant s_1 \leqslant 1.$$

Dla ustalonych liczb $s_1, \ldots, s_{k-1}$ oznaczmy

$$\begin{aligned} x(s_k) &= x_{i+1} + s_1(x_{i+2} - x_{i+1}) + \cdots + s_{k-1}(x_{i+k} - x_{i+k-1}) + s_k(x_i - x_{i+k}) \\ &= t_0 x_i + \cdots + t_k x_{i+k}. \end{aligned}$$

Pochodna (względem s_k) funkcji $g(s_k) \stackrel{\text{def}}{=} f^{(k-1)}\big(x(s_k)\big)$ jest równa

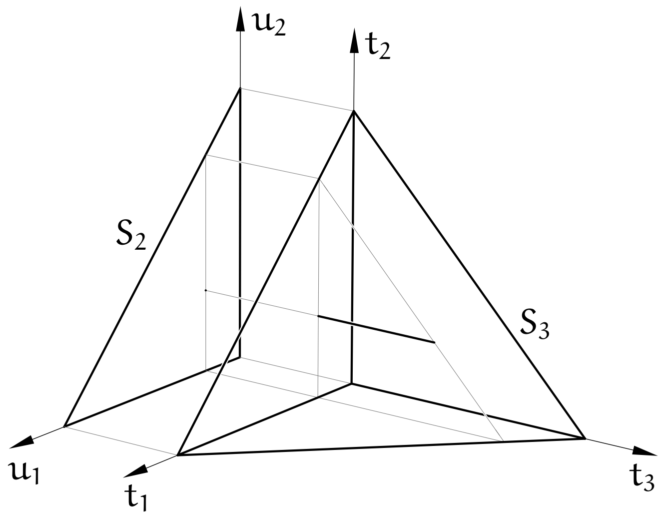
$$g'(s_k) = (x_i - x_{i+k}) f^{(k)}\big(x(s_k)\big). \tag{$\otimes$}$$

Całkę po prawej stronie wzoru (⁑) oznaczmy symbolem I. Jakobian przejścia od zmiennych $t_1, \ldots, t_k$ do $s_1, \ldots, s_k$ (tj. wartość bezwzględna wyznacznika macierzy we wzorze ($\oplus$)) jest równy 1. Dzięki temu możemy obliczyć

$$\begin{aligned} I = \int \ldots \int_{S_k} f^{(k)}(t_0 x_i + \cdots + t_k x_{i+k})\,dt_1 \ldots dt_k = \\ \int_0^1 \ldots \int_0^{s_{k-2}} \int_0^{s_{k-1}} f^{(k)}\big(x(s_k)\big)\,ds_k\,ds_{k-1} \ldots ds_1. \end{aligned}$$

Ponieważ $s_k = t_0$, $s_{k-1} = t_0 + t_k$, a ponadto liczby $t_1, \ldots, t_{k-1}$ nie zależą od s_k, na podstawie ($\otimes$) „wewnętrzna” całka jest równa

$$\int_0^{s_{k-1}} \frac{g'(s_k)}{x_i - x_{i+k}}\, ds_k = \frac{1}{x_i - x_{i+k}} \Big(f^{(k-1)}\big(x(s_{k-1})\big) - f^{(k-1)}\big(x(0)\big)\Big)$$
$$= \frac{1}{x_i - x_{i+k}} \big(f^{(k-1)}(t_0 x_i + \cdots + t_{k-1} x_{i+k-1}) - f^{(k-1)}(t_1 x_{i+1} + \cdots + t_k x_{i+k})\big).$$



Obszarem całkowania dla „zewnętrznych” całek jest zbiór

$$S_{k-1} = \{\, (u_0, \ldots, u_{k-1}) \colon\ \textstyle\sum_{j=0}^{k-1} u_j = 1,\ u_0, \ldots, u_{k-1} \geqslant 0 \},$$

gdzie $u_0 = t_0 + t_k$ oraz $u_j = t_j$ dla $j = 1, \ldots, k-1$. Jeśli przyjmiemy założenie indukcyjne, że

$$f[x_i, \ldots, x_{i+k-1}] = \int \underset{S_{k-1}}{\ldots} \int f^{(k-1)}(u_0 x_i + \cdots + u_{k-1} x_{i+k-1})\, du_1 \ldots du_{k-1}$$

i obliczymy te całki korzystając z niego, to dostaniemy

$$I = \frac{f[x_i, \ldots, x_{i+k-1}] - f[x_{i+1}, \ldots, x_{i+k}]}{x_i - x_{i+k}} = f[x_i, \ldots, x_{i+k}]. \quad \square$$

Wzór ($\overset{**}{*}$) ma nazwę <u>wzoru Hermite’a-Genocchiego</u>[16]. Możemy zauważyć, że funkcja podcałkowa, a więc także całka, zależy w sposób ciągły od liczb $x_i, \ldots, x_{i+k}$ także wtedy, gdy węzły te „sklejają się”, tzn. wtedy, gdy zmieniając węzły, doprowadzamy do pojawienia się węzłów krotnych.

[16]Wzór ten obowiązuje także dla funkcji zespolonych zmiennej zespolonej. Jeśli punkty $x_i, \ldots, x_{i+k}$ nie leżą na jednej prostej w płaszczyźnie zespolonej, to funkcja f musi być analityczna w otoczce wypukłej tych punktów — odpowiada ona przedziałowi A rozpatrywanemu w przypadku rzeczywistym.

<u>Wniosek</u>. *Jeśli funkcja* f *jest klasy* C^k *w otoczeniu punktu* x_i*, to*[17]

$$\lim_{x_{i+1},\dots,x_{i+k}\to x_i} f[x_i,\dots,x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

<u>Dowód</u>. Miara (objętość k-wymiarowa) zbioru całkowania S_k, który jest sympleksem k-wymiarowym, jest równa

$$|S_k| = \int\limits_{S_k}\!\!\dots\!\int 1\,dt_1\dots dt_k = \frac{1}{k!}.$$

Funkcja podcałkowa w ($\overset{*}{*}\!{*}$) jest ciągła, dlatego na podstawie twierdzenia o wartości średniej istnieje punkt $\xi = \sum_{j=0}^{k} t_j x_{i+j}$, gdzie $(t_0,\dots,t_k)\in S_k$, taki że całka we wzorze ($\overset{*}{*}\!{*}$) jest równa $f^{(k)}(\xi)|S_k| = f^{(k)}(\xi)/k!$. Liczba ξ, będąca kombinacją wypukłą liczb $x_i,\dots,x_{i+k}$, leży w najkrótszym przedziale zawierającym te liczby. Wystarczy zatem dokonać odpowiedniego przejścia granicznego. □

Dalej jeszcze raz wykażemy ten fakt, za pomocą wzoru na resztę interpolacyjną. Na jego podstawie możemy zdefiniować różnicę dzieloną rzędu $k \geqslant 1$ w przypadku, gdy $x_i = \dots = x_{i+k}$, wzorem

$$f[\underbrace{x_i,\dots,x_i}_{k+1}] \overset{\text{def}}{=} \frac{f^{(k)}(x_i)}{k!}, \qquad (\ast\ast\ast\ast)$$

natomiast w przypadku, gdy pewne węzły mają krotność większą niż 1, ale nie wszystkie węzły są jednakowe, możemy (dzięki symetrii) uporządkować je tak, aby było $x_i \neq x_{i+k}$, i użyć wzoru (*). Na przykład, jeśli $x_0 = x_1 = x_2 \neq x_3$ i funkcja f jest dwukrotnie różniczkowalna w otoczeniu x_0, to

$$\begin{aligned}
f[x_0,x_0,x_3] &= \frac{f'(x_0)-f[x_0,x_3]}{x_0-x_3} = \frac{f'(x_0)}{x_0-x_3} - \frac{f(x_0)}{(x_0-x_3)^2} + \frac{f(x_3)}{(x_0-x_3)^2},\\
f[x_0,x_0,x_0,x_3] &= \frac{f''(x_0)/2 - f[x_0,x_0,x_3]}{x_0-x_3}\\
&= \frac{f''(x_0)}{2(x_0-x_3)} - \frac{f'(x_0)}{(x_0-x_3)^2} + \frac{f(x_0)}{(x_0-x_3)^3} - \frac{f(x_3)}{(x_0-x_3)^3}.
\end{aligned}$$

W przypadku ogólnym różnica dzielona rzędu k, $f[x_i,\dots,x_{i+k}]$, jest kombinacją liniową wartości funkcji f i jej pochodnych w węzłach, przy czym jeśli pewien węzeł ma krotność r, to kombinacja obejmuje pochodne funkcji f w tym węźle do rzędu $r-1$.

[17]Szczególny przypadek tego wzoru, dla $k = 1$, służy do *zdefiniowania* pochodnej funkcji.

Algorytm różnic dzielonych

Przypuśćmy, że węzły $x_0, \ldots, x_n$ są parami różne. Obliczmy różnicę dzieloną wielomianu $p_k(x)$ należącego do bazy Newtona określonej dla tych węzłów:

$$p_k[x, x_0] = \frac{(x - x_0) \cdot \ldots \cdot (x - x_{k-1}) - (x_0 - x_0) \cdot \ldots \cdot (x_0 - x_{k-1})}{x - x_0} = \\ (x - x_1) \cdot \ldots \cdot (x - x_{k-1}).$$

Otrzymaliśmy wielomian stopnia $k - 1$. Obliczając różnice dzielone coraz wyższych rzędów, dostaniemy wielomiany coraz niższych stopni:

$$\begin{aligned} p_k[x, x_0, x_1] &= (x - x_2) \cdot \ldots \cdot (x - x_{k-1}), \\ &\vdots \\ p_k[x, x_0, \ldots, x_{k-2}] &= (x - x_{k-1}), \\ p_k[x, x_0, \ldots, x_{k-2}, x_{k-1}] &= 1. \end{aligned}$$

Po ostatnim kroku możemy oczywiście podstawić $x = x_k$, co nie zmieni wartości otrzymanego wielomianu stopnia 0. Różnice dzielone rzędów wyższych niż k są równe 0. Biorąc pod uwagę zbiór miejsc zerowych wielomianu p_k, mamy

$$p_k[x_0, \ldots, x_i] = \begin{cases} 0 & \text{dla } i \neq k, \\ 1 & \text{dla } i = k. \end{cases}$$

Podany wyżej rachunek „przechodzi” też na przypadek węzłów powtarzających się. Tak więc różnice dzielone $\cdot[x_0], \cdot[x_0, x_1], \ldots, \cdot[x_0, \ldots, x_n]$ — funkcjonały liniowe na przestrzeni $\mathbb{R}[x]_n$ — tworzą bazę sprzężoną do bazy $\{p_0, \ldots, p_n\}$ przestrzeni $\mathbb{R}[x]_n$.

Niech $h(x)$ będzie rozwiązaniem zadania interpolacyjnego Lagrange'a dla węzłów $x_0, \ldots, x_n$. Wielomian h możemy przedstawić jako kombinację liniową wielomianów $p_0, \ldots, p_n$: $h(x) = \sum_{k=0}^{n} b_k p_k(x)$. Z tego, że określone dla tych węzłów różnice dzielone tworzą bazę sprzężoną do bazy Newtona określonej dla tego samego ciągu węzłów wynika, że

$$h[x_0, \ldots, x_i] = \sum_{k=0}^{n} b_k p_k[x_0, \ldots, x_i] = b_i.$$

Znamy wartości wielomianu h w węzłach interpolacyjnych, są nimi liczby $y_0, \ldots, y_n$, a zatem możemy obliczyć współczynniki $b_0, \ldots, b_n$ wielomianu h

w bazie Newtona. Wygodnie jest przedstawić ich obliczenie za pomocą schematu

$$
\begin{array}{c|ccccccc}
x_0 & y_0 = b_0 & & & & & & \\
 & \searrow & & & & & & \\
x_1 & y_1 & \rightarrow & h[x_0, x_1] = b_1 & & & & \\
 & \searrow & & \searrow & & & & \\
x_2 & y_2 & \rightarrow & h[x_1, x_2] & \rightarrow & h[x_0, x_1, x_2] = b_2 & & \\
\vdots & \vdots & & \vdots & & \vdots & \ddots & \\
 & \searrow & & \searrow & & & \searrow & \\
x_n & y_n & \rightarrow & h[x_{n-1}, x_n] & \rightarrow & h[x_{n-2}, x_{n-1}, x_n] \ldots & \rightarrow & h[x_0, \ldots, x_n] = b_n
\end{array}
$$

Podprogram realizujący to obliczenie zastępuje w tablicy y dane wartości funkcji przez współczynniki $b_0, \ldots, b_n$:

```
for ( j = 1; j ⩽ n; j++ )
  for ( i = n; i ⩾ j; i-- )
    y[i] = (y[i] − y[i − 1])/(x[i] − x[i − j]);
```

Jeśli węzły są *uporządkowane monotonicznie*, to obliczone różnice dzielone są kombinacjami liniowymi zaburzonych wartości funkcji danej. Jeśli nie ma nadmiaru i niedomiaru, można dowieść, że (zobacz wzór (**))

$$\mathrm{fl}(h[x_i, \ldots, x_{i+k}]) = \sum_{j=i}^{i+k} c_{ij}^{(k)} y_j (1 + \delta_{ij}^{(k)}), \quad \text{gdzie } |\delta_{ij}^{(k)}| \mathrel{\dot{\leqslant}} 3k\nu.$$

W tej analizie skutki błędów zaokrągleń przedstawia się tylko jako zaburzenia danych wartości funkcji, bez zaburzania węzłów. Jej wynik oznacza, że dla każdego współczynnika $b_k = h[x_0, \ldots, x_k]$ istnieją takie dane zaburzone, $\tilde{y}_0 = y_0(1 + \delta_{00}^{(k)}), \ldots, \tilde{y}_k = y_k(1 + \delta_{0k}^{(k)})$, dla których otrzymany wynik $\tilde{b}_k$ jest dokładny. Ale dla każdego współczynnika b_k zaburzenia danych równoważne popełnionym błędom zaokrągleń mogą być inne — tak więc algorytm różnic dzielonych *jest* numerycznie poprawnym algorytmem obliczania *każdej* z liczb b_k, ale *nie jest* numerycznie poprawnym algorytmem obliczania *całego* wektora $[b_0, \ldots, b_n]$. Z tej tak zwanej <u>numerycznej prawie poprawności</u> wynika numeryczna stabilność algorytmu różnic dzielonych.

Aby rozwiązać zadanie interpolacyjne Hermite'a, należy zmodyfikować ten algorytm. Istotne jest uporządkowanie danych; wymagamy, aby w tablicy x wszystkie „egzemplarze” węzła krotnego występowały obok siebie. W tablicy y zadaną wartość funkcji podajemy w miejscu odpowiadającym pierwszemu wystąpieniu odpowiedniego węzła, a na kolejnych miejscach podajemy wartości kolejnych pochodnych. Algorytm można zrealizować w taki sposób:

```
k[0] = 0;
for ( i = 1; i ⩽ n; i++ )
  k[i] = x[i] == x[i − 1] ? k[i − 1] + 1 : 0;
for ( j = 1; j ⩽ n; j++ )
  for ( i = n; i ⩾ j; i-- )
    if ( k[i] == 0 )
      y[i] = (y[i] − y[i − 1 − k[i − 1]])/(x[i] − x[i − j]);
    else { y[i] /= j; k[i]--; }
```

W pierwszej pętli w miejscu i-tym pomocniczej tablicy k zapisujemy informację, którego rzędu pochodnej wartością jest dana liczba $y[i]$. W drugiej pętli używamy tej informacji do wybrania odpowiedniej instrukcji: obliczenia różnicy dzielonej za pomocą wzoru (*) lub podzielenia $y[i]$ przez odpowiednią liczbę całkowitą, co prowadzi do otrzymania silni w mianowniku wzoru (⁑).

Reszta interpolacyjna

Z uwagi na liczne zastosowania zadań interpolacyjnych Lagrange'a i Hermite'a w aproksymacji funkcji i w konstrukcji różnych metod numerycznych (np. rozwiązywania równań nieliniowych i obliczania całek), duże znaczenie ma wzór opisujący resztę interpolacyjną.

Twierdzenie. *Niech* $h(x)$ *oznacza wielomian interpolacyjny Hermite'a funkcji* f *dla węzłów* $x_0, \dots, x_n$ *w przedziale* $A \subset \mathbb{R}$. *Jeśli funkcja* f *jest klasy* $C^{n+1}(A)$, *to dla każdego* $x \in A$ *istnieje liczba* $\xi \in A$, *taka że*

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

Jeśli funkcja f *jest klasy* $C^n(A)$, *to dla każdego* $x \in A$ *istnieją liczby* $\xi_0, \xi_1 \in A$, *takie że*

$$f(x) - h(x) = \frac{f^{(n)}[\xi_0, \xi_1]}{(n+1)!} p_{n+1}(x).$$

Dowód. Jeśli $x = x_i$ dla pewnego $i \in \{0, \dots, n\}$, to $p_{n+1}(x) = 0$ i obie dowodzone równości są oczywiste (z dowolnym $\xi \in A$ albo $\xi_0, \xi_1 \in A$). Dla ustalonego $x \in A \setminus \{x_0, \dots, x_n\}$ uporządkujmy ciąg $x_0, \dots, x_n, x$ tak, aby otrzymać ciąg niemalejący $x_0^{(0)} \leqslant \cdots \leqslant x_{n+1}^{(0)}$. Niech $f \in C^{n+1}(A)$. Określamy funkcję

$$g_x(s) \stackrel{\text{def}}{=} f(s) - h(s) - z p_{n+1}(s),$$

z parametrem $z = z(x)$, który dobierzemy za chwilę. Korzystając z tego, że $p_{n+1}(x) \neq 0$, bierzemy

$$z = \frac{f(x) - h(x)}{p_{n+1}(x)},$$

i w ten sposób dostajemy $g_x(x) = 0$. Tak określona funkcja spełnia warunek $g_x(x_i^{(0)}) = 0$ dla $i = 0, \dots, n+1$, tzn. ma co najmniej $n+2$ miejsca zerowe[18].

Funkcja g_x jest klasy $C^{n+1}(A)$. Jej pochodna rzędu $k \leqslant n+1$ ma co najmniej $n+2-k$ miejsca zerowe, które tworzą ciąg niemalejący $x_0^{(k)} \leqslant \ldots \leqslant x_{n+1-k}^{(k)}$. Istotnie, jeśli $x_i^{(0)} < x_{i+1}^{(0)}$, to (z twierdzenia Rolle'a) funkcja g_x, która na końcach przedziału $[x_i^{(0)}, x_{i+1}^{(0)}]$ przyjmuje tę samą wartość 0, osiąga wewnątrz tego przedziału maksimum lub minimum, w punkcie $x_i^{(1)}$ będącym miejscem zerowym funkcji g_x'. Jeśli zaś funkcja g_x ma miejsce zerowe o krotności $r > 1$ (w węźle $x_i^{(0)} = \cdots = x_{i+r-1}^{(0)}$), to jej pochodna ma w tym punkcie miejsce zerowe o krotności $r-1$ (zatem mamy podciąg $x_i^{(1)} = \cdots = x_{i+r-2}^{(1)}$). Korzystając z indukcji, stosujemy to rozumowanie do kolejnych pochodnych. Wynika z niego, że pochodna rzędu $n+1$ funkcji g_x ma w przedziale A co najmniej jedno miejsce zerowe, $\xi = x_0^{(n+1)}$.

Podstawiając $s = \xi$, dostajemy

$$0 = g_x^{(n+1)}(\xi) = f^{(n+1)}(\xi) - h^{(n+1)}(\xi) - z p_{n+1}^{(n+1)}(\xi).$$

Pochodna rzędu $n+1$ wielomianu $h(s)$ (stopnia n) jest równa 0, zaś pochodna wielomianu $p_{n+1}(s)$ (stopnia $n+1$), którego współczynnik (w bazie potęgowej) przy s^{n+1} jest równy 1, jest dla każdego s, w tym dla $s = \xi$, równa $(n+1)!$. Zatem

$$z = \frac{f^{(n+1)}(\xi)}{(n+1)!}.$$

Dowód pierwszej równości zakończymy, wstawiając to do definicji funkcji g_x i biorąc $s = x$.

Jeśli funkcja f jest tylko klasy $C^n(A)$, to funkcja g_x określona jak wyżej jest klasy $C^n(A)$ i różniczkowanie należy zakończyć po obliczeniu pochodnej rzędu n. Wynika stąd istnienie co najmniej dwóch miejsc zerowych, $\xi_0 = x_0^{(n)}$ i $\xi_1 = x_1^{(n)}$ funkcji $g_x^{(n)}$, przy czym są one różne, ponieważ mogłyby być identyczne tylko wtedy, gdyby było $x^{(0)} = \cdots = x_{n+1}^{(0)}$, co wykluczyliśmy. Obliczamy

$$0 = g_x^{(n)}[\xi_0, \xi_1] = f^{(n)}[\xi_0, \xi_1] - h^{(n)}[\xi_0, \xi_1] - z p_{n+1}[\xi_0, \xi_1].$$

---

[18] Jeśli pewien węzeł ma krotność $r > 1$, to liczymy go r razy; funkcja g_x przyjmuje w tym punkcie wartość 0 razem z pochodnymi rzędu $1, \dots, r-1$; oczywiście ciąg $x^{(0)}, \dots, x_{n+1}^{(0)}$ nie jest wtedy ściśle rosnący.

W tym rachunku $h^{(n)}$ jest funkcją stałą (bo h ma stopień co najwyżej n), natomiast $p_{n+1}^{(n)}$ jest wielomianem stopnia 1; łatwo jest sprawdzić, że

$$p_{n+1}^{(n)}(s) = (n+1)!\Big(s - \frac{x_0 + \cdots + x_n}{n+1}\Big),$$

skąd wynika, że $p_{n+1}^{(n)}[\xi_0, \xi_1] = p_{n+1}^{(n+1)}(s) = (n+1)!$ dla każdego $s \in \mathbb{R}$. Tak więc

$$z = \frac{f^{(n)}[\xi_0, \xi_1]}{(n+1)!},$$

co należy podstawić do definicji funkcji g_x razem z argumentem $s = x$, aby otrzymać drugą dowodzoną równość. □

Przypuśćmy, że węzły $x_0, \ldots, x_n$ są parami różne i weźmy dowolną liczbę $x \notin \{x_0, \ldots, x_n\}$. Rozważmy wielomian interpolacyjny Lagrange'a $h(s)$ funkcji f dla węzłów $x_0, \ldots, x_n$ i wielomian $h_{n+1}(s)$ stopnia co najwyżej $n+1$, taki że $h_{n+1}(s) = f(s)$ dla $s \in \{x_0, \ldots, x_n, x\}$. Wtedy dla każdego $s \in \mathbb{R}$

$$h_{n+1}(s) = h(s) + f[x_0, \ldots, x_n, x]p_{n+1}(s),$$

gdzie $p_{n+1}(s) = (s - x_0) \cdot \ldots \cdot (s - x_n)$. Powyższa równość ma miejsce dla *dowolnej* funkcji f określonej *w punktach* $x_0, \ldots, x_n, x$. W szczególności, dla $s = x$ mamy

$$h_{n+1}(x) = f(x) = h(x) + f[x_0, \ldots, x_n, x]p_{n+1}(x).$$

W przypadku, gdy funkcja f jest klasy C^{n+1} *w przedziale* zawierającym wszystkie te punkty, na podstawie udowodnionego twierdzenia

$$f(x) = h(x) + \frac{f^{(n+1)}(\xi)}{(n+1)!}p_{n+1}(x),$$

a ponieważ $p_{n+1}(x) \neq 0$, z zestawienia ostatnich dwóch wzorów wynika równość

$$f[x_0, \ldots, x_n, x] = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

dla pewnego ξ należącego do tego przedziału. Stąd wynika fakt, który jest także wnioskiem ze wzoru Hermite'a-Genocchiego[19] i który umożliwia określenie różnic dzielonych dla ciągów, w których występują węzły o krotnościach większych niż 1: jeśli funkcja f jest klasy C^k w otoczeniu x_i, to

$$\lim_{x_{i+1}, \ldots, x_{i+k} \to x_i} f[x_i, \ldots, x_{i+k}] = \frac{f^{(k)}(x_i)}{k!}.$$

[19]Dowód na podstawie wzoru Hermite'a-Genocchiego w odróżnieniu od tego ma zastosowanie także do funkcji zespolonych.

Wzór na resztę interpolacyjną w szczególnym przypadku, gdy $x_0 = \cdots = x_n$, jest wzorem Taylora (z resztą w postaci Lagrange'a). Inny przypadek szczególny, dla dwóch węzłów jednokrotnych, wykorzystaliśmy już w analizie metody siecznych w pierwszym wykładzie. Dalsze zastosowania nastąpią.

Zadania i problemy

1. Rozwiąż zadanie interpolacyjne Lagrange'a (tj. skonstruuj bazę Newtona i oblicz współczynniki wielomianu interpolacyjnego w tej bazie) dla węzłów i wartości funkcji podanych w tabelce:

x_i	-2	-1	0	2	3
$f(x_i)$	17	1	1	1	37

x_i	-2	-1	0	1	2	3
$f(x_i)$	-63	-6	-1	0	21	188

2. Rozwiąż zadanie interpolacyjne Hermite'a dla węzłów i wartości funkcji i pochodnych podanych w tabelce:

x_i	0	1	3
$f(x_i)$	-1	-2	80
$f'(x_i)$		-2	
$f''(x_i)$		12	

x_i	-2	-1	0	2
$f(x_i)$	-64	-7	-2	20
$f'(x_i)$			1	
$f''(x_i)$			-2	

3. Znajdź wyrażenia opisujące współczynniki kombinacji liniowej wartości funkcji

$$f[a, b, c, d] = Af(a) + Bf(b) + Cf(c) + Df(d)$$

oraz wartości funkcji i pochodnej

$$f[a, b, b, c] = Af(a) + B_0f(b) + B_1f'(b) + Cf(c)$$

w punktach a, b, c, d parami różnych.

4. Przejście między bazą Newtona i potęgową może być dokonane za pomocą schematu Hornera; zobaczmy, w jaki sposób. Rozważmy dzielenie z resztą wielomianu $w(x) = a_nx^n + \cdots + a_1x + a_0$ przez dwumian $(x - x_0)$:

$$w(x) = \sum_{i=0}^{n} a_ix^i = \left(\sum_{i=0}^{n-1} c_{i+1}x^i\right)(x - x_0) + c_0 = c_nx^n + \sum_{i=0}^{n-1}(c_i - c_{i+1}x_0)x^i.$$

Stąd $c_n = a_n$ oraz $c_i = c_{i+1}x_0 + a_i$ dla $i < n$. Zauważamy, że liczby c_i są kolejno nadawanymi wartościami zmiennej w podczas obliczania za pomocą (zwykłego) schematu Hornera wartości $w(x_0)$. Otrzymujemy resztę z dzielenia, $b_0 = c_0 = w(x_0)$, i współczynniki w bazie potęgowej ilorazu $c(x)$. Jeśli określimy bazę Newtona, której elementami są wielomiany

$$q_0(x) = 1,$$

$$q_i(x) = \prod_{k=1}^{i-1}(x - x_k) \quad \text{dla } i = 1, \ldots, n-1,$$

to współczynniki wielomianu $c(x)$ w tej bazie są identyczne ze współczynnikami $b_1, \ldots, b_n$ wielomianu w w bazie Newtona $\{p_0, \ldots, p_n\}$. Aby otrzymać b_1, można następnie podzielić wielomian $c(x)$ przez $(x - x_1)$, i tak dalej, rekurencyjnie. Mamy zatem algorytm przejścia od bazy potęgowej do bazy Newtona:

```
for ( j = 0; j < n; j++ )
  for ( i = n − 1; i ⩾ j; i-- )
    a[i] += a[i + 1] ∗ x_j;
```

(algorytm ten zastępuje w tablicy a współczynniki wielomianu w bazie potęgowej współczynnikami w bazie Newtona). Algorytm przejścia w drugą stronę wykonuje operacje przeciwne w odwrotnej kolejności:

```
for ( j = n − 1; j ⩾ 0; j-- )
  for ( i = j; i < n; i++ )
    a[i] -= a[i + 1] ∗ x_j;
```

5. Niech $p_0(x) = 1$, $p_1(x) = x - 1$, $p_2(x) = (x - 1)^2$, $p_3 = (x - 1)^2(x - 2)$, $p_4(x) = (x - 1)^2(x - 2)(x - 3)$ oraz $w(x) = p_0(x) + 2p_2(x) - 3p_3(x) + p_4(x)$. Oblicz $w[2, 3, 4]$ oraz $w[3, 4, 4]$.
6. Algorytm Aitkena. Wartość w dowolnym punkcie x wielomianu interpolacyjnego Lagrange'a określonego przez węzły $x_0, \ldots, x_n$ i wartości w tych węzłach $y_0, \ldots, y_n$, może być obliczona (kosztem $O(n^2)$ operacji) bezpośrednio na podstawie tych danych. Przypuśćmy, że dwa wielomiany stopnia co najwyżej $k - 1$, $p_i^{(k-1)}(x)$ i $p_{i+1}^{(k-1)}(x)$, są rozwiązaniami zadań interpolacyjnych Lagrange'a odpowiednio dla węzłów $x_i, \ldots, x_{i+k-1}$ oraz $x_{i+1}, \ldots, x_{i+k}$. Wtedy wielomian

$$p_i^{(k)}(x) = \frac{x_{i+k} - x}{x_{i+k} - x_i} p_i^{(k-1)}(x) + \frac{x - x_i}{x_{i+k} - x_i} p_{i+1}^{(k-1)}(x)$$

ma stopień co najwyżej k i jest rozwiązaniem zadania dla węzłów $x_i, \ldots, x_{i+k}$. Istotnie, dla $x = x_i$ pierwszy z ułamków w powyższym wzorze ma wartość 1, a drugi 0 (zatem $p_i^{(k)}(x_i) = p_i^{(k-1)}(x_i) = y_i$, dla $x = x_{i+k}$ ułamki mają wartości 0 i 1 (skąd wynika $p_i^{(k)}(x_{i+k}) = p_{i+1}^{(k-1)}(x_{i+k}) = y_{i+k}$), zaś dla każdego $x \in \mathbb{R}$ (w tym dla $x \in \{x_{i+1}, \ldots, x_{i+k-1}\}$) suma ułamków jest równa 1, a więc $p_i^{(k)}(x_j) = y_j$ dla $j = i + 1, \ldots, i + k - 1$.

Wartości wielomianów stopnia 0 w punkcie x, $p_i^{(0)}(x) = y_i$, mamy za darmo. Możemy zatem wpisać liczby $y_0, \ldots, y_n$ do tablicy p, i wykonać algorytm

```
for ( k = 1; k ⩽ n; k++ )
  for ( i = 0; i ⩽ n − k; i++ )
    p[i] = ((x_{i+k} − x) ∗ p[i] + (x − x_i) ∗ p[i + 1])/(x_{i+k} − x_i);
```

otrzymując w ten sposób wartość wielomianu interpolacyjnego w zmiennej $p[0]$.

7. Odpowiednikiem znanego wzoru Leibniza dla pochodnych iloczynu $f(x) = g(x)h(x)$ funkcji odpowiednio gładkich g i h,

$$f^{(k)}(x) = \sum_{l=0}^{k} \binom{k}{l} g^{(l)}(x)h^{(k-l)}(x),$$

jest wzór Leibniza dla różnic dzielonych: jeśli funkcje g i h są odpowiedniej klasy w otoczeniu każdego węzła (stosownie do krotności) i $f(x) = g(x)h(x)$, to

$$f[x_0, \ldots, x_k] = \sum_{l=0}^{k} g[x_0, \ldots, x_l]h[x_l, \ldots, x_k].$$

Dowód I: Indukcja. Oczywiście $f[x_0] = g[x_0]h[x_0]$, załóżmy zatem, że $k > 0$ i wzór obowiązuje dla różnic dzielonych rzędu mniejszego niż k, a ponadto węzły są parami różne. Możemy napisać

$$\begin{aligned}
f[x_0, \ldots, x_k] &\underset{1}{=} \frac{f[x_0, \ldots, x_{k-1}] - f[x_1, \ldots, x_k]}{x_0 - x_k} \\
&\underset{2}{=} \frac{\sum_{l=0}^{k-1} g[x_0, \ldots, x_l]h[x_l, \ldots, x_{k-1}] - \sum_{l=1}^{k} g[x_1, \ldots, x_l]h[x_l, \ldots, x_k]}{x_0 - x_k} \\
&\underset{3}{=} \frac{1}{x_0 - x_k}\Bigg(\sum_{l=0}^{k-1}\Big(g[x_0, \ldots, x_l]h[x_l, \ldots, x_{k-1}] - g[x_0, \ldots, x_l]h[x_{l+1}, \ldots, x_k]\Big) \\
&\qquad - \sum_{l=1}^{k}\Big(g[x_1, \ldots, x_l]h[x_l, \ldots, x_k] - g[x_0, \ldots, x_{l-1}]h[x_l, \ldots, x_k]\Big)\Bigg) \\
&\underset{4}{=} \frac{1}{x_0 - x_k}\Bigg(\sum_{l=0}^{k-1} g[x_0, \ldots, x_l](x_l - x_k)h[x_l, \ldots, x_k] \\
&\qquad + \sum_{l=1}^{k}(x_0 - x_l)g[x_0, \ldots, x_l]h[x_l, \ldots, x_k]\Bigg) \\
&\underset{5}{=} \frac{1}{x_0 - x_k}\Bigg((x_0 - x_k)g[x_0]h[x_0, \ldots, x_k] + \sum_{l=1}^{k-1}(x_0 - x_k)g[x_0, \ldots, x_l]h[x_l, \ldots, x_k] \\
&\qquad + (x_0 - x_k)g[x_0, \ldots, x_k]h[x_k]\Bigg).
\end{aligned}$$

Równość 1 wynika z definicji różnic dzielonych, a równość 2 z założenia indukcyjnego. Do lewej strony równości 3 zostało dodane i odjęte to samo wyrażenie, przez co powstała jej prawa strona. Równosć 4 też wynika z definicji różnic dzielonych, a końcowe wyrażenie powstało przez wyodrębnienie pierwszego składnika pierwszej sumy i ostatniego składnika sumy drugiej oraz połączenia pozostałych składników obu sum w pary. Równość tego wyrażenia i prawej strony wzoru Leibniza jest oczywista.

Dowód, że wzór Leibniza ma zastosowanie także dla przypadku węzłów krotnych, wymaga dokonania odpowiedniego przejścia granicznego. □

Dowód II: Określamy bazy Newtona:

$$p_l(x) = \prod_{j=0}^{l-1}(x - x_j), \qquad q_m(x) = \prod_{j=k-m+1}^{k}(x - x_j).$$

Niech $F(x)$ oznacza iloczyn wielomianów interpolacyjnych Hermite'a funkcji g i h opartych na węzłach $x_0, \dots, x_k$:

$$F(x) = \sum_{l=0}^{k} g[x_0, \dots, x_l]p_l(x) \sum_{m=0}^{k} h[x_m, \dots, x_k]q_{k-m}(x).$$

Dla dowolnego węzła x_i o krotności r funkcje F i f mają w tym węźle takie same wartości i pochodne do rzędu $r-1$ włącznie. Możemy napisać

$$F(x) = \sum_{0 \leqslant l \leqslant m \leqslant k} \big(g[x_0, \dots, x_l]h[x_m, \dots, x_k]p_l(x)q_{k-m}(x)\big) + \sum_{0 \leqslant m < l \leqslant k} (\dots)$$

(składniki obu sum są opisane tym samym wzorem). Dla $m < l$ iloczyn wielomianów $p_l(x)q_{k-m}(x)$ ma w węźle x_i o krotności r miejsce zerowe o krotności co najmniej r, a zatem pierwsza suma (której wszystkie składniki mają stopień nie większy niż k) reprezentuje wielomian interpolacyjny Hermite'a funkcji f, oparty na węzłach $x_0, \dots, x_k$. Współczynnik tego wielomianu odpowiadający wielomianowi p_k w bazie Newtona i jednocześnie x^k w bazie potęgowej jest równy $f[x_0, \dots, x_k]$. Składniki stopnia k w pierwszej sumie mają $l = m$, a pozostałe mają stopień niższy. Pozostaje zauważyć, że współczynnik przy x^k (w bazie potęgowej) pierwszej sumy jest równy wyrażeniu po prawej stronie wzoru Leibniza. $\square$

Interpolacja funkcjami sklejanymi

Motywacja dla stosowania funkcji sklejanych

Funkcje sklejane są to funkcje określone w ten sposób, że pewien przedział $[a, b] \subset \mathbb{R}$ (albo, jeśli jest taka potrzeba, cały zbiór liczb rzeczywistych) dzielimy na podprzedziały, wybierając węzły. W każdym przedziale, którego końcami są węzły, funkcja sklejana stopnia n jest wielomianem stopnia co najwyżej n.

Uwaga. W tym wykładzie posługuję się pojęciem stopnia funkcji sklejanej, tj. największego dopuszczalnego przez reprezentację stopnia wielomianów opisujących tę funkcję w poszczególnych przedziałach, ale w wielu publikacjach i bibliotekach podprogramów jest w użyciu tzw. rząd funkcji sklejanej, tj. liczba o 1 większa od stopnia. Zatem, np. funkcje sklejane pierwszego rzędu to funkcje stopnia 0, czyli kawałkami stałe, wykresem funkcji sklejanej rzędu 2, czyli stopnia 1, jest łamana.

W wielu zastosowaniach funkcje sklejane są wygodniejsze niż funkcje wielomianowe. W szczególności, kształt wykresu funkcji sklejanej nawet niskiego stopnia może być dowolnie skomplikowany, łatwo jest więc aproksymować różne funkcje z dobrą dokładnością. Zastosowanie funkcji sklejanych w interpolacji ma również przewagę nad wielomianami. Występujący we wzorze opisującym rozwiązanie zadania interpolacji Lagrange'a wielomian

$$\Phi_i(x) = \prod_{j\in\{0,\dots,n\}\setminus\{i\}} \frac{x - x_j}{x_i - x_j},$$

który przyjmuje wartość 1 dla $x = x_i$ oraz 0 dla $x = x_j \neq x_i$, między swoimi miejscami zerowymi oscyluje i (zależnie od n i od liczb $x_0, \dots, x_n$) może przyjmować wartości wychodzące daleko poza przedział $[0, 1]$. Interpolacyjne funkcje sklejane nie mają tej wady, dzięki czemu np. wykres funkcji sklejanej przechodzącej przez zadane punkty wydaje się zwykle zgodny z oczekiwaniami.

Rozwiązując zadania interpolacji za pomocą funkcji sklejanych, mamy do czynienia z *dwoma rodzajami* węzłów. Węzły interpolacyjne to punkty w dziedzinie (przedziale $[a, b]$), w których zadajemy wartości funkcji. Węzły funkcji sklejanej to punkty rozdzielające przedziały, w których funkcja może być opisana przez różne wielomiany. Często mamy do czynienia z sytuacją, w której węzły interpolacyjne są jednocześnie węzłami funkcji sklejanej (tej sytuacji poświęcimy szczególną uwagę), ale w ogólności to są *różne pojęcia.*

Z wielu różnych względów w zastosowaniach dominują funkcje sklejane trzeciego stopnia (tzw. kubiczne) i teraz na nich skupimy uwagę.

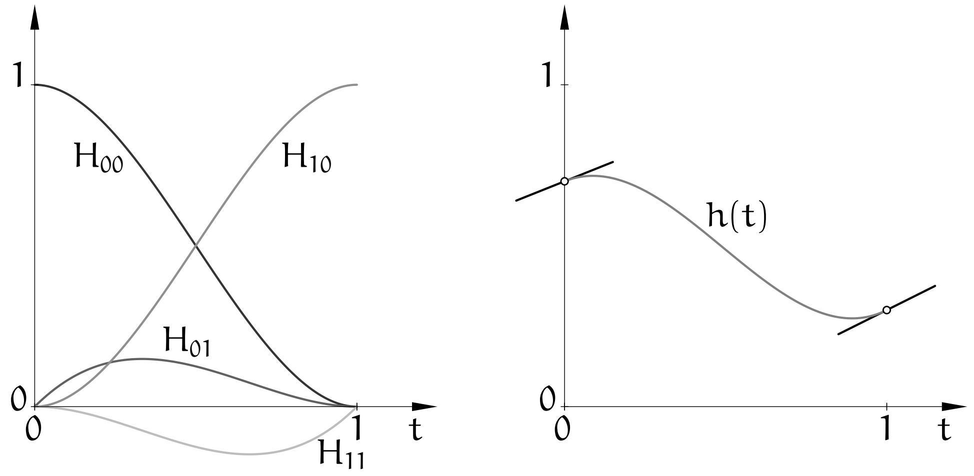
Reprezentacja Hermite'a funkcji sklejanych trzeciego stopnia

Określamy cztery wielomiany:

$$H_{00}(t) = 2t^3 - 3t^2 + 1, \qquad H_{10}(t) = -2t^3 + 3t^2,$$
$$H_{01}(t) = t^3 - 2t^2 + t, \qquad H_{11}(t) = t^3 - t^2.$$

Wielomiany te są rozwiązaniami zadań interpolacyjnych Hermite'a dla dwóch dwukrotnych węzłów, 0 i 1. Mianowicie, zachodzą równości

$$H_{00}(0) = 1, \quad H_{00}(1) = H'_{00}(0) = H'_{00}(1) = 0,$$
$$H_{10}(1) = 1, \quad H_{10}(0) = H'_{10}(0) = H'_{10}(1) = 0,$$
$$H'_{01}(0) = 1, \quad H_{01}(0) = H_{01}(1) = H'_{01}(1) = 0,$$
$$H'_{11}(1) = 1, \quad H_{11}(0) = H_{11}(1) = H'_{11}(0) = 0.$$



Dzięki temu rozwiązanie zadania interpolacyjnego Hermite'a z tymi węzłami dla dowolnej funkcji f możemy zapisać wzorem

$$h(t) = f(0)H_{00}(t) + f'(0)H_{01}(t) + f(1)H_{10}(t) + f'(1)H_{11}(t).$$

Co więcej, przez zamianę zmiennej możemy znaleźć rozwiązanie zadania interpolacyjnego dla dowolnych dwóch węzłów interpolacyjnych o krotności 2. Jeśli węzłami tymi są liczby u_i i u_{i+1}, i oznaczymy $h_i = u_{i+1} - u_i$ (zakładamy, że $h_i > 0$), to mamy stąd wzór

$$h(x) = f(u_i)H_{i,00}(x) + f'(u_i)H_{i,01}(x) + f(u_{i+1})H_{i,10}(x) + f'(u_{i+1})H_{i,11}(x),$$

w którym użyliśmy funkcji

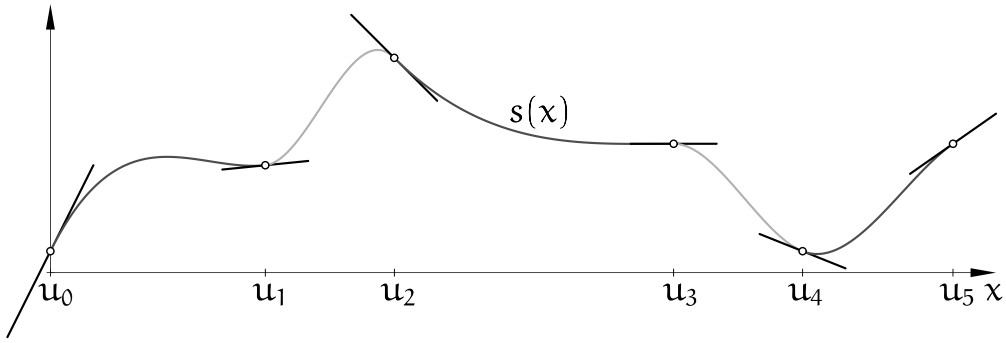
$$H_{i,00}(x) = H_{00}(t), \qquad H_{i,01}(x) = h_i H_{01}(t),$$
$$H_{i,10}(x) = H_{10}(t), \qquad H_{i,11}(x) = h_i H_{11}(t),$$

przy czym $t = (x - u_i)/h_i$.

Dla ustalonych węzłów $u_0, \ldots, u_N$ (tworzących ciąg rosnący), mając dowolne liczby $a_0, \ldots, a_N$ oraz $b_0, \ldots, b_N$, możemy określić funkcję kawałkami wielomianową w przedziale $[u_0, u_N]$ wzorem

$$s(x) = p_i(x) = a_i H_{i,00}(x) + b_i H_{i,01}(x) + a_{i+1} H_{i,10}(x) + b_{i+1} H_{i,11}(x) \qquad \text{dla } x \in [u_i, u_{i+1}].$$

Jest oczywiste, że funkcja ta jest klasy $C^1[u_0, u_N]$, i w węźle u_i ma wartość a_i, a jej pochodna w u_i jest równa b_i, dla $i = 0, \ldots, N$.



Funkcja s jest funkcją sklejaną opartą na ciągu węzłów, z których każdy należy liczyć dwukrotnie. Funkcja ta jest (sklejanym) rozwiązaniem zadania interpolacyjnego Hermite'a, w którym dla każdego węzła zadajemy *dwa* warunki interpolacyjne: wartość funkcji i pochodnej. Funkcja s jest skonstruowana za pomocą wielomianów spełniających warunki interpolacyjne Hermite'a dla dwóch węzłów o krotności 2 (końców każdego przedziału (u_i, u_{i+1})) i dlatego jej reprezentacja w tej postaci jest nazywana reprezentacją Hermite'a[20].

Kubiczne interpolacyjne funkcje sklejane klasy C^2

Aby określić funkcję sklejaną trzeciego stopnia, która jest rozwiązaniem zadania interpolacyjnego Lagrange'a (tj. dla każdego węzła u_i chcemy podać tylko jedną liczbę, a_i, będącą wartością funkcji), skorzystamy z warunku ciągłości pochodnej drugiego rzędu. Zatem, pochodne drugiego rzędu wielomianów $p_{i-1}(x)$ i $p_i(x)$, opisujących poszukiwaną funkcję s w przedziałach (u_{i-1}, u_i) oraz (u_i, u_{i+1}), mają przyjmować w punkcie u_i tę samą wartość (reprezentacja Hermite'a gwarantuje, że będą miały w tym punkcie identyczne wartości i pochodne pierwszego rzędu). Na tej podstawie możemy obliczyć liczby b_i, tj. wartości pochodnej pierwszego rzędu funkcji s w węzłach interpolacyjnych.

[20] W czasach, gdy żył Charles Hermite (1822–1901 r.), funkcje sklejane nie były jeszcze znane; odkrył je Schoenberg w 1946 r.

Aby wyprowadzić odpowiednie równania, należy obliczyć pochodne wielomianów, za pomocą których przedstawiamy rozwiązanie:

$$
\begin{aligned}
&H''_{i-1,00}(u_i) = \frac{6}{h_{i-1}^2}, \quad H''_{i,00}(u_i) = \frac{-6}{h_i^2}, \quad H''_{i-1,10}(u_i) = \frac{-6}{h_{i-1}^2}, \quad H''_{i,10}(u_i) = \frac{6}{h_i^2},\\
&H''_{i-1,01}(u_i) = \frac{2}{h_{i-1}}, \quad H''_{i,01}(u_i) = \frac{-4}{h_i}, \quad H''_{i-1,11}(u_i) = \frac{4}{h_{i-1}}, \quad H''_{i,11}(u_i) = \frac{-2}{h_i}.
\end{aligned}
$$

Zatem, warunek ciągłości drugiej pochodnej w punkcie u_i, $p''_{i-1}(u_i) = p''_i(u_i)$, ma postać

$$
\frac{6}{h_{i-1}^2}a_{i-1} - \frac{6}{h_{i-1}^2}a_i + \frac{2}{h_{i-1}}b_{i-1} + \frac{4}{h_{i-1}}b_i = -\frac{6}{h_i^2}a_i + \frac{6}{h_i^2}a_{i+1} - \frac{4}{h_i}b_i - \frac{2}{h_i}b_{i+1}.
$$

Po pomnożeniu stron przez $h_{i-1}h_i/2$ i uporządkowaniu, dostajemy równanie

$$
h_ib_{i-1} + 2(h_{i-1}+h_i)b_i + h_{i-1}b_{i+1} = 3\biggl(\frac{h_i}{h_{i-1}}(a_i - a_{i-1}) + \frac{h_{i-1}}{h_i}(a_{i+1}-a_i)\biggr). \tag{*}
$$

Uwaga: Powyższe równanie można wyprowadzić także bez jawnego wypisywania i różniczkowania wielomianów — elementów bazy. W tym celu posłużymy się schematem różnic dzielonych. Dla wielomianu p_i opisującego funkcję sklejaną w przedziale $[u_i, u_{i+1}]$ mamy

$$
\begin{array}{l|lllllllll}
u_i & a_i \\
 & & \searrow \\
u_i & a_i & \to & b_i \\
 & & \searrow & & \searrow \\
u_i & a_i & \to & b_i & \to & \frac{1}{2}p''_i(u_i) \\
 & & \searrow & & \searrow & & \searrow \\
u_{i+1} & a_{i+1} & \to & p_i[u_i,u_{i+1}] & \to & p_i[u_i,u_i,u_{i+1}] & \to & p_i[u_i,u_i,u_i,u_{i+1}] \\
 & & \searrow & & \searrow & & \searrow & & \searrow \\
u_{i+1} & a_{i+1} & \to & b_{i+1} & \to & p_i[u_i,u_{i+1},u_{i+1}] & \to & p_i[u_i,u_i,u_{i+1},u_{i+1}] & \to & 0
\end{array}
$$

Różnica dzielona rzędu 4 wielomianu stopnia 3 jest (dla dowolnych węzłów) równa 0, dlatego ma miejsce równość

$$
p_i[u_i,u_i,u_i,u_{i+1}] = p_i[u_i,u_i,u_{i+1},u_{i+1}].
$$

Po pomnożeniu stron przez $h_i = u_{i+1} - u_i$ możemy napisać

$$
p_i[u_i,u_i,u_{i+1}] - \frac{1}{2}p''_i(u_i) = p_i[u_i,u_{i+1},u_{i+1}] - p_i[u_i,u_i,u_{i+1}],
$$

czyli

$$
\frac{1}{2}p''_i(u_i) = 2p_i[u_i,u_i,u_{i+1}] - p_i[u_i,u_{i+1},u_{i+1}].
$$

Do tego wzoru podstawimy

$$
p_i[u_i,u_i,u_{i+1}] = \frac{p_i[u_i,u_{i+1}] - b_i}{h_i}, \quad p_i[u_i,u_{i+1},u_{i+1}] = \frac{b_{i+1} - p_i[u_i,u_{i+1}]}{h_i},
$$

skąd po uporządkowaniu dostaniemy

$$\frac{1}{2}p_i''(u_i) = \frac{1}{h_i}\big(3p_i[u_i, u_{i+1}] - 2b_i - b_{i+1}\big).$$

Zamieniając w tym rachunku wielomian p_i na p_{i-1} i węzeł u_{i+1} na u_{i-1}, znajdziemy

$$\frac{1}{2}p_{i-1}''(u_i) = \frac{1}{h_{i-1}}\big(3p_{i-1}[u_i, u_{i-1}] - 2b_i - b_{i-1}\big).$$

Równanie (*) otrzymamy, żądając aby było

$$\frac{1}{h_{i-1}}\big(3p_{i-1}[u_i, u_{i-1}] - 2b_i - b_{i-1}\big) = \frac{1}{h_i}\big(3p_i[u_i, u_{i+1}] - 2b_i - b_{i+1}\big),$$

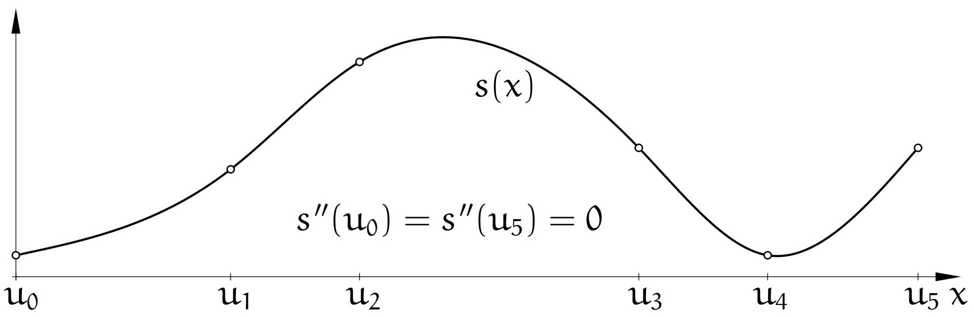
co oznacza równość pochodnych drugiego rzędu obu wielomianów w punkcie u_i, mnożąc strony tej równości przez $h_{i-1}h_i$ i odpowiednio przenosząc składniki. □

W ten sposób otrzymaliśmy równania ciągłości pochodnych drugiego rzędu funkcji s w „wewnętrznych” węzłach $u_1, \ldots, u_{N-1}$; dla ustalonych liczb $a_0, \ldots, a_N$ musimy znaleźć liczby $b_0, \ldots, b_N$ spełniające te równania. Zauważamy, że liczba niewiadomych jest o 2 większa niż liczba równań. Aby mieć rozwiązanie jednoznaczne, trzeba dołożyć dodatkowe dwa równania.

Dodatkowe równania w konstrukcji sklejanych funkcji interpolacyjnych opisują zwykle tzw. <u>warunki brzegowe</u>, tj. pewne warunki narzucone na pochodne funkcji s w skrajnych węzłach u_0 i u_N. Najprostszy sposób, to arbitralne określenie wartości pochodnych pierwszego rzędu, tj. liczb b_0 i b_N. To jednak może być kłopotliwe dla użytkownika programu. Często stosowanym rozwiązaniem jest żądanie, aby pochodna drugiego rzędu funkcji s była w punktach u_0 i u_N równa 0. Powstaje wtedy tzw. <u>naturalna funkcja sklejana</u>. Na podstawie wypisanych wcześniej wartości funkcji $H_{i,jk}$, równania $p_0''(u_0) = 0$ i $p_{N-1}''(u_N) = 0$ możemy przedstawić w postaci

$$\begin{aligned} 2h_0b_0 + h_0b_1 &= 3(a_1 - a_0),\\ h_{N-1}b_{N-1} + 2h_{N-1}b_N &= 3(a_N - a_{N-1}). \end{aligned}$$

Po dołączeniu tych równań otrzymujemy układ równań liniowych z macierzą trójdiagonalną. Możemy zauważyć, że dla dowolnego rozmieszczenia węzłów, tj. dla dowolnych dodatnich liczb $h_0, \ldots, h_{N-1}$, macierz ta jest diagonalnie dominująca. Zatem, mamy układ o jednoznacznym rozwiązaniu, które możemy znaleźć za pomocą eliminacji Gaussa kosztem $O(N)$ działań arytmetycznych.



Istnieje wiele innych sposobów określania warunków brzegowych, np. można zażądać, aby pochodna trzeciego rzędu wielomianów p_0 i p_{N-1} była równa 0 (zatem, aby były to wielomiany drugiego stopnia), lub też, aby wielomian p_0 był identyczny z p_1, a wielomian p_{N-1} z p_{N-2} (a więc, aby węzły interpolacyjne u_1 i u_{N-1} *nie były* węzłami funkcji sklejanej — po angielsku nazywa się to warunek *not a knot*).

Jeszcze inna możliwość, to przyjęcie $b_N = b_0$ i wymaganie, aby pochodna drugiego rzędu funkcji s w węzłach u_0 i u_N była taka sama. Wtedy, jeśli $a_0 = a_N$, tj. funkcja s ma tę samą wartość w punktach u_0 i u_N, skonstruujemy tzw. okresową funkcję sklejaną — nakładając warunek, że dla każdego $x \in \mathbb{R}$ $s(x + T) = s(x)$, gdzie $T = u_N - u_0$, otrzymujemy okresową funkcję klasy $C^2(\mathbb{R})$. W układzie równań (*) zastępujemy niewiadomą b_N przez b_0 i dołączamy równanie

$$h_0 b_{N-1} + 2(h_{N-1} + h_0)b_0 + h_{N-1}b_1 = 3\left(\frac{h_0}{h_{N-1}}(a_0 - a_{N-1}) + \frac{h_{N-1}}{h_0}(a_1 - a_0)\right),$$

otrzymując w ten sposób układ równań z macierzą cykliczną trójdiagonalną, diagonalnie dominującą.

Twierdzenie Holladaya

Dla dużej liczby węzłów wielomiany interpolacyjne Lagrange'a „źle się zachowują", tj. ich wartości między sąsiednimi węzłami mogą wystawać daleko poza przedział, którego końcami są wartości wielomianu w tych węzłach. Udowodnimy twierdzenie, które można zinterpretować w ten sposób, że pod tym względem najlepiej, jak tylko się da, zachowuje się naturalna kubiczna interpolacyjna funkcja sklejana. W tym celu określimy funkcjonał, który nazwiemy energią, i który przyjmiemy za miarę „powyginania" wykresów funkcji klasy $C^2[u_0, u_N]$:

$$E(f) \stackrel{\text{def}}{=} \int_{u_0}^{u_N} \big(f''(x)\big)^2\,dx.$$

<u>Twierdzenie Holladaya.</u> *W zbiorze funkcji klasy* $C^2[u_0, u_N]$ *spełniających warunki interpolacyjne Lagrange'a określone w węzłach* $u_0, \ldots, u_N$ *najmniejszą energię ma naturalna kubiczna funkcja sklejana.*

<u>Dowód.</u> Niech s oznacza naturalną kubiczną funkcję sklejaną spełniającą zadane warunki interpolacyjne, i niech f oznacza dowolną inną funkcję klasy C^2, przyjmującą w węzłach te same wartości. Mamy $f = (f - s) + s$, zatem

$$E(f) = \int_{u_0}^{u_N} \bigl(f''(x) - s''(x)\bigr)^2\,dx + 2\int_{u_0}^{u_N} \bigl(f''(x) - s''(x)\bigr)s''(x)\,dx + \int_{u_0}^{u_N} \bigl(s''(x)\bigr)^2\,dx.$$

Obliczymy drugą całkę w powyższym wzorze, całkując przez części:

$$\int_{u_0}^{u_N} \bigl(f''(x) - s''(x)\bigr)s''(x)\,dx = \bigl(f'(x) - s'(x)\bigr)s''(x)\Big|_{u_0}^{u_N} - \int_{u_0}^{u_N} \bigl(f'(x) - s'(x)\bigr)s'''(x)\,dx.$$

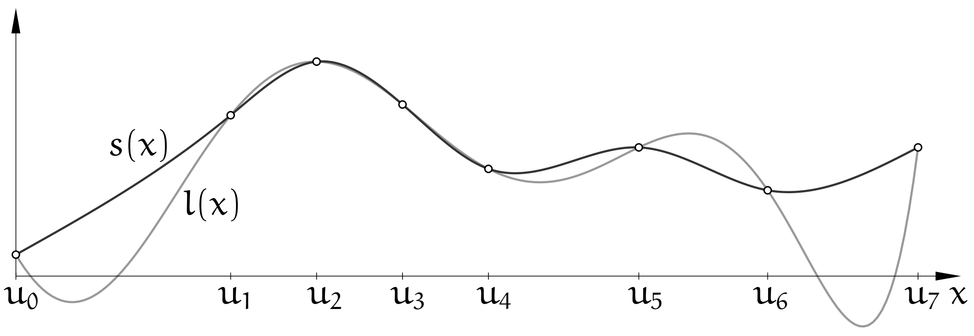
Dla naturalnej funkcji sklejanej s jest $s''(u_0) = s''(u_N) = 0$, ponadto w każdym przedziale (u_i, u_{i+1}) pochodna trzeciego rzędu funkcji s jest stała; oznaczmy ją symbolem s_i. Rozpatrywana całka jest więc równa

$$-\sum_{i=0}^{N-1} \int_{u_i}^{u_{i+1}} \bigl(f'(x) - s'(x)\bigr)s_i\,dx = -\sum_{i=0}^{N-1} s_i\bigl(f(x) - s(x)\bigr)\Big|_{u_i}^{u_{i+1}} = 0,$$

bo dla każdego i jest $f(u_i) = s(u_i)$. Mamy stąd

$$E(f) = \int_{u_0}^{u_N} \bigl(f''(x) - s''(x)\bigr)^2\,dx + \int_{u_0}^{u_N} \bigl(s''(x)\bigr)^2\,dx.$$

Jeśli funkcje f i s mają taką samą pochodną drugiego rzędu w przedziale $[u_0, u_N]$ i są różne, to ich różnica jest wielomianem stopnia 0 lub 1, ale wtedy nie mogą przyjmować tych samych wartości we wszystkich węzłach. Zatem, jeśli przyjmują i są różne, to zachodzi nierówność $E(f) > E(s)$, co pragnęliśmy udowodnić. □



Na rysunku są wykresy dwóch funkcji interpolacyjnych dla tych samych danych: kubicznej naturalnej funkcji sklejanej $s(x)$ klasy C^2 i wielomianu interpolacyjnego Lagrange'a $l(x)$ stopnia 7. Niewątpliwie wykres tego wielomianu jest bardziej „powyginany".

Funkcje B-sklejane

Kubiczne funkcje sklejane są odpowiednie w większości zastosowań praktycznych, ale w pewnych przypadkach potrzebne są też funkcje sklejane innych stopni. Reprezentacje Hermite'a, takie jak opisana wcześniej reprezentacja trzeciego stopnia, są w miarę wygodne dla funkcji stopni nieparzystych, ale jeszcze wygodniejsza, dla dowolnego stopnia, jest reprezentacja B-sklejana. Jej podstawą są tzw. unormowane funkcje B-sklejane które mają *kilka* równoważnych definicji[21]. Jedna z nich wykorzystuje tzw. obciętą funkcję potęgową stopnia n, określoną wzorem

$$(x-u)_+^n \stackrel{\text{def}}{=} \begin{cases} (x-u)^n & \text{dla } x \geqslant u, \\ 0 & \text{dla } x < u. \end{cases}$$

Możemy zauważyć, że dla ustalonego u funkcja ta jest klasy $C^{n-1}(\mathbb{R})$. Co więcej, jeśli dwa wielomiany stopnia co najwyżej n, $p_1(x)$ i $p_2(x)$, i ich pochodne rzędu $1, \ldots, n-1$ mają w punkcie u odpowiednio takie same wartości, to istnieje stała c, taka że funkcję sklejaną s klasy C^{n-1} opisaną przez p_1 dla $x < u$ i przez p_2 dla $x \geqslant u$ można przedstawić w postaci sumy: $s(x) = p_1(x) + c(x-u)_+^n$. Jeszcze jedno: *umawiamy się*, że jeśli $x = u$, to $(x-u)_+^0 = 1$ (dzięki temu wiemy, co robić z wyrażeniem 0^0 w tym zastosowaniu).

Def. Niech $u_0 \leqslant u_1 \leqslant \cdots \leqslant u_{N-1} \leqslant u_N$. Niech $f_x(u) = (x-u)_+^n$. Unormowana funkcja B-sklejana stopnia n z węzłami $u_i, \ldots, u_{i+n+1}$ jest określona wzorem

$$N_i^n(x) \stackrel{\text{def}}{=} (-1)^{n+1}(u_{i+n+1} - u_i) f_x[u_i, \ldots, u_{i+n+1}].$$

Jeśli $u_i = \cdots = u_{i+n+1}$, to[22] funkcja N_i^n jest zerowa.

W rozważaniach teoretycznych przyjmuje się też nieskończone ciągi węzłów (np. złożone z wszystkich liczb całkowitych), co bywa wygodne. Dla uporządkowanego

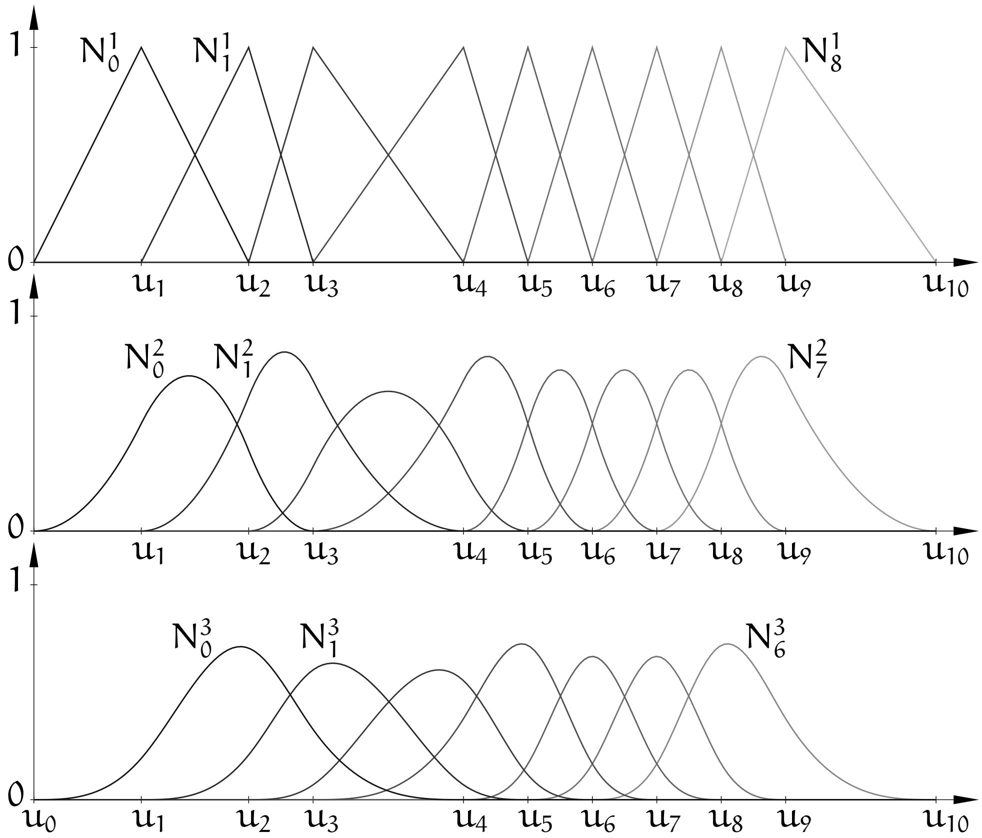
[21] Niezależnie od sposobów definiowania, których jest wiele, w literaturze można spotkać też wiele różnych sposobów numerowania i oznaczania węzłów i funkcji B-sklejanych. Tu używam jednego z nich.

[22] Różnica dzielona funkcji f_x jest pochodną rzędu $n+1$ i dla $x = u_i$ jest nieokreślona. Do opisu takich sytuacji przydają się dystrybucje — tu tzw. delta Diraca — które nie są funkcjami w zwykłym sensie. Przyjmując, że $N_i^n = 0$, pozbywamy się kłopotu.

ciągu $N+1$ węzłów, biorąc podciągi złożone z kolejnych $n+2$ węzłów, można określić $N-n$ funkcji B-sklejanych stopnia n, $N_0^n, \dots, N_{N-n-1}^n$. Funkcję sklejaną stopnia n z węzłami $u_0, \dots, u_N$ przedstawia się w postaci

$$s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x).$$

Wykresy przykładowych funkcji B-sklejanych stopnia 1, 2 i 3 wyglądają tak:



Nazwa „funkcje B-sklejane” ma dwa źródła: kształty wykresów funkcji, które mają jedno maksimum, przy odrobinie wyobraźni przypominają przekrój dzwonu (stąd angielskie określenie *bell-shaped*), a ponadto funkcje te (dla węzłów spełniających opisany dalej warunek progresywności) tworzą bazę rozpiętej przez nie przestrzeni liniowej funkcji sklejanych. Pokażę kilka własności funkcji B-sklejanych:

Własność 1. *Funkcja* N_i^n *poza przedziałem* $[u_i, u_{i+n+1})$ *ma wartość* 0.

Dowód. Liczby u_i i u_{i+n+1} to najmniejszy i największy argument różnicy dzielonej w definicji funkcji N_i^n. Jeśli zatem $x < u_i$, to funkcja f_x ma wartość 0 dla każdego $u \in \{u_i, \dots, u_{i+n+1}\}$ i jej różnica dzielona jest różnicą dzieloną funkcji zerowej. Z drugiej strony, dla $u \leqslant x$ funkcja $f_x(u)$ jest wielomianem stopnia n. Jeśli więc $x \geqslant u_{i+n+1} \geqslant \dots \geqslant u_i$, to jej różnica dzielona rzędu $n+1$ też jest równa 0. $\square$

Jeśli $u_i = u_{i+n+1}$, to funkcja N_i^n jest funkcją zerową. Dlatego zwykle, mając ustalony stopień n, narzucamy na ciąg węzłów warunek progresywności: oprócz tego, że ciąg jest niemalejący, ma być $u_i < u_{i+n+1}$ dla $i = 0, \ldots, N-n-1$.

Własność 2. *Wartości funkcji N_i^n można obliczać za pomocą wzoru Mansfielda-de Boora-Coxa:*

$$N_i^0(x) = \begin{cases} 1 & \text{dla } x \in [u_i, u_{i+1}), \\ 0 & \text{w przeciwnym razie,} \end{cases}$$

$$N_i^n(x) = \frac{x - u_i}{u_{i+n} - u_i} N_i^{n-1}(x) + \frac{u_{i+n+1} - x}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x) \quad \text{dla } n > 0.$$ [23]

Dowód. Dla funkcji N_i^0 mamy $f_x(u) = (x-u)_+^0$, skąd wynika

$$N_i^0(x) = -(u_{i+1} - u_i) \frac{(x-u_i)_+^0 - (x-u_{i+1})_+^0}{u_i - u_{i+1}},$$

i wystarczy zbadać trzy przypadki: $x < u_i$, $u_i \leqslant x < u_{i+1}$ i $u_{i+1} \leqslant x$.

Jeśli $n > 0$, to możemy przedstawić funkcję $f_x(u)$ w postaci iloczynu funkcji $g(u) = x - u$ i $h(u) = (x-u)_+^{n-1}$ i skorzystać ze wzoru Leibniza (ćwiczenie 7.7). Ponieważ pierwszy czynnik jest wielomianem stopnia 1, jego różnice dzielone rzędów wyższych niż 1 są równe 0 i wzór Leibniza dla tego iloczynu ma postać

$$f_x[u_i, \ldots, u_{i+n+1}] = g[u_i]h[u_i, \ldots, u_{i+n+1}] + g[u_i, u_{i+1}]h[u_{i+1}, \ldots, u_{i+n+1}],$$

przy czym $g[u_i, u_{i+1}] = -1$. Na podstawie definicji funkcji B-sklejanych

$$h[u_i, \ldots, u_{i+n}] = \frac{(-1)^n N_i^{n-1}(x)}{u_{i+n} - u_i}, \quad h[u_{i+1}, \ldots, u_{i+n+1}] = \frac{(-1)^n N_{i+1}^{n-1}(x)}{u_{i+n+1} - u_{i+1}}.$$

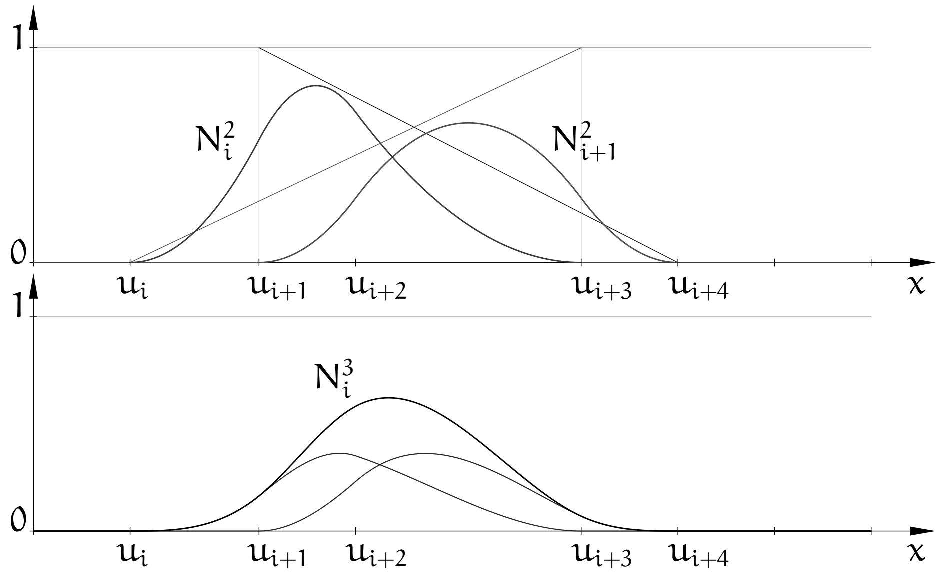
Możemy na tej podstawie obliczyć

$$N_i^n(x) = (-1)^{n+1}(u_{i+n+1} - u_i) \times$$
$$\times \left((x - u_i) \frac{\frac{(-1)^n N_i^{n-1}(x)}{u_{i+n} - u_i} - \frac{(-1)^n N_{i+1}^{n-1}(x)}{u_{i+n+1} - u_{i+1}}}{u_i - u_{i+n+1}} - \frac{(-1)^n N_{i+1}^{n-1}(x)}{u_{i+n+1} - u_{i+1}} \right).$$

Aby dokończyć dowód, wystarczy uporządkować wyrażenie po prawej stronie powyższej równości. □

Na rysunku jest pokazany przykład, tj. wykresy funkcji B-sklejanych stopnia 2 i pozostałych funkcji, za pomocą których jest skonstruowana funkcja B-sklejana trzeciego stopnia.

[23] Jeśli $u_{i+n} = u_i$ lub $u_{i+1} = u_{i+n+1}$, to funkcja N_i^{n-1} lub N_{i+1}^{n-1} jest funkcją zerową; odpowiedni składnik wyrażenia opisującego funkcję N_i^n znika (a zatem nie dzielimy przez 0).



W wielu książkach, zwłaszcza na temat grafiki komputerowej, wzór Mansfielda-de Boora-Coxa jest użyty do zdefiniowania funkcji B-sklejanych. W odróżnieniu od definicji opartej na różnicach dzielonych i obciętej funkcji potęgowej ten wzór jest podstawą dla efektywnych algorytmów, podanych dalej.

<u>Własność 3</u>. *W każdym przedziale* $(u_j, u_{j+1}) \subset [u_i, u_{i+n+1}]$ *funkcja* $N_i^n(x)$ *jest wielomianem stopnia* n. *Jeśli liczba* u_j *w ciągu* $u_i, \ldots, u_{i+n+1}$ *występuje* r *razy, to funkcja* N_i^n *ma w punkcie* u_j *ciągłe pochodne rzędu* $1, \ldots, n-r$.

<u>Dowód</u>. To, że między sąsiednimi węzłami funkcja B-sklejana jest wielomianem stopnia co najwyżej n, natychmiast wynika ze wzoru Mansfielda-de Boora-Coxa. Fakt, że stopień niezerowego wielomianu opisującego funkcję N_i^n w takim przedziale jest równy n można udowodnić, na przykład badając jej pochodną rzędu n (wzór opisujący pochodną będzie dalej).

Jeśli pewien węzeł ma krotność $r > 1$, to różnica dzielona funkcji f_x ma składniki, które są iloczynami stałych (określonych przez węzły) i pochodnych funkcji $f_x(u)$ rzędu $1, \ldots, r-1$ w punkcie u_j. Dla $k \leqslant n$ pochodna rzędu k funkcji f_x w punkcie u_j jest równa $(-1)^k \frac{n!}{(n-k)!}(x-u_j)_+^{n-k}$. Ponieważ dla każdego składnika jest $k < r$, wszystkie składniki mają pochodną rzędu $n-r$ względem x ciągłą w u_j. $\square$

W szczególności jeśli węzły są jednokrotne, to funkcje B-sklejane stopnia n są klasy $C^{n-1}(\mathbb{R})$. Jeśli zaś $u_i = \cdots = u_{i+n} < u_{i+n+1}$, to funkcja N_i^n jest nieciągła w punkcie u_i i podobnie jeśli $u_i < u_{i+1} = \cdots = u_{i+n+1}$, to funkcja N_i^n jest nieciągła w punkcie u_{i+1}.

Własność 4 (rozkład jedynki)[24]. *Funkcje* N_i^n *są nieujemne i dla każdego* $x \in [u_n, u_{N-n})$ *zachodzi równość*

$$\sum_{i=0}^{N-n-1} N_i^n(x) = 1.$$

Co więcej, jeśli $x \in [u_k, u_{k+1}) \subset [u_n, u_{N-n})$, *to*

$$\sum_{i=k-n}^{k} N_i^n(x) = 1.$$

Dowód. Prosty dowód indukcyjny na podstawie wzoru Mansfielda-de Boora-Coxa polecam jako ćwiczenie. □

Własność 5 (otoczki wypukłej). *Jeśli* $x \in [u_k, u_{k+1}) \subset [u_n, u_{N-n})$, *to wartość funkcji sklejanej*

$$s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x) = \sum_{i=k-n}^{k} d_i N_i^n(x)$$

spełnia nierówności $\min_{i=k-n,\dots,k} d_i \leqslant s(x) \leqslant \max_{i=k-n,\dots,k} d_i$.

Dowód. To jest natychmiastowy wniosek z własności 1 i 4. □

Z uwagi na te własności zwykle w zastosowaniach używa się funkcji B-sklejanych określonych dla ciągu węzłów wybranego tak, aby przedział $[u_n, u_{N-n}]$ był potrzebną w konkretnym zastosowaniu dziedziną[25].

Własność 6. *Pochodna funkcji B-sklejanej stopnia* $n > 0$, *w punktach* x, *w których istnieje, wyraża się wzorem*

$$\frac{d}{dx} N_i^n(x) = \frac{n}{u_{i+n} - u_i} N_i^{n-1}(x) - \frac{n}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x).$$

Dowód. Aby otrzymać pochodną funkcji N_i^n, możemy we wzorze ją definiującym zastąpić funkcję $f_x(u)$ przez jej pochodną ze względu na parametr x, tj. przez funkcję $p_x(u) = n(x-u)_+^{n-1}$. Dostajemy wtedy równość

$$\begin{aligned} N_i^{n\prime}(x) &= (-1)^{n+1}(u_{i+n+1} - u_i) p_x[u_i, \dots, u_{i+n+1}] \\ &= (-1)^{n+1}(u_{i+n+1} - u_i) \frac{p_x[u_i, \dots, u_{i+n}] - p_x[u_{i+1}, \dots, u_{i+n+1}]}{u_i - u_{i+n+1}}, \end{aligned}$$

której przekształcenie do potrzebnej postaci pozostawiam jako ćwiczenie. □

---

[24] Właśnie dla osiągnięcia tej własności został w definicji funkcji N_i^n wprowadzony czynnik $(-1)^{n+1}(u_{i+n+1} - u_i)$ i w tym sensie funkcje te są „unormowane”.

[25] Dziedzinę domykamy z prawej strony, przyjmując, że wartość funkcji sklejanej w u_{N-n} jest wartością wielomianu opisującego tę funkcję w przedziale $[u_{N-n-1}, u_{N-n})$.

Pochodna funkcji N_i^n w punkcie x nie istnieje, jeśli liczba x występuje n lub $n+1$ razy w ciągu $u_i, \ldots, u_{i+n+1}$. Z własności 1, 3 i 6 wynika, że jeśli $s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x)$ oraz $x \in (u_n, u_{N-n})$, i x nie jest węzłem o krotności większej niż $n-1$, to

$$s'(x) = \sum_{i=0}^{N-n-2} \frac{n(d_{i+1} - d_i)}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x),$$

przy czym jeśli $x \in (u_k, u_{k+1}) \subset (u_n, u_{N-n})$, to

$$s'(x) = \sum_{i=k-n}^{k-1} \frac{n(d_{i+1} - d_i)}{u_{i+n+1} - u_{i+1}} N_{i+1}^{n-1}(x). \qquad (\square)$$

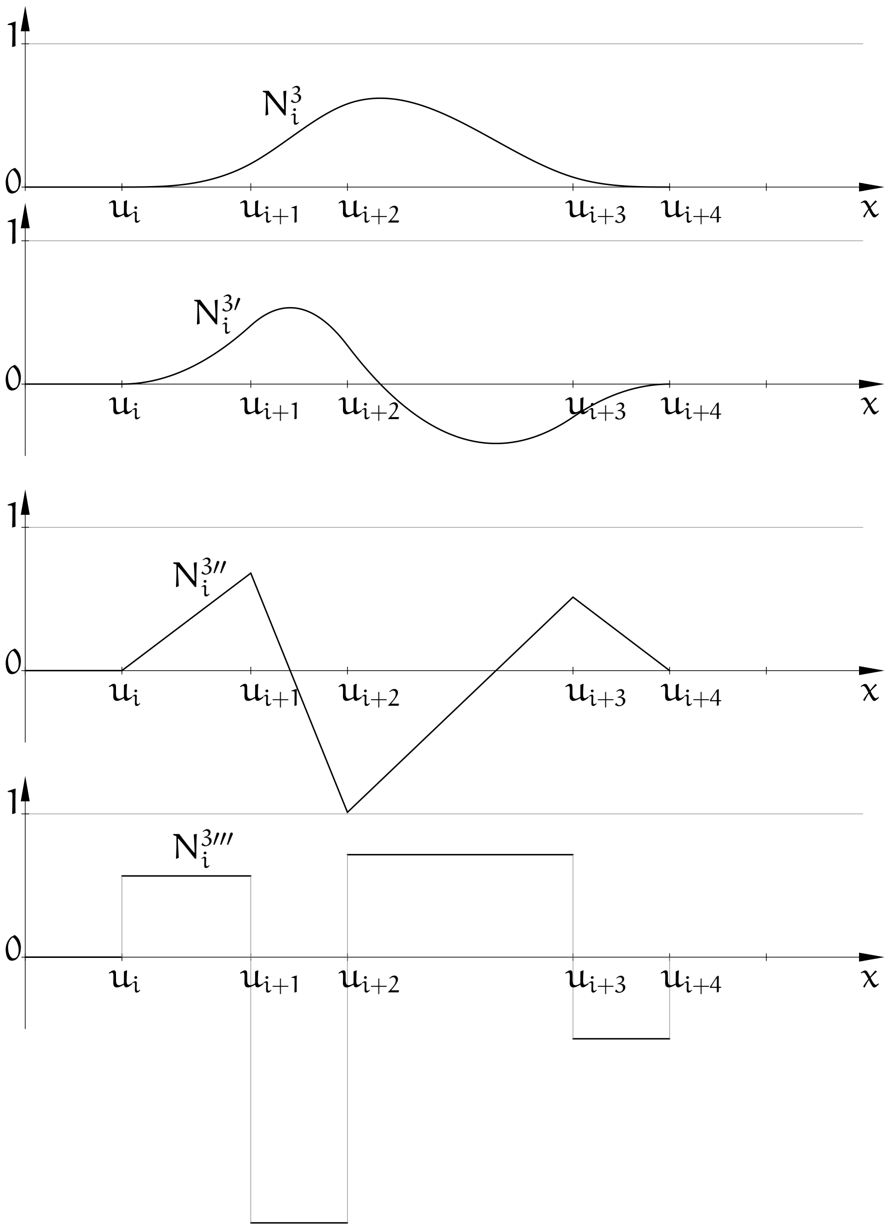
Dowód tych wzorów też polecam jako ćwiczenie. Trzeba w nim skorzystać ze spostrzeżenia, że funkcje N_0^{n-1} i N_{N-n}^{n-1} są w przedziale $[u_n, u_{N-n})$ równe 0.

<u>Własność 7 (lokalna liniowa niezależność)</u>. *Jeśli* $u_k < u_{k+1}$, *to funkcje* $N_{k-n}^n, \ldots, N_k^n$ *w przedziale* $[u_k, u_{k+1})$ *są wielomianami liniowo niezależnymi.*

<u>Dowód</u>. Wystarczy wykazać, że jeśli funkcja $s(x) = \sum_{i=k-n}^{k} d_i N_i^n(x)$ jest w całym przedziale $[u_k, u_{k+1})$ równa zeru, to $d_{k-n} = \cdots = d_k = 0$. Dla $n = 0$ to jest oczywiste. Przypuśćmy zatem, że $n > 0$ i że funkcje $N_{k-n+1}^{n-1}, \ldots, N_k^{n-1}$ w przedziale $[u_k, u_{k+1})$ są liniowo niezależnymi wielomianami stopnia $n-1$. Jeśli funkcja s znika w tym przedziale, to również jej pochodna jest w nim równa 0, ale ze wzoru ($\square$) i z przypuszczenia wynika, że $d_{i+1} - d_i = 0$ dla $i = k-n, \ldots, k-1$. Stąd funkcja zerowa s ma współczynniki $d_{k-n} = \cdots = d_k$. Ale z własności otoczki wypukłej wynika, że dla każdego $x \in [u_k, u_{k+1})$ jest $\sum_{i=k-n}^{k} d_k N_i^n(x) = d_k$, więc musi być $d_k = 0$ i tym sposobem krok indukcyjny został zrobiony. $\square$

Na podstawie własności 7 łatwo jest dowieść (ćwiczenie), że zbiór funkcji $N_0^n, \ldots, N_{N-n-1}^n$ określony dla progresywnego ciągu węzłów $u_0, \ldots, u_N$ jest liniowo niezależny, a nawet że zbiór nieskończony $\{\, N_i^n \colon i \in \mathbb{Z} \,\}$ funkcji określonych dla nieskończonego progresywnego ciągu węzłów jest liniowo niezależny. Zatem zbiór funkcji B-sklejanych jest istotnie bazą przestrzeni funkcji sklejanych, którą rozpina.

<u>Własność 8 (minimalnego nośnika)</u>. *Nie istnieje niezerowa funkcja sklejana stopnia* $\leqslant n$ *z węzłami* $u_i, \ldots, u_{i+n+1}$, *która w otoczeniu każdego węzła o krotności* r *jest klasy* C^{n-r} *i której nośnik (domknięcie zbioru, w którym funkcja przyjmuje niezerowe wartości) jest podzbiorem właściwym przedziału* $[u_i, u_{i+n+1}]$, *tj. nośnika funkcji* N_i^n.



Dowód. Przypadki $n = 0$ i $n = 1$ są trywialne. Niech $n > 1$. Załóżmy, że wszystkie węzły są jednokrotne i przypuśćmy, że taka funkcja f (klasy $C^{n-1}(\mathbb{R})$) istnieje i (bez straty ogólności) przyjmuje wartość dodatnią w pewnym punkcie $x_0^{(0)} \in (u_i, u_{i+n+1})$. Ponieważ $f(u_i) = f(u_{i+n+1}) = 0$, w pewnym przedziale funkcja f jest rosnąca, a po nim następuje przedział, w którym f maleje. To oznacza, że istnieją punkty $x_0^{(1)}$ i $x_1^{(1)}$, takie że $u_i < x_0^{(1)} < x_1^{(1)} < u_{i+n+1}$ i $f'(x_0^{(1)}) > 0$, $f'(x_1^{(1)}) < 0$. Z ciągłości f' jest też $f'(u_i) = f'(u_{i+n+1}) = 0$.

Stosując to rozumowanie do kolejnych ciągłych pochodnych (rzędu $2, \ldots, n-1$), stwierdzamy, że $f^{(n-1)}$ ma co najmniej $n+1$ przedziałów wewnątrz $[u_i, u_{i+n+1}]$, w których jest na przemian rosnąca i malejąca. Ale pochodna rzędu $n-1$ funkcji sklejanej stopnia n jest funkcją sklejaną stopnia 1; pochodna rzędu n jest funkcją kawałkami stałą, która musi w co najmniej $n+1$ przedziałach przyjmować na przemian wartości dodatnie i ujemne. Jedynymi punktami nieciągłości funkcji $f^{(n)}$ mogą być węzły $u_i, \ldots, u_{i+n+1}$, które wyznaczają potrzebne $n+1$ przedziałów. Ich suma (po domknięciu) daje cały przedział $[u_i, u_{i+n+1}]$. Dowód dla przypadku, gdy pewne węzły mają krotność większą niż 1, zostawiam jako ćwiczenie. □

Jako ciekawostkę, bez dowodu, podam jeszcze

<u>Własność 9</u>.

$$\int_{\mathbb{R}} N_i^n(x)\,dx = \int_{u_i}^{u_{i+n+1}} N_i^n(x)\,dx = \frac{1}{n+1}(u_{i+n+1} - u_i).$$

Spośród funkcji B-sklejanych stopnia n określonych dla progresywnego ciągu węzłów $u_0, \ldots, u_N$ w przedziale $[u_k, u_{k+1}) \subset [u_n, u_{N-n})$ niezerowe wartości przyjmują tylko funkcje $N_{k-n}^n, \ldots, N_k^n$. Dla ustalonego x można je obliczyć za pomocą <u>algorytmu de Boora</u>, który realizuje obliczenie na podstawie wzoru Mansfielda-de Boora-Coxa. Oznaczmy

$$\alpha_i^{(j)} = \frac{x - u_i}{u_{i+j} - u_i}, \qquad \beta_{i+1}^{(j)} = \frac{u_{i+j+1} - x}{u_{i+j+1} - u_{i+1}} = 1 - \alpha_{i+1}^{(j)}.$$

Przed wykonaniem podanej niżej procedury należy znaleźć przedział $[u_k, u_{k+1})$, do którego należy liczba x; można w tym celu zastosować wyszukiwanie binarne, lub, jeśli węzły $u_n, \ldots, u_{N-n}$ są równoodległe, posłużyć się dzieleniem (tj. obliczyć $k = n + \lfloor (x - u_n)/(u_{n+1} - u_n) \rfloor$).

```
/* x ∈ [u_k, u_{k+1}) ⊂ [u_n, u_{N−n}) */
b[k] = 1;                                   /* N_k^0 = 1 */
for ( j = 1; j ⩽ n; j++ ) {
  β = (u_{k+1} − x)/(u_{k+1} − u_{k−j+1});    /* β = β_{k−j+1}^{(j)} */
  b[k − j] = β ∗ b[k − j + 1];              /* N_{k−j}^j = βN_{k−j+1}^{j−1} */
  for ( i = k − j + 1; i < k; i++ ) {
    α = 1 − β;                              /* α = α_i^{(j)} */
    β = (u_{i+j+1} − x)/(u_{i+j+1} − u_{i+1}); /* β = β_{i+1}^{(j)} */
    b[i] = α ∗ b[i] + β ∗ b[i + 1];         /* N_i^j = αN_i^{j−1} + βN_{i+1}^{j−1} */
  }
  b[k] ∗= (1 − β);                          /* N_k^j = αN_k^{j−1} */
}
/* b[i] = N_i^n(x) dla i = k − n, . . . , k */
```

Niech $x \in [u_k, u_{k+1}) \subset [u_n, u_{N-n})$. Wtedy mamy również

$$s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x) = \sum_{i=k-n}^{k} d_i N_i^n(x),$$

Jeśli $n > 0$, to podstawiając prawą stronę wzoru Mansfielda-de Boora-Coxa, dostajemy

$$\begin{aligned}
s(x) &= \sum_{i=k-n}^{k} d_i \Bigg(\underbrace{\frac{x-u_i}{u_{i+n}-u_i}}_{\alpha_i^{(n)}} N_i^{n-1}(x) + \underbrace{\frac{u_{i+n+1}-x}{u_{i+n+1}-u_{i+1}}}_{1-\alpha_{i+1}^{(n)}} N_{i+1}^{n-1}(x) \Bigg) \\
&= \sum_{i=k-n+1}^{k} \alpha_i^{(n)} d_i N_i^{n-1}(x) + \sum_{i=k-n}^{k-1} (1-\alpha_{i+1}^{(n)}) d_i N_{i+1}^{n-1}(x) \\
&= \sum_{i=k-n+1}^{k} \alpha_i^{(n)} d_i N_i^{n-1}(x) + \sum_{i=k-n+1}^{k} (1-\alpha_i^{(n)}) d_{i-1} N_i^{n-1}(x) \\
&= \sum_{i=k-n+1}^{k} \big(\alpha_i^{(n)} d_i + (1-\alpha_i^{(n)}) d_{i-1}\big) N_i^{n-1}(x).
\end{aligned}$$

Oznaczmy $d_i^{(1)} = \alpha_i^{(n)} d_i + (1-\alpha_i^{(n)}) d_{i-1}$. Jeśli $n > 1$, to powyższe przekształcenie możemy zastosować rekurencyjnie, otrzymując na końcu

$$s(x) = \sum_{i=k}^{k} d_i^{(n)} N_i^0(x) = d_k^{(n)}.$$

Na tym rachunku opiera się <u>algorytm de Boora</u> obliczania wartości funkcji sklejanej danej za pomocą współczynników w bazie B-sklejanej stopnia n:

/* $d_i^{(0)} = d_i$ *dla* $i = k-n, \dots, k$, $x \in [u_k, u_{k+1}) \subset [u_n, u_{N-n})$ */
<u>for</u> ($j = 1$; $j \leqslant n$; j++)
  <u>for</u> ($i = k-n+j$; $i \leqslant k$; i++) {
    $\alpha = (x-u_i)/(u_{i+n+1-j}-u_i)$;   /* $\alpha = \alpha_i^{(n+1-j)}$ */
    $d_i^{(j)} = (1-\alpha) * d_{i-1}^{(j-1)} + \alpha * d_i^{(j-1)}$;
  }
/* $d_k^{(n)} = s(x)$ */

Koszty obu algorytmów de Boora są rzędu n^2; dla funkcji niskich stopni, zazwyczaj stosowanych w praktyce, to są małe koszty. Ponadto można wykazać, że oba algorytmy mają bardzo dobre własności numeryczne (tj. niedokładności wyników będące skutkiem błędów zaokrągleń w implementacjach korzystających z arytmetyki zmiennopozycyjnej są małe). Jest to konsekwencją faktu, że dla $x \in [u_k, u_{k+1})$ wszystkie wartości przypisywane zmiennym α i β są liczbami z przedziału $[0, 1]$. Dzięki temu zaburzenia będące skutkami błędów zaokrągleń przenoszą się na końcowe wyniki z czynnikami co najwyżej 1.

Reprezentacja B-sklejana funkcji sklejanych jest bardzo elastyczna i wygodna w zastosowaniach, nie tylko w interpolacji, ale też w aproksymacji. Bardzo dużo zastosowań funkcje te mają w grafice komputerowej i systemach projektowania dla inżynierów. Zależnie od potrzeb, można dobrać stopień i węzły tak, aby otrzymać funkcje sklejane o odpowiednich własnościach.

Kubiczne funkcje interpolacyjne w reprezentacji B-sklejanej

Konstrukcja B-sklejanej reprezentacji kubicznej funkcji interpolacyjnej klasy C^2, tj. obliczenie współczynników d_i, polega na rozwiązaniu układu równań liniowych (podobnie jak dla reprezentacji Hermite'a). W tym przypadku przyjmujemy, że wartości y_i funkcji są zadane w węzłach $u_3, \dots, u_{N-3}$, tworzących ciąg rosnący. Ponieważ w każdym z tych węzłów tylko trzy funkcje B-sklejane są niezerowe, mamy równania

$$N^3_{i-3}(u_i)d_{i-3} + N^3_{i-2}(u_i)d_{i-2} + N^3_{i-1}(u_i)d_{i-1} = y_i.$$

Wartości funkcji B-sklejanych w węzłach możemy obliczyć za pomocą algorytmu de Boora. Do tych równań (dla $i = 3, \dots, N-3$) należy dołączyć jeszcze dwa równania opisujące warunki brzegowe (np. prowadzące do otrzymania naturalnej funkcji sklejanej s, tj. spełniającej warunek $s''(u_3) = s''(u_{N-3}) = 0$). Dla dowolnych sensownych warunków brzegowych równania można przekształcić tak, aby otrzymać układ równoważny z macierzą trójdiagonalną.

Można też skonstruować kubiczną sklejaną funkcję okresową klasy $C^2(\mathbb{R})$ o okresie T. Ciąg węzłów musi być rosnący i jeśli $T = u_{N-3} - u_3$, to trzeba przyjąć $u_{N-6+i} = u_i + T$ dla $i = 1, \dots, 5$. Okresowa funkcja sklejana dla takiego ciągu ma współczynniki $d_0 = d_{N-6}$, $d_1 = d_{N-5}$, $d_2 = d_{N-4}$. Warunki interpolacyjne $s(u_i) = y_i$ zadajemy w węzłach $u_3, \dots, u_{N-3}$, przy czym przyjmujemy $y_{N-3} = y_3$. Na tej podstawie otrzymuje się układ $N-6$ równań liniowych, którego rozwiązaniem są współczynniki $d_0, \dots, d_{N-7}$ okresowej sklejanej funkcji interpolacyjnej, z macierzą cykliczną trójdiagonalną.

Twierdzenie Schoenberga-Whitney

Przypomnijmy, że słowo „węzły” ma co najmniej dwa różne znaczenia. Po pierwsze, węzły interpolacyjne, czyli punkty, w których zadajemy wartości funkcji. Drugie znaczenie to węzły funkcji sklejanej, czyli punkty rozgraniczające przedziały, w których funkcja jest (a dokładniej, może być) opisana za pomocą różnych wielomianów. Do tej pory wybieraliśmy węzły interpolacyjne pokrywające się z węzłami funkcji sklejanej, ale możemy dopuścić inny ich wybór. Trzeba jednak wiedzieć, jaki wybór jest dopuszczalny, aby rozwiązanie zadania interpolacyjnego Lagrange'a istniało.

Oznaczmy zatem symbolami $u_0, \ldots, u_N$ węzły funkcji sklejanej, a konkretniej progresywny (tj. niemalejący i spełniający warunek $u_i < u_{i+n+1}$ dla $i = 0, \ldots, N-n-1$) ciąg węzłów, których użyjemy do określenia funkcji B-sklejanych stopnia n. Liczba tych funkcji to $N-n$. Węzły interpolacyjne oznaczymy symbolami $v_0, \ldots, v_{N-n-1}$. Liczba warunków interpolacyjnych, które nakładamy, jest równa wymiarowi przestrzeni funkcji sklejanych rozpiętej przez nasze funkcje B-sklejane, dzięki czemu warunki brzegowe są zbędne. Założymy, że węzły interpolacyjne są ponumerowane tak, aby tworzyły ciąg rosnący (jest jasne, że węzły interpolacyjne muszą być parami różne).

Twierdzenie Schoenberga-Whitney. *Funkcja sklejana stopnia n, oparta na ciągu węzłów $u_0, \ldots, u_N$ i przyjmująca zadane wartości $y_0, \ldots, y_{N-n-1}$ odpowiednio w punktach $v_0, \ldots, v_{N-n-1}$, takich że $v_0 < v_1 < \cdots < v_{N-n-1}$, istnieje i jest jednoznacznie określona wtedy i tylko wtedy, gdy $N_i^n(v_i) \neq 0$ dla $i = 0, \ldots, N-n-1$.*

Dowód tego twierdzenia polega na wykazaniu że odpowiednia macierz Vandermonde'a (tj. macierz V o współczynnikach $a_{ij} = N_j^n(v_i)$) jest nieosobliwa wtedy i tylko wtedy, gdy warunek $N_i^n(v_i) \neq 0$ dla $i = 0, \ldots, N-n-1$ jest spełniony. Łatwo jest dowieść konieczności tego warunku (polecam to jako ćwiczenie), natomiast dowód, że jest to też warunek dostateczny, jest żmudny (nie polecam, chyba że ktoś sam tego chce). Twierdzenie rozstrzyga problem z punktu widzenia algebry, ale nie gwarantuje, że zadanie interpolacji jest dobrze uwarunkowane (i np. macierz V jest dobrze uwarunkowana). Aby tak było, dla każdego $i \in \{0, \ldots, N-n-1\}$ węzeł interpolacyjny v_i powinien leżeć w pobliżu punktu, w którym odpowiadająca mu funkcja B-sklejana N_i^n osiąga wartość maksymalną.

Przypuśćmy, że dane są węzły interpolacyjne, $v_0, \ldots, v_{N-n-1}$, ustawione w ciąg rosnący. Jeśli n jest nieparzyste, to dobrym (ale *nie jedynym dobrym*) wyborem jest przyjęcie węzłów funkcji sklejanej $u_0 = \cdots = u_n = v_0$, oraz $u_i = v_{i-(n+1)/2}$ dla $i = n+1, \ldots, N-n-1$, i $u_{N-n} = \cdots = u_N = v_{N-n-1}$.

Dla parzystego n można wybrać $u_0 = \cdots = u_n = v_0$, i $u_i = (v_{i-n/2-1} + v_{i-n/2})/2$ dla $i = n+1, \ldots, N-n-1$, oraz $u_{N-n} = \cdots = u_N = v_{N-n-1}$.

Wspomniana wyżej macierz V jest wstęgowa, a dokładniej, ma w każdym wierszu co najwyżej $n+1$ niezerowych współczynników. Dzięki temu znalezienie sklejanej funkcji interpolacyjnej może być bardzo mało kosztowne (koszt eliminacji Gaussa w tym przypadku to $O(Nn^2)$ operacji).

Zadania i problemy

1. Napisz układ równań, którego rozwiązaniem są liczby $b_0, \dots, b_N$, będące wartościami pochodnej naturalnej kubicznej sklejanej funkcji interpolacyjnej w węzłach $0, \dots, N$. Rozwiąż ten układ dla $N = 4$ i danych wartości funkcji $a_0 = \dots = a_3 = 0$, $a_4 = 1$.

   Podaj oszacowanie wskaźnika uwarunkowania macierzy tego układu w normie drugiej, na podstawie twierdzenia Gerszgorina.

2. Napisz układ równań dla okresowej kubicznej funkcji interpolacyjnej (warunki interpolacyjne, tj. wartości funkcji, zadane są w węzłach funkcji sklejanej), z węzłami równoodległymi $u_i = i$ dla $i = 0, \dots, 5$, i rozwiąż ten układ.

3. Wyprowadź równania opisujące warunki brzegowe dla kubicznej funkcji sklejanej s w reprezentacji Hermite'a, określone przez zadanie wartości pochodnych drugiego rzędu, $c_0 = s''(u_0)$ i $c_N = s''(u_N)$.

   (dla sprawdzenia: $2h_0b_0 + h_0b_1 = 3(a_1 - a_0) - \dfrac{h_0^2c_0}{2}$,

   $h_{N-1}b_{N-1} + 2h_{N-1}b_N = 3(a_N - a_{N-1}) + \dfrac{h_{N-1}^2c_N}{2}$).

4. Wyprowadź równania opisujące warunki brzegowe Bessela dla kubicznej funkcji sklejanej s: ma być $s'(u_0) = b_0 = w_L'(u_0)$ oraz $s'(u_N) = b_N = w_P'(u_N)$, gdzie w_L i w_P to odpowiednio wielomiany interpolacyjne Lagrange'a skonstruowane dla pierwszych trzech i ostatnich trzech węzłów: $w_L(u_i) = a_i$ dla $i = 0, 1, 2$ oraz $w_P(u_i) = a_i$ dla $i = N-2, N-1, N$.

   (dla sprawdzenia: $(h_0 + h_1)b_0 = \left(2 + \dfrac{h_1}{h_0}\right)(a_1 - a_0) - \dfrac{h_0}{h_1}(a_2 - a_1)$,

   $(h_{N-2} + h_{N-1})b_N = \left(2 + \dfrac{h_{n-2}}{h_{N-1}}\right)(a_N - a_{N-1}) - \dfrac{h_{N-1}}{h_{N-2}}(a_{N-1} - a_{N-2})$).

5. Wyprowadź równania opisujące warunki brzegowe, które narzucają pochodną trzeciego rzędu równą 0 dla wielomianów opisujących kubiczną funkcję sklejaną w przedziałach $[u_0, u_1]$ i $[u_{N-1}, u_N]$.

   (dla sprawdzenia: $h_0b_0 + h_1b_1 = 2(a_1 - a_0)$,

   $h_{N-1}b_{N-1} + h_{N-1}b_N = 2(a_N - a_{N-1})$).

6. Wyprowadź równania opisujące warunki *not a knot* dla kubicznej interpolacyjnej funkcji sklejanej klasy C^2: węzły interpolacyjne u_1 oraz u_{N-1} mają nie być węzłami funkcji sklejanej. Wskazówka: wystarczy zapewnić, że wielomiany p_0 i p_1 opisujące funkcję s w przedziałach $[u_0, u_1]$ i $[u_1, u_2]$, mają taką samą pochodną trzeciego rzędu, podobnie wielomiany p_{N-2} i p_{N-1}.

(dla sprawdzenia: $h_1^2 b_0 + (h_1^2 - h_0^2) b_1 - h_0^2 b_2 = \frac{2h_1^2}{h_0}(a_1 - a_0) - \frac{2h_0^2}{h_1}(a_2 - a_1)$,

$$h_{N-1}^2 b_{N-2} + (h_{N-1}^2 - h_{N-2}^2) b_{N-1} - h_{N-2}^2 b_N = \frac{2h_{N-1}^2}{h_{N-2}}(a_{N-1} - a_{N-2}) - \frac{2h_{N-2}^2}{h_{N-1}}(a_N - a_{N-1})\ ).$$

W równaniach opisujących te warunki brzegowe występują niewiadome b_2 i b_{N-2}, przez co macierz układu równań opisujących ciągłość pochodnych drugiego rzędu w węzłach i te warunki nie jest trójdiagonalna. Użyj równań ciągłości s'' w węzłach u_1 i u_{N-1} do wyeliminowania tych niewiadomych.

(dla sprawdzenia: $h_1 b_0 + (h_0 + h_1) b_1 = \frac{h_1(3h_0 + 2h_1)}{h_0(h_0 + h_1)}(a_1 - a_0) + \frac{h_0^2}{h_1(h_0 + h_1)}(a_2 - a_1)$,

$$(h_{N-2} + h_{N-1}) b_{N-1} + h_{N-2} b_N = \frac{h_{N-1}^2}{h_{N-2}(h_{N-2} + h_{N-1})}(a_{N-1} - a_{N-2}) + \frac{h_{N-2}(2h_{N-2} + h_{N-1})}{h_{N-1}(h_{N-2} + h_{N-1})}(a_N - a_{N-1})\ ).$$

7. Wyprowadź równania opisujące warunki brzegowe „animacyjne”: należy zadać wartości pochodnej pierwszego rzędu w węzłach u_0 i u_N i pochodną drugiego rzędu w tych węzłach *zamiast* wartości funkcji w węzłach u_1 i u_{N-1}. Jeśli otrzymana w ten sposób funkcja opisuje położenie obiektu w czasie, a narzucone pochodne są równe 0, to początkowo nieruchomy obiekt płynnie rozpędzi się, przejdzie przez zadane położenia w chwilach $u_2, \dots, u_{N-2}$ i łagodnie zahamuje w położeniu końcowym w chwili u_N.

   (dla sprawdzenia: $3a_1 - h_0 b_1 = 3a_0 + 2h_0 b_0 + \frac{h_0^2}{2} c_0$,

   $$h_{N-1} b_{N-1} + 3a_{N-1} = 3a_N - 2h_{N-1} b_N + \frac{h_{N-1}^2}{2} c_N.$$

   Liczby a_1 i a_{N-1}, tj. wartości funkcji w węzłach u_1 i u_{N-1} są tu niewiadome.

8. Wyprowadź alternatywną postać równania umożliwiającego skonstruowanie kubicznej sklejanej funkcji interpolacyjnej klasy C^2, z niewiadomymi wartościami c_{i-1}, c_i, c_{i+1} pochodnej drugiego rzędu odpowiednio w węzłach u_{i-1}, u_i i u_{i+1}.

   Wskazówka: Pochodne drugiego rzędu wielomianów p_{i-1} i p_i, opisujących funkcję sklejaną w przedziałach $[u_{i-1}, u_i]$ oraz $[u_i, u_{i+1}]$, są wielomianami pierwszego stopnia, interpolującymi odpowiednio liczby c_{i-1}, c_i oraz c_i, c_{i+1}. Znajdź te wielomiany i scałkuj dwukrotnie, dobierając stałe całkowania tak, aby otrzymane wyrażenia opisywały wielomiany p_{i-1} i p_i i w szczególności przyjmowały w węźle u_i (daną) wartość a_i, a ich pochodne pierwszego rzędu przyjmowały (niewiadomą) wartość b_i. Następnie oblicz $a_{i-1} = p_{i-1}(u_{i-1})$ oraz $a_{i+1} = p_i(u_{i+1})$ i stąd otrzymaj dwa wyrażenia na b_i. Postaw znak równości między tymi wyrażeniami i uporządkuj otrzymane równanie.

9. Wykaż, że naturalna kubiczna funkcja sklejana klasy $C^2(\mathbb{R})$ z węzłami $u_0, \ldots, u_N$ może być przedstawiona w postaci

$$s(x) = p_{-1}(x) + \sum_{i=0}^{N} c_i(x - u_i)_+^3,$$

gdzie $\sum_{i=0}^{N} c_i = \sum_{i=0}^{N} c_i u_i = 0$ oraz p_{-1} jest wielomianem stopnia co najwyżej 1.

10. Napisz układ równań opisujący kubiczną naturalną funkcją sklejaną s klasy $C^2(\mathbb{R})$ z węzłami $u_0 < \cdots < u_4$, taką że $s(u_0) = s'(u_0) = s(u_4) = s'(u_4) = 0$, $s(u_2) = c \neq 0$; przedstaw funkcję s w postaci z poprzedniego zadania. Wykaż, że z dokładnością do czynnika stałego jest to funkcja B-sklejana N_0^3.

11. Znajdź wielomiany stopnia 5, będące rozwiązaniami sześciu zadań interpolacyjnych Hermite'a, dla dwóch węzłów trzykrotnych 0 i 1, analogicznych do zadań, których rozwiązaniami są wielomiany trzeciego stopnia z wykładu. Podaj sposób określenia interpolacyjnej funkcji sklejanej stopnia 5 dla węzłów $u_0, \ldots, u_N$ przy użyciu tych wielomianów.
Wskazówka: Wystarczy znaleźć trzy wielomiany przez rozwiązanie zadań interpolacyjnych. Pozostałe trzy są do nich symetryczne, tj. $H_{10}(x) = H_{00}(1 - x)$, $H_{11}(x) = -H_{01}(1 - x)$, $H_{12}(x) = H_{02}(1 - x)$.

12. Udowodnij, że jeśli węzły $u_0, \ldots, u_N$ tworzą ciąg arytmetyczny (tj. są równoodległe), to funkcje B-sklejane ustalonego stopnia n dla tego ciągu węzłów są identyczne z dokładnością do przesunięcia argumentu.

13. Znajdź wielomiany opisujące funkcje B-sklejane N_i^2 oraz N_i^3, określone dla węzłów będących kolejnymi liczbami naturalnymi.
Wskazówka: Na podstawie poprzedniego zadania, dla każdego $i \in \mathbb{Z}$ oraz $x \in \mathbb{R}$ jest $N_i^n(x) = N_0^n(x - i)$, wystarczy zatem znaleźć wielomiany opisujące tylko jedną funkcję B-sklejaną ustalonego stopnia.

14. Napisz układ równań, którego rozwiązaniem jest wektor współczynników $d_0, \ldots, d_{N-4}$ naturalnej kubicznej sklejanej funkcji interpolacyjnej w bazie $\{N_0^3, \ldots, N_{N-4}^3\}$, przy czym funkcje B-sklejane, z których składa się ta baza, są określone za pomocą ciągu węzłów równoodległych: $u_i = i$ dla $i = 0, \ldots, N$ (warunki interpolacyjne, tj. wartości funkcji, są określone w węzłach $3, \ldots, N - 3$).

15. Napisz układ równań, którego rozwiązaniem jest wektor współczynników (w bazie B-sklejanej) funkcji sklejanej stopnia 2, przy czym węzłami funkcji sklejanej są liczby naturalne $0, 1, \ldots, N$, a węzłami interpolacyjnymi są liczby $2, 2\frac{1}{2}, 3\frac{1}{2}, \ldots, N - 2\frac{1}{2}, N - 2$ (z wyjątkiem węzłów skrajnych odległości między kolejnymi węzłami są równe 1).

16. Na podstawie wzoru opisującego pochodną funkcji B-sklejanej wyprowadź wzór opisujący pochodną funkcji sklejanej $s(x) = \sum_{i=0}^{N-n-1} d_i N_i^n(x)$.

Interpolacja trygonometryczna

Trygonometryczne zadanie interpolacji Lagrange'a

Def. Wielomian trygonometryczny stopnia n jest to funkcja o postaci

$$w(t) = a_0 + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt). \tag{*}$$

Wielomiany trygonometryczne występują w różnych zastosowaniach, zwłaszcza takich, w których pojawiają się funkcje okresowe. Często powstają z obcięcia szeregów Fouriera, tj. szeregów opisanych wzorem podobnym do wzoru (*), w którym zamiast n jest ∞.

Wzór (*) opisuje funkcję o okresie 2π. Aby otrzymać funkcję o dowolnym okresie $T > 0$, można dokonać zamiany zmiennych. Będziemy mieli wtedy

$$f(x) = a_0 + \sum_{k=1}^{n} (a_k \cos kt + b_k \sin kt),$$

gdzie $t = 2\pi(x - x_0)/T$, dla dowolnie wybranego x_0.

Trygonometryczne zadanie interpolacyjne Lagrange'a polega na znalezieniu wielomianu trygonometrycznego stopnia n, którego wartości $f_0, \ldots, f_{2n}$ są określone w $2n+1$ węzłach interpolacyjnych, $x_0, \ldots, x_{2n} \in \mathbb{R}$.

Twierdzenie. *Warunek konieczny i dostateczny istnienia i jednoznaczności rozwiązania tego zadania dla dowolnych liczb* $f_0, \ldots, f_{2n} \in \mathbb{R}$ *jest taki, że* $(x_j - x_k)/T \notin \mathbb{Z}$ *dla* $j \neq k$.

Dowód. Z uwagi na to, że rozwiązanie zadania musi być funkcją okresową, konieczność tego warunku jest oczywista. To, że ten warunek jest dostateczny, wystarczy udowodnić dla przypadku szczególnego $T = 2\pi$. Dla węzłów interpolacyjnych x_j i wartości funkcji f_j podanych dla tych węzłów, określamy liczby zespolone $z_j = e^{ix_j}$ oraz $h_j = z_j^n f_j$. Jeśli węzły $x_0, \ldots, x_{2n}$ spełniają rozważany warunek, to liczby z_j są parami różne. Jak wiemy, zadanie interpolacyjne Lagrange'a, tj. wyznaczenie wielomianu $h(z)$ stopnia co najwyżej $2n$, takiego że $h(z_j) = h_j$ dla $j = 0, \ldots, 2n$, ma rozwiązanie[26]. Możemy je przedstawić w bazie potęgowej w taki sposób: $h(z) = \sum_{k=0}^{2n} c_{k-n} z^k$.

[26] Można je znaleźć za pomocą algorytmu różnic dzielonych; pod tym względem zadanie interpolacji wielomianowej dla zespolonych węzłów i wartości funkcji nie różni się od przypadku rzeczywistego.

Zespolona funkcja wymierna

$$g(z) \stackrel{\text{def}}{=} \sum_{k=-n}^{n} c_k z^k$$

w węzłach z_j przyjmuje wartości f_j, i jest tylko jedna taka funkcja o tej postaci (bo liczby $c_{-n}, \ldots, c_n$ są określone jednoznacznie przez warunki interpolacyjne nałożone na wielomian h).

Niech $\hat{g}(z) \stackrel{\text{def}}{=} \overline{g(z)}$. Dla każdego $z \in \mathbb{C}$ takiego że $|z| = 1$, w tym dla każdego z_j, jest $\bar{z} = \frac{1}{z}$ i stąd

$$\hat{g}(z) = \overline{g(z)} = \sum_{k=-n}^{n} \bar{c}_k \bar{z}^k = \sum_{k=-n}^{n} \bar{c}_k z^{-k} = \sum_{k=-n}^{n} \bar{c}_{-k} z^k.$$

Ponieważ liczby f_j są rzeczywiste, zachodzą równości $\hat{g}(z_j) = f_j = g(z_j)$ dla każdego j. Spełniająca te warunki interpolacyjne funkcja $\hat{g}(z)$ też jest tylko jedna (tj. liczby $\bar{c}_{-n}, \ldots, \bar{c}_n$ są jednoznacznie określone przez te warunki), ale to oznacza, że $c_{-k} = \bar{c}_k$ dla $k = -n, \ldots, n$, a stąd wynika, że w zbiorze $\{\, z \in \mathbb{C}\colon |z| = 1 \,\}$ funkcje $g(z)$ i $\hat{g}(z)$ są identyczne.

Zatem, funkcja $w(t) = g(e^{it}) = \overline{g(e^{it})}$ ma dla każdego $t \in \mathbb{R}$ wartość rzeczywistą i spełnia warunki $w(x_j) = f_j$ dla $j = 0, \ldots, 2n$. Tak więc

$$\begin{aligned}
w(t) = \sum_{k=-n}^{n} c_k e^{ikt} &= c_0 + \sum_{k=1}^{n} (c_k e^{ikt} + \bar{c}_k e^{-ikt}) = \\
& c_0 + \sum_{k=1}^{n} \big(c_k(\cos kt + i \sin kt) + \bar{c}_k(\cos kt - i \sin kt)\big) = \\
& c_0 + \sum_{k=1}^{n} \big((c_k + \bar{c}_k)\cos kt + i(c_k - \bar{c}_k)\sin kt\big) = \\
& c_0 + \sum_{k=1}^{n} (2\,\mathrm{Re}\, c_k \cos kt - 2\,\mathrm{Im}\, c_k \sin kt).
\end{aligned}$$

Mamy stąd współczynniki wielomianu trygonometrycznego (*): $a_0 = \mathrm{Re}\, c_0$ (jest $\mathrm{Im}\, c_0 = 0$), oraz $a_k = 2\,\mathrm{Re}\, c_k$ i $b_k = -2\,\mathrm{Im}\, c_k$ dla $k > 0$. □

W praktycznych zastosowaniach najczęściej wybiera się węzły $x_0, \ldots, x_{2n}$, które dzielą przedział $[x_0, x_0 + T)$ (o długości okresu T interpolowanej funkcji) na części o jednakowych długościach. Zamiast bezpośrednio rozwiązywać zadania interpolacji trygonometrycznej, zwykle sprowadza się problem do konstrukcji zespolonego wielomianu algebraicznego, w sposób podobny do użytego w powyższym dowodzie.

Dyskretna transformata Fouriera

Def. Dyskretną transformatą Fouriera ciągu zespolonego $(a_k)_{k\in\mathbb{Z}}$ o okresie n (tj. spełniającego warunek $a_{k+n} = a_k$ dla każdego $k \in \mathbb{Z}$) jest ciąg zespolony $(b_j)_{j\in\mathbb{Z}}$ określony wzorem

$$b_j = \sum_{k=0}^{n-1} a_k e^{-2\pi ijk/n}.$$

Odwrotną dyskretną transformatą Fouriera ciągu $(a_k)_{k\in\mathbb{Z}}$ nazywamy ciąg $(c_j)_{j\in\mathbb{Z}}$ określony wzorem

$$c_j = \frac{1}{n}\sum_{k=0}^{n-1} a_k e^{2\pi ijk/n}.$$

Ciągi $(b_j)_{j\in\mathbb{Z}}$ i $(c_j)_{j\in\mathbb{Z}}$ są okresowe o okresie n. Oba przekształcenia zdefiniowane wyżej są liniowe i każde z nich jest jest odwrotnością tego drugiego, co uzasadnia nazwę. Mamy bowiem

$$d_l = \frac{1}{n}\sum_{j=0}^{n-1}\Bigl(\sum_{k=0}^{n-1} a_k e^{-2\pi ijk/n}\Bigr)e^{2\pi ilj/n} = \frac{1}{n}\sum_{k=0}^{n-1} a_k \sum_{j=0}^{n-1} e^{2\pi i(l-k)j/n}.$$

Jeśli $k = l$, to $e^{2\pi i(l-k)j/n} = e^0 = 1$, zaś jeśli $k \neq l$, to liczby $e^{2\pi i(l-k)j/n}$ są pierwiastkami zespolonymi z 1; ich suma dla $j \in \{0, \ldots, n-1\}$ jest równa 0. Stąd wynika, że $d_l = a_l$.

Interpolacja trygonometryczna i dyskretna transformata Fouriera występuje w wielu problemach związanych z analizą, transmisją i przetwarzaniem sygnałów (np. akustycznych lub obrazów), a także w rozwiązywaniu równań różniczkowych.

Algorytm FFT

Możemy zauważyć, że ciąg $(b_j)_{j\in\mathbb{Z}}$, który jest transformatą ciągu $(a_k)_{k\in\mathbb{Z}}$, składa się z wartości wielomianu stopnia $n-1$ o współczynnikach $a_0, \ldots, a_{n-1}$ w punktach $e^{-2\pi ij/n}$, $j = 0, \ldots, n-1$. Dyskretną transformatę Fouriera można wyznaczyć za pomocą schematu Hornera; wyznaczenie pełnej transformaty kosztowałoby wtedy $n^2 - n$ mnożeń i dodawań zespolonych. Okazuje się, że można to zadanie rozwiązać kosztem $\Theta(n(p_1 + \cdots + p_r))$ działań, gdzie $p_1, \ldots, p_r$ są liczbami pierwszymi, takimi że $n = p_1 \cdot \ldots \cdot p_r$. Odkrycia tego dokonali w 1952 r. Cooley i Tukey.

Zauważmy, że ciąg okresowy $(a_k)_{k\in\mathbb{Z}}$ o okresie 1 jest ciągiem stałym i jest on identyczny ze swoją dyskretną transformatą Fouriera. Dalej, przypuśćmy, że liczba n jest podzielna przez $p > 1$. Oznaczmy $w_j = e^{-2\pi i j/n}$. Wtedy wzór definiujący dyskretną transformatę Fouriera można przedstawić w postaci

$$b_j = \sum_{k=0}^{n/p-1} a_{pk} w_j^{pk} + w_j \sum_{k=0}^{n/p-1} a_{pk+1} w_j^{pk} + \cdots + w_j^{p-1} \sum_{k=0}^{n/p-1} a_{pk+p-1} w_j^{pk}.$$

Podzieliliśmy tu ciąg $a_0, \ldots, a_{n-1}$ na podciągi n/p-elementowe, wybierając do każdego z nich co p-ty element. Możemy dalej zauważyć, że sumy mnożone przez kolejne potęgi liczby w_j są wyrażeniami opisującymi transformaty tych podciągów, a dokładniej ich obustronnie nieskończonych rozszerzeń o okresie n/p. Obliczenie dyskretnej transformaty Fouriera dla ciągu o okresie n może być zatem wykonane przez następujący algorytm rekurencyjny:

- Jeśli $n = 1$, to przyjmij $b_0 = a_0$ (dla $n = 1$ przekształceniu poddajemy ciąg stały, którego obrazem jest ten sam ciąg).
- Jeśli n jest liczbą pierwszą, to zastosuj wzór podany jako definicja dyskretnej transformaty Fouriera i użyj schematu Hornera.
- Jeśli $n > 1$ jest podzielne przez liczbę pierwszą $p < n$, to podziel ciąg na p podciągów (zgodnie z opisem wyżej), oblicz transformaty tych podciągów i „scal” je, stosując wzór podany wyżej i schemat Hornera.

Wzór opisujący transformatę odwrotną może być przekształcony podobnie; zamiast $w_j = (\cos\frac{2\pi j}{n}, -\sin\frac{2\pi j}{n})$ występuje w nim liczba $\overline{w_j} = (\cos\frac{2\pi j}{n}, \sin\frac{2\pi j}{n})$. Możemy zatem użyć takiego samego algorytmu, zostawiając mnożenie wyniku działania procedury rekurencyjnej przez czynnik $\frac{1}{n}$ na sam koniec. Koszt algorytmu w istotny sposób zależy od możliwości rozłożenia liczby n na czynniki.

Algorytm jest najbardziej efektywny, jeśli liczba n jest potęgą liczby 2 i często określenie FFT (od angielskiego *Fast Fourier Transform*) dotyczy takiego wariantu algorytmu. Zbadamy go dokładniej. Dla parzystej liczby n transformatę otrzymamy przez „scalenie” transformat dwóch podciągów, złożonych odpowiednio z elementów parzystych i nieparzystych ciągu danego. Transformaty te oznaczymy symbolami $(p_j)_{j\in\mathbb{Z}}$ i $(q_j)_{j\in\mathbb{Z}}$. Przypomnijmy, że transformaty te są ciągami obustronnie nieskończonymi, o okresie $n/2$, reprezentowanymi przez podciągi $p_0, \ldots, p_{n/2-1}$ i $q_0, \ldots, q_{n/2-1}$. Możemy napisać

$$b_j = \sum_{k=0}^{n/2-1} a_{2k} w_j^{2k} + w_j \sum_{k=0}^{n/2-1} a_{2k+1} w_j^{2k} = p_j + w_j q_j.$$

Podstawiając $j + n/2$ w miejsce j, i biorąc pod uwagę, że $w_{j+n/2} = e^{-2\pi i(j+n/2)/n} = e^{-2\pi ij/n}e^{-2\pi in/(2n)} = -w_j$ oraz $w_{j+n/2}^{2k} = w_j^{2k}$, dostajemy

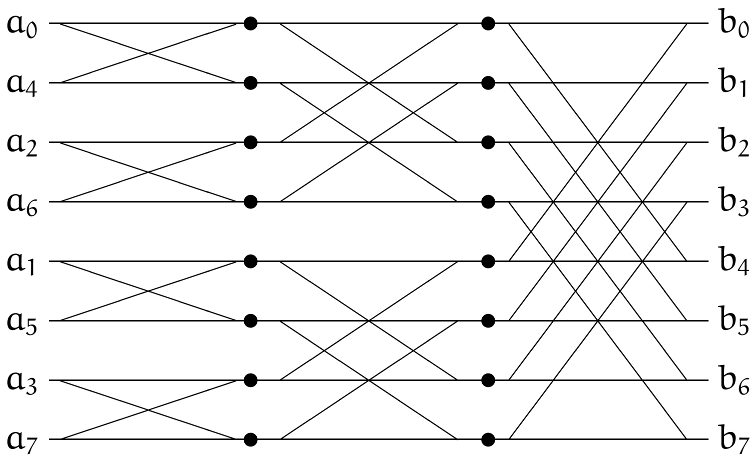
$$b_{j+n/2} = \sum_{k=0}^{n/2-1} a_{2k}w_{j+n/2}^{2k} + w_{j+n/2}\sum_{k=0}^{n/2-1} a_{2k+1}w_{j+n/2}^{2k} = p_j - w_jq_j.$$

Implementacja algorytmu FFT w postaci procedury rekurencyjnej jest następująca:

```
void rFFT ( int n, complex a[] )
{
  complex *p, *q, u, w, t; int j;

  if ( n > 1 ) {
    p = malloc ( n*sizeof(complex) );
    q = &p[n/2];
    for ( j = 0; j < n/2; j++ ) {
      p[j] = a[2*j];
      q[j] = a[2*j+1];
    }
    rFFT ( n/2, p );
    rFFT ( n/2, q );
    u = 1;
    w = e^{-2πi/n};
    for ( j = 0; j < n/2; j++ ) {
      t = u*q[j];
      a[j] = p[j] + t;
      a[j+n/2] = p[j] - t;
      u = u*w;
    }
    free ( p );
  }
} /*rFFT*/
```

Oglądając tę implementację, widzimy, że choć procedura umieszcza transformatę w tej samej tablicy, w której są początkowo dane liczby $a_0, \ldots, a_{n-1}$, potrzebuje ona sporo pamięci dodatkowej (w rzeczywistości potrzeba dodatkowych tablic o sumarycznej długości $2n$). Można jednak zaprojektować taką implementację, która wszystkie obliczenia wykonuje „w miejscu”, tj. która oprócz tablicy z danym ciągiem, który należy zastąpić przez jego transformatę, potrzebuje tylko niewielkiej ustalonej liczby zmiennych prostych. Aby otrzymać taką procedurę, nierekurencyjną i dodatkowo oszczędzającą pewne działania, przyjrzymy się „przepływowi danych", to znaczy zbadamy, od których współczynników zależą transformaty obliczane „po drodze". Dla $n = 8$ „przepływ danych" jest przedstawiony na rysunku (najlepiej go oglądać od prawej do lewej strony).



Krawędzie łączą dane z wynikami, tj. każda liczba (z wyjątkiem danych) jest obliczana na podstawie liczb znajdujących się w kolumnie na lewo od niej, połączonych z nią kreskami. Widzimy, że w każdym przypadku obliczenie polega na zastąpieniu pary liczb przez inną parę, obliczoną tylko na jej podstawie (i dana para liczb nie jest do niczego innego potrzebna). Jeśli zatem ustawimy dane wejściowe w odpowiedniej kolejności, to można całe obliczenie wykonać bez potrzeby rezerwowania dodatkowej tablicy.

Ostatnia transformata powstaje z transformat podciągów „parzystego" i „nieparzystego". Każda z tych dwóch transformat jest obliczana na podstawie transformat „parzystego" i „nieparzystego" podciągu odpowiedniego podciągu itd.; zatem ogólna reguła porządkowania danych wejściowych polega na ustawieniu ich *w kolejności odwróconych bitów*. Jeśli indeks j danego współczynnika a_j przedstawimy w układzie dwójkowym, przy użyciu $l = \log_2 n$ cyfr dwójkowych (bitów), to indeks miejsca w tablicy, na którym ma się on znaleźć, otrzymamy wypisując te bity w odwrotnej kolejności. Procedura FFT, która realizuje to obliczenie, ma postać

```
void FFT ( int n, complex a[] )
{
  complex t, u, w;
  int i, j, k, l, m, p;

  l = log2n; m = n/2;
```

```
       /* przestawianie danych w tablicy */
  for ( i = 1, j = m; i < n-1; i++ ) {
    if ( i < j ) przestaw ( &a[i], &a[j] );
    k = m;
    while ( k <= j ) { j -= k; k /= 2; }
    j += k;
  }
       /* obliczanie transformaty */
  for ( k = 1; k <= l; k++ ) {
    m = 2^k; p = m/2;
    u = 1; w = e^{-πi/p};
    for ( j = 0; j < p; j++ ) {
      i = j;
      do {
        t = a[i+p]*u;
        a[i+p] = a[i]-t; a[i] = a[i]+t;
        i += m;
      } while ( i <= n );
      u *= w;
    }
  }
} /*FFT*/
```

Algorytm ten opublikowali w 1965 r. Cooley, Lewis i Welch. Pierwsza pętla, `for ( i = ... ) ...`, dokonuje przestawienia elementów w tablicy zgodnie z kolejnością odwróconych bitów. Kolejne przebiegi drugiej pętli, `for ( k = ... ) ...`, mają na celu obliczenie $n/2$ transformat podciągów o okresie 2, $n/4$ transformat podciągów o okresie 4, itd. Liczba $e^{-2\pi i/n}$ (wartość zmiennej `w`) i jej potęgi, czyli liczby w_j (kolejne wartości zmiennej `u`) są obliczane tylko raz dla wszystkich transformat podciągów o tym samym okresie. W każdym przebiegu pętli `for ( j = ... ) ...` obliczane są pary współczynników o numerach j oraz $j+p$ we wszystkich transformatach podciągów o okresie $m = 2p$, ponieważ pętla najbardziej wewnętrzna (`do`...`while`) przebiega przez wszystkie te transformaty.

Można udowodnić, że algorytm FFT, także w wersji ogólnej (dla dowolnego n), jest numerycznie stabilny, tj. istnieje stała K (zależna od n), taka że współczynniki $\tilde{b}_j$ obliczone przy użyciu arytmetyki zmiennopozycyjnej przybliżają dokładne współczynniki b_j dyskretnej transformaty Fouriera z błędem spełniającym nierówność

$$\max_j |\tilde{b}_j - b_j| \leqslant K\nu \max_j |b_j|,$$

gdzie $\nu = 2^{-t}$. Dla liczby n będącej potęgą 2 można przyjąć

$$K = \bigl(\sqrt{2}\,\log_2 n + (\log_2 n - 1)(3 + 2\varepsilon)\bigr)\sqrt{n},$$

gdzie ε jest oszacowaniem błędu bezwzględnego obliczonych kosinusów i sinusów, tj. części rzeczywistych i urojonych liczb w_j.

Szybkie mnożenie wielomianów

Zajmiemy się następującym zadaniem: dane są współczynniki $a_0, \ldots, a_n$ i $b_0, \ldots, b_m$ wielomianów $a(x) = \sum_{k=0}^{n} a_k x^k$ i $b(x) = \sum_{k=0}^{m} b_k x^k$. Należy obliczyć współczynniki $c_0, \ldots, c_{n+m}$ wielomianu $c(x) = \sum_{k=0}^{n+m} c_k x^k = a(x)b(x)$. „Zwykły” algorytm mnożenia wielomianów można zrealizować za pomocą podprogramu

```
for ( k = 0; k <= n+m; k++ ) c[k] = 0;
for ( i = 0; i <= n; i++ )
  for ( j = 0; j <= m; j++ ) c[i+j] += a[i]*b[j];
```

Operacją dominującą w tym algorytmie jest mnożenie współczynników; operacji tych należy wykonać $(n+1)(m+1)$; jeśli $m \approx n$, to złożoność obliczeniowa ma rząd $\Theta(n^2)$, choć zarówno danych, jak i wyników jest $\Theta(n)$.

Alternatywny sposób rozwiązywania tego zadania polega na wybraniu liczb $x_0, \ldots, x_{n+m}$, obliczeniu wartości wielomianów a i b, obliczeniu wartości $c(x_j) = a(x_j)b(x_j)$ wielomianu c i znalezieniu jego współczynników w bazie potęgowej, przez rozwiązanie zadania interpolacyjnego Lagrange'a. Mnożenie wielomianów — w postaci mnożenia ich wartości w wybranych punktach — wymaga wykonania tylko $n + m + 1$ mnożeń. Trzeba tylko umieć szybko obliczyć wartości wielomianów a i b i szybko rozwiązać zadanie interpolacyjne.

Do tego celu możemy użyć algorytmu FFT; jeśli przyjmiemy, że $x_j = e^{-2\pi i j/N}$, gdzie liczba N jest najmniejszą całkowitą potęgą liczby 2 większą niż $n + m$, to ciąg wartości wielomianu a w tych punktach jest dyskretną transformatą Fouriera ciągu współczynników $a_0, \ldots, a_n, 0, \ldots, 0$ o długości (a raczej okresie) N. Mając wartości wielomianu c w punktach x_j, możemy obliczyć jego współczynniki w bazie potęgowej, wyznaczając odwrotną dyskretną transformatę Fouriera. Całe to obliczenie jest wykonalne za pomocą $\Theta(N \log N) = \Theta((n+m)\log(n+m))$ działań zmiennopozycyjnych.

Zadania i problemy

1. Znajdź złożoność obliczeniową algorytmu FFT dla ciągu o długości $n = p^k$, gdzie p jest liczbą pierwszą, a k jest liczbą naturalną.

   Czy istnieje monotoniczna funkcja $f(n)$, taka że złożoność algorytmu FFT jest równa $\Theta(f(n))$? Odpowiedź uzasadnij.

2. Dwuwymiarowa dyskretna transformata Fouriera układu liczb $a_{jk} \in \mathbb{C}$, gdzie $j, k \in \mathbb{Z}$, takiego że istnieją liczby naturalne n, m, takie że dla każdego $j, k \in \mathbb{Z}$ jest $a_{jk} = a_{j+n,k} = a_{j,k+m}$, jest układem liczb

$$b_{pq} = \sum_{j=0}^{n} \sum_{k=0}^{m} a_{jk} e^{-2\pi i jp/n - 2\pi i kq/m}.$$

   Znajdź wzór opisujący dwuwymiarową dyskretną transformatę odwrotną.

   Podaj algorytm obliczania dwuwymiarowej dyskretnej transformaty Fouriera za pomocą algorytmu FFT (dla transformaty jednowymiarowej) i podaj złożoność tego algorytmu w zależności od liczb n i m.

3. Za normę ciągu okresowego $(a_k)_{k\in\mathbb{Z}}$ o okresie n możemy uznać dowolną normę wektora w $\mathbb{C}^n$, którego współrzędnymi są liczby $a_0, \ldots, a_{n-1}$. Niech to będzie norma Höldera z $p = 2$. Znajdź wskaźnik uwarunkowania zadania obliczania dyskretnej transformaty Fouriera w tej normie.

4. Napisz podprogram realizujący algorytm FFT dla przypadku, gdy $n = 3^k$, $k \in \mathbb{N}$.

Aproksymacja funkcji

Niech f oznacza funkcję określoną w przedziale $[a, b]$. Definicja tej funkcji może nie być wygodnym algorytmem obliczania jej wartości (np. funkcja f może być granicą nieskończonego ciągu), ewentualnie możemy mieć tylko „czarną skrzynkę" w postaci podprogramu obliczającego wartość funkcji f, przy czym koszt „sięgnięcia do tej skrzynki" może być bardzo duży, jeśli na przykład obliczenie wartości funkcji f w punkcie x polega na przeprowadzeniu eksperymentu fizycznego z parametrem x i dokonaniu pomiaru.

Zadanie aproksymacji polega na znalezieniu w ustalonej przestrzeni liniowej V, której elementy są funkcjami określonymi na przedziale $[a, b]$, funkcji g przybliżającej funkcję f (która w ogólności *nie jest* elementem przestrzeni V). Zwykle przestrzeń V wybieramy tak, aby koszt obliczania wartości należących do niej funkcji był mały, bo w zamierzeniu będziemy wielokrotnie obliczać wartości funkcji g, której chcemy używać zamiast f w jakimś celu.

Aby funkcja g mogła skutecznie „udawać" funkcję f, *musi* być skonstruowana w oparciu o wiedzę na temat własności funkcji f i na temat *zamierzonej jakości aproksymacji*. Jeśli na przykład funkcja f ma ciągłą pochodną, to możemy chcieć, aby nie tylko wartości funkcji g były bliskie wartościom funkcji f, ale także aby pochodna funkcji g przybliżała pochodną funkcji f (samo przybliżanie wartości funkcji f tego *nie zapewnia*). Przyjmiemy, że funkcja f jest elementem pewnej przestrzeni liniowej U, na przykład przestrzeni funkcji klasy $C^k[a, b]$ dla pewnego k. Będziemy rozpatrywać algorytmy dobierania funkcji z przestrzeni $V \subset U$.

Błąd aproksymacji będziemy mierzyć za pomocą pewnej normy określonej w przestrzeni U. Zwykle normy określa się za pomocą całek i bardzo często bierze się normy Höldera; wtedy miarą błędu przybliżenia funkcji f przez g jest wyrażenie (dla ustalonego $p \geqslant 1$)

$$\|f - g\|_p = \left(\int_a^b |f(x) - g(x)|^p \, dx\right)^{1/p}.$$

Dwa szczególnie ważne przypadki to $p = 2$ (mówimy wtedy o aproksymacji średniokwadratowej) oraz przypadek graniczny dla $p \to \infty$, gdy błąd jest określony wzorem

$$\|f - g\|_\infty = \max_{x \in [a,b]} |f(x) - g(x)|.$$

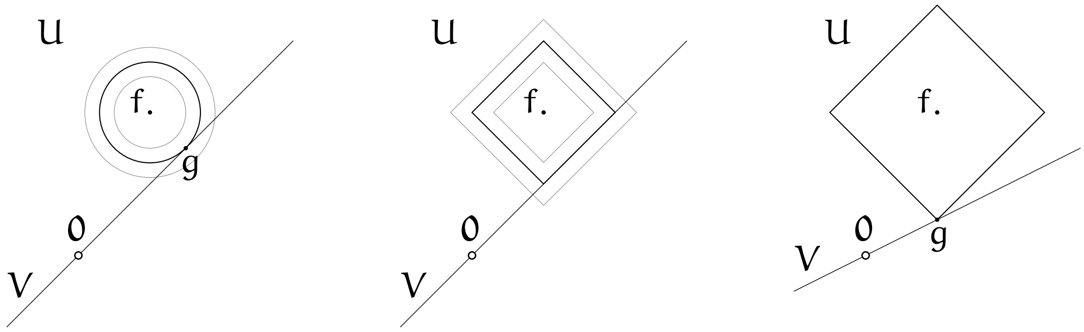
Ten przypadek nazywa się aproksymacją jednostajną.

Gdybyśmy byli zainteresowani także przybliżaniem pochodnej funkcji f, to mierzylibyśmy błąd innymi sposobami, np. obliczając wyrażenie

$$\max\{\|f-g\|_\infty, c\|f'-g'\|_\infty\},$$

z jakoś wybraną zawczasu stałą $c > 0$.

Zadania aproksymacji często stawia się dla funkcji, które nie są dokładnie znane; wiemy tylko, że funkcja f należy do *pewnej klasy* (o której coś wiemy, im więcej, tym lepiej) i mamy do dyspozycji *skończoną informację* na temat tej funkcji, np. jej wartości w pewnych punktach (mogą być otrzymane w wyniku pomiarów). Rozważamy też ciąg skończenie wymiarowych podprzestrzeni $V_1 \subset V_2 \subset \cdots$ przestrzeni U (np. przestrzeni wielomianów coraz wyższych stopni), w których poszukujemy przybliżeń. Minimalna ilość informacji, które należy podać (np. liczba punktów, w których trzeba zmierzyć wartości funkcji, przy odpowiednim rozmieszczeniu tych punktów), aby móc przybliżyć dowolną funkcję z rozpatrywanej klasy z zadaną dokładnością (przez wybranie elementu którejś przestrzeni V_k), jest nazywane złożonością informacyjną zadania aproksymacji.



Jeśli $f \in U$ (czyli funkcja f ma skończoną normę, uwaga: to jest nietrywialne, jeśli możemy tylko obliczać lub mierzyć wartości funkcji f w kolejnych punktach) i wymiar podprzestrzeni $V \subset U$ jest skończony, to rozwiązanie zadania aproksymacji istnieje: wybieramy najmniejszą kulę (tj. zbiór $B_{f,r} = \{\, h\colon h \in U, \|f-h\| \leqslant r\}$) o środku f, która ma niepuste przecięcie z podprzestrzenią V — kula taka istnieje, bo zbiór takich kul jest niepusty (jest w nim kula o promieniu $\|f\|$) i każda z nich jest domknięta, a podprzestrzeń V też jest domknięta; rozwiązaniem zadania jest dowolny element tego przecięcia. Rozwiązanie to może nie być jednoznaczne, jeśli brzeg dowolnej (a zatem każdej) kuli zawiera odcinek. Rozwiązanie *jest niejednoznaczne* wtedy, gdy taki odcinek ma kierunek dowolnego niezerowego wektora należącego do podprzestrzeni V, co jest pokazane na rysunkach.

Def. Przestrzeń unormowana U jest silnie wypukła, jeśli brzeg kuli nie zawiera żadnego odcinka.

Def. Norma w przestrzeni liniowej jest ostra jeśli nierówność trójkąta, $\|f+g\| \leqslant \|f\|+\|g\|$ jest równością wtedy i tylko wtedy, gdy f lub g jest wektorem zerowym, lub istnieje liczba dodatnia a, taka że $f = ag$.

Okazuje się, że przestrzeń unormowana U jest silnie wypukła wtedy, gdy norma jest ostra. Zadanie aproksymacji w każdej skończenie wymiarowej podprzestrzeni $V \subset U$ ma wtedy jednoznaczne rozwiązanie. Dla każdego $p > 1$ norma p-ta Höldera jest ostra, ale nie są ostre normy dla $p = 1$ i $p = \infty$.

Aproksymacja jednostajna

Z analizy znamy twierdzenie aproksymacyjne Weierstrassa: *Jeśli funkcja* f *jest ciągła na przedziale* $[a,b]$*, to dla każdego* $\varepsilon > 0$ *istnieje wielomian* p_n *pewnego stopnia* n*, taki że* $\|f-p_n\|_\infty \leqslant \varepsilon$.

Twierdzenie Weierstrassa ma konstruktywny dowód (Bernstein, 1912 r.), ale konstrukcja użyta w tym dowodzie nie nadaje się do praktycznego stosowania, bo nawet dla „łatwych" funkcji i niezbyt małego ε wynikające z dowodu oszacowanie liczby n może być rzędu wielu tysięcy, podczas gdy wystarczy n mniejsze niż 10. Jedną z przyczyn tak słabych wyników konstrukcji jest to, że poza ciągłością o funkcji f niczego się nie zakłada.

Jeśli funkcja f jest klasy $C^{n+1}[a,b]$, to zadanie aproksymacji możemy rozwiązać przez skonstruowanie wielomianu interpolacyjnego Lagrange'a lub Hermite'a. W tym celu wybieramy węzły interpolacyjne $x_i \in [a,b]$ dla $i = 0,\ldots,n$, obliczamy wartości funkcji f (i ewentualnie pochodnych, jeśli są krotne węzły) i stosujemy algorytm różnic dzielonych. Dla tak skonstruowanego wielomianu $h_n(x)$, na podstawie wzoru opisującego resztę, mamy

$$\|f-h_n\|_\infty = \max_{x\in[a,b]} \frac{|f^{(n+1)}(\xi(x))|}{(n+1)!}|p_{n+1}(x)|,$$

gdzie $p_{n+1}(x) = (x-x_0)\cdot\ldots\cdot(x-x_n)$. Mamy zatem problem, jak dobrać węzły, aby opisany powyższym wzorem błąd aproksymacji był jak najmniejszy.

Wielomiany i węzły Czebyszewa

Możemy ustalić $\varepsilon > 0$, a następnie starać się dobrać węzły interpolacyjne w przedziale $[a,b]$ w dowolny sposób zapewniający, że błąd aproksymacji jest mniejszy niż ε. Jeśli się to uda, to nie przejmujemy się tym, że inny wybór mógłby dać jeszcze mniejszy błąd. Jeśli $\max_{x\in[a,b]}|f^{(n+1)}(x)| \leqslant M_{n+1}$, to

$$\|f-h_n\|_\infty \leqslant \frac{M_{n+1}}{(n+1)!}\|p_{n+1}\|_\infty. \tag{*}$$

Możemy wybrać węzły tak, aby zminimalizować czynnik $\|p_{n+1}\|_\infty$. Aby to zrobić, zbadamy tzw. wielomiany Czebyszewa, zdefiniowane za pomocą wzorów

$$T_0(u) = 1,$$
$$T_1(u) = u,$$
$$T_k(u) = 2uT_{k-1}(u) - T_{k-2}(u) \quad \text{dla } k > 1.$$

Jest jasne, że funkcja $T_k(u)$ jest wielomianem stopnia k. Wzór rekurencyjny dla $k > 1$ to tak zwana formuła trójczłonowa, która umożliwia m.in. numeryczne obliczanie wartości tych wielomianów i ich kombinacji liniowych dla ustalonego u. Wielomiany Czebyszewa można określić także innymi sposobami, z których nam się przyda taki:

$$T_k(u) = \cos(k\arccos u) \quad \text{dla } u \in [-1,1].$$

Sprawdźmy, że to jest równoważna definicja: oznaczmy $u = \cos t$. Wtedy $T_0(u) = \cos 0 = 1$ oraz $T_1(u) = \cos t = u$, zaś dla $k > 1$, podstawiając $\alpha = kt$ i $\beta = (k-2)t$ do tożsamości trygonometrycznej

$$\cos\alpha + \cos\beta = 2\cos\frac{\alpha+\beta}{2}\cos\frac{\alpha-\beta}{2},$$

otrzymujemy równość

$$\cos kt = 2\cos(k-1)t\cos t - \cos(k-2)t,$$

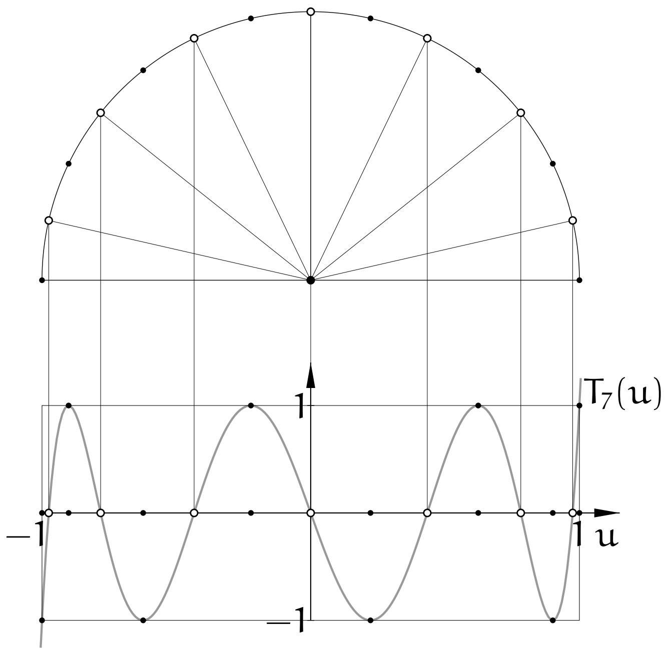
czyli formułę trójczłonową.

Na podstawie trygonometrycznego wzoru określającego wielomiany Czebyszewa możemy stwierdzić, że k miejsc zerowych wielomianu T_k (czyli wszystkie) znajduje się w przedziale $[-1,1]$, mianowicie są nimi liczby

$$z_j = \cos\frac{2j+1}{2k}\pi \quad \text{dla } j = 0,\dots,k-1,$$

a ponadto wielomian T_k w przedziale $[-1,1]$ przyjmuje wartości ekstremalne, na przemian $+1$ i -1, w punktach

$$y_j = \cos\frac{j}{k}\pi \quad \text{dla } j = 0,\dots,k.$$



Mając ustalony przedział $[a, b]$ oraz liczbę $k > 0$, możemy określić wielomian

$$q_k(x) = \frac{(b-a)^k}{2^{2k-1}} T_k(u),$$

gdzie $u = 2(x-a)/(b-a) - 1$, czyli $x = \frac{b+a}{2} + \frac{b-a}{2}u$. Z formuły trójczłonowej łatwo można wywnioskować, że wielomian $T_k(u)$ jest sumą wyrażenia $2^{k-1}u^k$ i pewnego wielomianu stopnia mniejszego niż k. Zatem współczynnik w bazie potęgowej przy x^k, czyli współczynnik wiodący wielomianu $q_k(x)$ jest równy 1. Wielomian q_k ma k miejsc zerowych w przedziale $[a, b]$ i w $k+1$ punktach tego przedziału, w tym w obu jego końcach, przyjmuje wartości ekstremalne, równe $\pm(b-a)^k/2^{2k-1}$.

Udowodnimy, że żaden wielomian stopnia k ze współczynnikiem wiodącym równym 1 nie może mieć mniejszych co do modułu wartości w całym przedziale $[a, b]$. Istotnie, gdyby taki wielomian, $w(x)$, istniał, to wielomian $r(x) = q_k(x) - w(x)$ miałby stopień mniejszy niż k, ale musiałby mieć co najmniej k miejsc zerowych w przedziale $[a, b]$, bo wykres wielomianu w przecinałby wykres wielomianu q_k co najmniej raz między każdymi jego sąsiednimi punktami ekstremalnymi (sąsiednie ekstrema mają tę samą wartość bezwzględną i przeciwne znaki, a wielomian w ma mieć w przedziale $[a, b]$ mniejsze wartości bezwzględne). Zatem, taki wielomian w nie istnieje. □

Dla dowolnych węzłów interpolacyjnych wielomian p_{n+1} występujący w oszacowaniu błędu ma współczynnik wiodący równy 1. Mamy zatem narzędzie do rozwiązywania zadania aproksymacji: aby przybliżyć funkcję klasy C^{n+1} w przedziale $[a, b]$, wybieramy tzw. węzły Czebyszewa, określone wzorem

$$x_j = \frac{b+a}{2} + \frac{b-a}{2} \cos \frac{2j+1}{2n+2}\pi \quad \text{dla } j = 0, \ldots, n,$$

i konstruujemy wielomian interpolacyjny Lagrange'a h_n stopnia n z tymi węzłami. Wtedy otrzymamy $p_{n+1} = q_{n+1}$ i

$$\|f - h_n\|_\infty \leqslant \frac{M_{n+1}}{(n+1)!} \frac{(b-a)^{n+1}}{2^{2n+1}}.$$

Wyrażenie po prawej stronie tej nierówności możemy porównać z przyjętym progiem ε, aby sprawdzić, czy błąd jest dostatecznie mały. Jeśli nie, ale funkcja f ma ciągłe pochodne wyższych rzędów (i umiemy znaleźć ich oszacowania), to możemy spróbować szczęścia z wielomianem interpolacyjnym wyższego stopnia.

Alternans i algorytm Remeza

Teraz zajmiemy się następującym problemem: dla ustalonej funkcji rzeczywistej f należy dobrać taki wielomian $g^\star$ stopnia co najwyżej n, aby błąd aproksymacji w normie maksimum w przedziale $[a, b]$ był najmniejszy. Nieco uogólniając zadanie, rozważymy problem aproksymacji przez określone w przedziale $[a, b]$ funkcje, które są elementami ustalonej przestrzeni V o wymiarze k; zatem, mając taką przestrzeń, chcemy w niej znaleźć element najlepiej przybliżający daną funkcję f, o której założymy, że jest ciągła.

Def. Przestrzeń liniowa V o wymiarze k, której elementami są rzeczywiste funkcje ciągłe określone w przedziale $[a, b]$, spełnia warunek Haara (albo: ma własność Haara), jeśli z faktu, że funkcja $g \in V$ ma k różnych miejsc zerowych w przedziale $[a, b]$ wynika, że jest to funkcja zerowa.

Własność Haara, dla dowolnie wybranego przedziału $[a, b]$, ma zatem przestrzeń liniowa $\mathbb{R}[x]_n$, której elementy są wielomianami stopnia co najwyżej n, ale nie tylko: weźmy przestrzeń wielomianów trygonometrycznych stopnia co najwyżej n i ustalmy dowolny przedział $[a, b]$ krótszy niż 2π (tj. krótszy niż okres wszystkich tych funkcji). Przestrzeń ta ma wymiar $2n + 1$, i jak wiemy, zadanie interpolacji Lagrange'a dla $2n + 1$ dowolnie wybranych w przedziale $[a, b]$ (parami różnych) węzłów ma w tej przestrzeni jednoznaczne rozwiązanie (zobacz rozdział 9 tych notatek). Jeśli więc pewien wielomian trygonometryczny stopnia n ma $2n + 1$ miejsc zerowych w przedziale $[a, b]$, to jest on funkcją zerową. Natomiast *nie mają* własności Haara przestrzenie, których elementami są funkcje sklejane: istnieją niezerowe funkcje sklejane, które mają nieskończenie wiele miejsc zerowych.

Twierdzenie Czebyszewa o alternansie: *Jeśli przestrzeń V o wymiarze k spełnia warunek Haara, to dla dowolnej funkcji ciągłej f zadanie aproksymacji jednostajnej ma w przestrzeni V jednoznaczne rozwiązanie, $g^\star$. Funkcja $f - g^\star$, opisująca błąd aproksymacji, ma w przedziale $[a, b]$ co najmniej $k + 1$ punktów, w których przyjmuje maksymalną wartość bezwzględną, przy czym znaki wartości funkcji $f - g^\star$ w kolejnych punktach z tego zbioru są przeciwne.*

Dowód twierdzenia Czebyszewa, który pominiemy, zawiera krok podobny do przeprowadzonego wcześniej dowodu stwierdzenia, że wielomian q_k ma najmniejszą normę $\|\cdot\|_\infty$ dla przedziału $[a, b]$ wśród wszystkich wielomianów stopnia k o współczynniku wiodącym 1 (i jest to jedyny taki wielomian).

Zbiór punktów, w których funkcja $f - g^\star$ przyjmuje na przemian minimalną i maksymalną wartość (wszystkie o tej samej wartości bezwzględnej $\|f - g^\star\|_\infty$) nazywany jest alternansem. Rozwiązanie zadania aproksymacji polega na znalezieniu takiego wielomianu $g^\star$ stopnia co najwyżej n, aby funkcja $f - g^\star$ przyjmowała w $n + 2$ punktach przedziału $[a, b]$ wartości ekstremalne o zmieniających się znakach. Jeśli funkcja f jest wypukła albo wklęsła i poszukujemy optymalnego wielomianu stopnia 1, to alternans składa się z trzech punktów, z których dwa są końcami przedziału $[a, b]$, dzięki czemu zadanie jest dosyć łatwe.

Jeśli poszukujemy optymalnego wielomianu wyższego stopnia, to możemy użyć opisanego niżej algorytmu Remeza, w którym konstruuje się pewien ciąg wielomianów $(g^{(j)})_{j\in\mathbb{N}}$ stopnia n, zbieżny do poszukiwanego wielomianu $g^\star$.

Za $g^{(0)}$ można przyjąć wielomian interpolacyjny Lagrange'a z $n+1$ węzłami Czebyszewa w przedziale $[a, b]$. Istotne jest, aby funkcja $f - g^{(0)}$ miała w przedziale $[a, b]$ co najmniej $n+2$ lokalne minima i maksima, rozmieszczone na przemian (wartości bezwzględne tych ekstremów mogą być różne), i na ogół taki wybór funkcji $g^{(0)}$ to zapewnia: funkcja $f - g^{(0)}$ ma minimum lub maksimum między każdymi dwoma węzłami interpolacyjnymi, a także przed pierwszym i za ostatnim węzłem. Jeśli znaki kolejnych ekstremów są takie same (o co jest bardzo trudno), to zamiast jednego z nich można przyjąć węzeł między nimi.

W j-tej iteracji na podstawie wielomianu $g^{(j-1)}$ należy skonstruować $g^{(j)}$. W tym celu trzeba znaleźć *wszystkie* ekstrema funkcji $f - g^{(j-1)}$ w przedziale $[a, b]$. To może być bardzo trudnym zadaniem obliczeniowym. Mając pewne informacje o funkcji f, możemy ustalić gęstość, z jaką wystarczy stablicować tę funkcję i wielomian $g^{(j-1)}$ w przedziale $[a, b]$, aby nie „zgubić” żadnego ekstremum (to może być np. 100, 1000, lub nawet więcej punktów), potem trzeba zastosować jakąś metodę numeryczną znajdowania punktów ekstremalnych z dużą dokładnością[27]. Następnie tworzymy j-te przybliżenie alternansu: wybieramy $n+2$ punkty w przedziale $[a, b]$, w których funkcja $f - g^{(j-1)}$ przyjmuje wartości ekstremalne, przy czym jeśli lokalnych ekstremów jest więcej niż $n+2$, to trzeba wybrać punkty, w których ekstrema mają największe wartości bezwzględne, z zachowaniem warunku zmieniających się znaków. Oznaczmy wybrane punkty symbolami $y_0^{(j)}, \dots, y_{n+1}^{(j)}$ (lub lepiej w skrócie $y_0, \dots, y_{n+1}$). Założymy, że są one uporządkowane monotonicznie.

Wielomian $g^{(j)}$ ma spełniać następujący warunek: dla $i = 0, \dots, n+1$ ma być $f(y_i) - g^{(j)}(y_i) = (-1)^i r_j$, gdzie r_j jest niewiadomą liczbą. Zatem, zachodzi równość $f(x) - g^{(j)}(x) = r_j h^{(j)}(x)$ dla pewnej funkcji $h^{(j)}$, takiej że $h^{(j)}(y_i) = (-1)^i$ dla $i = 0, \dots, n+1$. Obliczając różnicę dzieloną rzędu $n+1$, otrzymamy

$$f[y_0, \dots, y_{n+1}] = r_j h^{(j)}[y_0, \dots, y_{n+1}],$$

bo różnica dzielona rzędu $n+1$ wielomianu $g^{(j)}$ stopnia n jest zerem. Ale stąd

[27] Metody znajdowania ekstremów, nieobecne w tym wykładzie, mają związek z metodami rozwiązywania równań nieliniowych.

możemy obliczyć

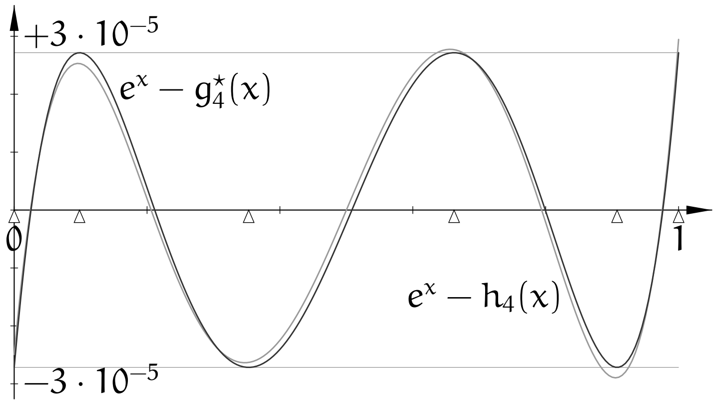
$$r_j = \frac{f[y_0, \dots, y_{n+1}]}{h^{(j)}[y_0, \dots, y_{n+1}]},$$

a następnie użyć r_j do obliczenia wartości wielomianu $g^{(j)}$ w punktach y_i i znaleźć ten wielomian przez rozwiązanie zadania interpolacyjnego Lagrange'a (mamy tu o 1 węzeł i warunek interpolacyjny za dużo, ale to nie szkodzi).

Ciąg wielomianów $g^{(j)}$ zwykle dość szybko zbiega do wielomianu $g^\star$, który przybliża funkcję f z najmniejszym błędem w przestrzeni $\mathbb{R}[x]_n$, przy czym ciąg liczb $|r_j|$ zbiega do normy błędu, tj. maksymalnej wartości bezwzględnej różnicy $f(x) - g^\star(x)$ w przedziale $[a, b]$. Jak widać z opisu (który jest dosyć uproszczony), to jest kosztowny algorytm, którego stosowanie może być opłacalne tylko wtedy, gdy wartości wielomianu $g^\star$ mają być obliczane *bardzo wiele razy*.

Przykład 1. Przybliżamy funkcję $f(x) = e^x$ w przedziale $[0, 1]$. Symbolem h_n oznaczymy wielomian interpolacyjny stopnia n oparty na węzłach Czebyszewa, a symbolem $g_n^\star$ wielomian optymalny znaleziony przy użyciu algorytmu Remeza. W ostatniej kolumnie tabeli podana jest liczba wykonanych iteracji (w każdej iteracji zostało znalezione nowe przybliżenie alternansu); punktem początkowym w każdym przypadku był wielomian h_n. Na rysunku są przedstawione wykresy błędów dla przypadku $n = 4$. Na osi poziomej są zaznaczone punkty alternansu.

n	$\lVert f - h_n \rVert_\infty$	$\lVert f - g_n^\star \rVert_\infty$	k
1	$1.24 \cdot 10^{-1}$	$1.06 \cdot 10^{-1}$	2
2	$9.87 \cdot 10^{-3}$	$8.76 \cdot 10^{-3}$	2
3	$6.00 \cdot 10^{-4}$	$5.45 \cdot 10^{-4}$	2
4	$2.95 \cdot 10^{-5}$	$2.72 \cdot 10^{-5}$	2
5	$1.21 \cdot 10^{-6}$	$1.13 \cdot 10^{-6}$	2
6	$4.28 \cdot 10^{-8}$	$4.03 \cdot 10^{-8}$	2

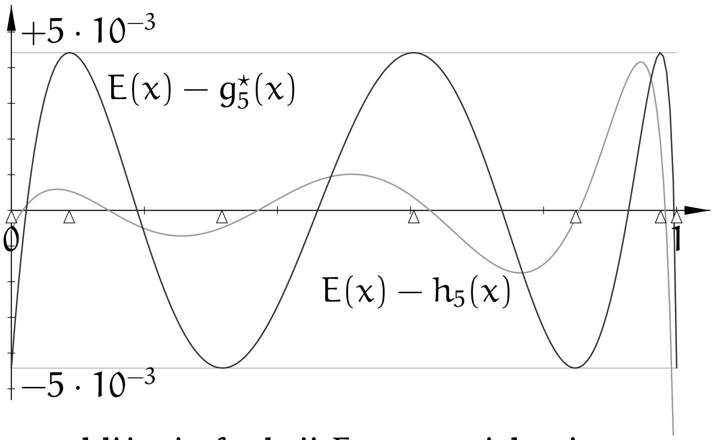


Przykład 2. W przedziale $[0, 1]$ przybliżamy funkcję

$$E(x) \stackrel{\text{def}}{=} \int_0^1 \sqrt{\frac{1 - x^2 t^2}{1 - t^2}}\, dt.$$

Jest to tak zwana zupełna całka eliptyczna drugiego rodzaju; występuje ona w różnych zastosowaniach (m.in. w mechanice). Funkcja ta maleje monotonicznie w przedziale $[0, 1]$, przyjmując na jego końcach wartości $E(0) = \frac{\pi}{2}$, $E(1) = 1$. Podobnie jak dla funkcji wykładniczej, *nie istnieje* wzór umożliwiający obliczanie $E(x)$ dla danego $x \in (0, 1)$ przy użyciu *skończenie wielu* działań arytmetycznych, co więcej, całek eliptycznych nie można wyrazić za pomocą funkcji wykładniczych i trygonometrycznych i ich odwrotności. Mając dane x, można konstruować rozmaite ciągi nieskończone, których granicą jest $E(x)$. W eksperymencie, którego wyniki są tu przedstawione, został użyty podprogram obliczający pewien wyraz takiego ciągu, przybliżający wartość funkcji E z błędem mniejszym niż 10^{-6}.

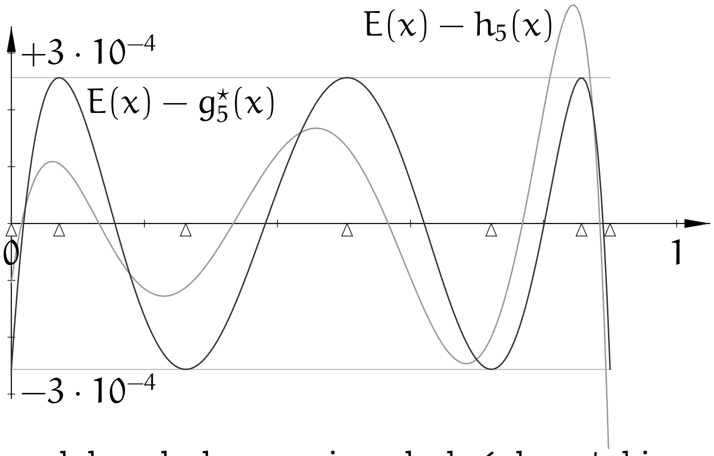
n	$\|E - h_n\|_\infty$	$\|E - g_n^\star\|_\infty$	k
1	$1.54 \cdot 10^{-1}$	$9.49 \cdot 10^{-2}$	2
2	$5.43 \cdot 10^{-2}$	$2.41 \cdot 10^{-2}$	3
3	$3.00 \cdot 10^{-2}$	$1.18 \cdot 10^{-2}$	3
4	$1.88 \cdot 10^{-2}$	$6.81 \cdot 10^{-3}$	3
5	$1.30 \cdot 10^{-2}$	$4.42 \cdot 10^{-3}$	3
6	$9.48 \cdot 10^{-3}$	$3.09 \cdot 10^{-3}$	4



Na rysunku wyżej są wykresy błędów przybliżenia funkcji E przez wielomiany piątego stopnia: h_5, interpolującego funkcję E w węzłach Czebyszewa, i $g_5^\star$. Możemy zauważyć, że funkcja E jest znacznie trudniejsza do aproksymacji — błędy przybliżeń znacznie wolniej maleją ze wzrostem stopnia niż w przypadku funkcji e^x. Powód jest taki, że funkcja E ma w przedziale $[0, 1)$ nieograniczoną pochodną; jest $\lim_{x\to 1} E'(x) = -\infty$. Wprawdzie maksymalny błąd przybliżenia funkcji E w przedziale $[0, 1]$ przez wielomian $g_5^\star$ jest mniejszy niż błąd przybliżenia przez h_5, ale możemy zauważyć, że gdybyśmy się ograniczyli do przedziału np. $[0, 0.9]$, to wielomian h_5 byłby znacznie lepszym przybliżeniem funkcji $E(x)$.

Wielomiany interpolacyjne funkcji E z węzłami Czebyszewa i wielomiany optymalne w przedziale $[0, 0.9]$ przybliżają funkcję E z błędami pokazanymi w następnej tabelce i (dla $n = 5$) na rysunku.

n	$\|E - h_n\|_\infty$	$\|E - g_n^\star\|_\infty$	k
1	$7.82 \cdot 10^{-2}$	$5.92 \cdot 10^{-2}$	2
2	$1.33 \cdot 10^{-2}$	$7.85 \cdot 10^{-3}$	3
3	$4.10 \cdot 10^{-3}$	$2.34 \cdot 10^{-3}$	3
4	$1.34 \cdot 10^{-3}$	$7.23 \cdot 10^{-4}$	3
5	$4.90 \cdot 10^{-4}$	$2.57 \cdot 10^{-4}$	3
6	$1.88 \cdot 10^{-4}$	$9.60 \cdot 10^{-5}$	3



Praktyczne wnioski z eksperymentów podobnych do powyższych dwóch są takie: jeśli funkcja, której przybliżenie wielomianowe należy skonstruować, ma pochodne wspólnie ograniczone, to zwykle nie sprawia kłopotów, ale zastąpienie wielomianu interpolacyjnego z węzłami Czebyszewa przez wynik działania algorytmu Remeza niewiele poprawia aproksymację; skuteczniejszym sposobem zmniejszenia błędu jest zwykle znalezienie wielomianu interpolacyjnego wyższego stopnia (z węzłami Czebyszewa). Jeśli funkcja ma w rozpatrywanym przedziale osobliwość, to żaden z tych sposobów nie jest dobry. Dobrym sposobem na pokonanie trudności jest zwykle podzielenie przedziału na krótsze podprzedziały i poszukiwanie wielomianów aproksymacyjnych w tych podprzedziałach. Wynikiem takiego postępowania jest aproksymacyjna funkcja sklejana. Jeśli jednak w pewnym podprzedziale jest osobliwość (np. nieciągła pochodna), to warto się zastanowić nad innym sposobem przybliżania funkcji, niż za pomocą wielomianów. Jedna z możliwości to użycie funkcji wymiernych. Aby skutecznie aproksymować, *zawsze* należy wiedzieć, jakie (jakiego rodzaju) osobliwości ma dana funkcja.

Aproksymacja jednostajna przez funkcje sklejane

Zauważmy, że gdyby funkcja f miała być przybliżana w przedziale dwukrotnie krótszym (np. w połowie przedziału $[a, b]$), w którym przyjęlibyśmy węzły interpolacyjne rozmieszczone w dwukrotnie mniejszych odstępach, to czynnik $\|p_{n+1}\|_\infty$ we wzorze (*) dla tego krótszego przedziału byłby 2^{n+1} razy mniejszy. Zatem skrócenie przedziału $[a, b]$ jest radykalnym sposobem zmniejszenia błędu aproksymacji jednostajnej i czasami jest to jedyny skuteczny sposób. Mając długi przedział, możemy podzielić go na krótsze podprzedziały i aproksymować funkcję f w każdym z nich innym wielomianem niskiego stopnia.

Zwykle potrzebna jest aproksymacyjna funkcja ciągła razem z pochodnymi pewnego rzędu — konstruujemy ją w całości, a nie osobno poszczególne wielomiany w podprzedziałach. Stopień i węzły dobieramy odpowiednio do zastosowania. Znanych jest wiele twierdzeń na temat aproksymacji funkcjami sklejanymi różnych stopni. Na przykład

<u>Twierdzenie</u>. *Niech* $f \in C^2[a, b]$ *i niech* s *oznacza kubiczną funkcję sklejaną klasy* $C^2[a, b]$ *z węzłami* $u_0 = a < u_1 < \cdots < u_N = b$*, taką że* $s(u_i) = f(u_i)$ *dla* $i = 0, \ldots, N$*. Niech* M_2 *oznacza stałą, taką że* $|f''(x)| \leqslant M_2$ *dla każdego* $x \in [a, b]$*, oraz* $|s''(a)| \leqslant 3M_2$ *i* $|s''(b)| \leqslant 3M_2$*. Wtedy dla każdego* $x \in [a, b]$ *zachodzą nierówności*

$$|f(x) - s(x)| \leqslant \frac{1}{2}M_2h^2, \qquad |f'(x) - s'(x)| \leqslant 2M_2h,$$

gdzie $h = \max_i(u_{i+1} - u_i)$*.*

Dowód. Niech $a_i = f(u_i) = s(u_i)$ dla $i = 0, \ldots, N$. Symbolami b_i i c_i oznaczymy odpowiednio wartości pochodnych rzędu 1 i 2 funkcji sklejanej s w węźle u_i. Dla $i = 0, \ldots, N-1$ niech p_i oznacza wielomian opisujący funkcję s w przedziale $[u_i, u_{i+1}]$, którego długość oznaczymy symbolem h_i.

Rozważamy wielomiany p_{i-1} i p_i. Pochodna drugiego rzędu każdego z nich jest wielomianem stopnia co najwyżej 1. Możemy napisać

$$p''_{i-1}(x) = \frac{c_i - c_{i-1}}{h_{i-1}}t + c_i, \qquad\qquad p''_i(x) = \frac{c_{i+1} - c_i}{h_i}t + c_i,$$

gdzie $t = x - u_i$ jest nową zmienną, wprowadzoną dla wygody. Całkując dwa razy, z odpowiednio dobranymi stałymi całkowania, otrzymamy te wielomiany:

$$\begin{aligned} p'_{i-1}(x) &= \frac{c_i - c_{i-1}}{2h_{i-1}}t^2 + c_i t + b_i, & p'_i(x) &= \frac{c_{i+1} - c_i}{2h_i}t^2 + c_i t + b_i, \\ p_{i-1}(x) &= \frac{c_i - c_{i-1}}{6h_{i-1}}t^3 + \frac{c_i}{2}t^2 + b_i t + a_i, & p_i(x) &= \frac{c_{i+1} - c_i}{6h_i}t^3 + \frac{c_i}{2}t^2 + b_i t + a_i. \end{aligned}$$

Podstawiając odpowiednio $t = -h_{i-1}$ i $t = h_i$, możemy obliczyć wartości tych wielomianów w punktach u_{i-1} oraz u_{i+1}:

$$\begin{aligned} a_{i-1} &= -\frac{c_i - c_{i-1}}{6}h_{i-1}^2 + \frac{c_i}{2}h_{i-1}^2 - b_i h_{i-1} + a_i, \\ a_{i+1} &= \frac{c_{i+1} - c_i}{6}h_i^2 + \frac{c_i}{2}h_i^2 + b_i h + a_i. \end{aligned}$$

Na podstawie każdej z tych równości można obliczyć b_i; oba otrzymane w ten sposób wyrażenia są równe:

$$\frac{a_i - a_{i-1}}{h_{i-1}} + \frac{c_i h_{i-1}}{3} + \frac{c_{i-1}h_{i-1}}{6} = b_i = \frac{a_{i+1} - a_i}{h_i} - \frac{c_i h_i}{3} - \frac{c_{i+1}h_i}{6}.$$

Zauważamy w tych wzorach różnice dzielone, $\frac{a_i - a_{i-1}}{h_{i-1}} = f[u_{i-1}, u_i]$ oraz $\frac{a_{i+1} - a_i}{h_i} = f[u_i, u_{i+1}]$. Przenosimy je na prawą stronę, pozostałe składniki na lewą i mnożymy strony przez 6:

$$2h_{i-1}c_i + h_{i-1}c_{i-1} + 2h_i c_i + h_i c_{i+1} = 6f[u_i, u_{i+1}] - 6f[u_{i-1}, u_i].$$

Dzielimy strony przez $h_{i-1} + h_i = u_{i+1} - u_{i-1}$, lewą stronę porządkujemy, a po prawej zauważamy różnicę dzieloną drugiego rzędu:

$$\frac{h_{i-1}}{h_{i-1} + h_i}c_{i-1} + 2c_i + \frac{h_i}{h_{i-1} + h_i}c_{i+1} = 6f[u_{i-1}, u_i, u_{i+1}]. \qquad (\diamond)$$

Otrzymaliśmy równanie liniowe z niewiadomymi c_{i-1}, c_i, c_{i+1}. Układ tych równań dla $i = 1, \ldots, N-1$ daje alternatywny sposób skonstruowania kubicznej sklejanej funkcji interpolacyjnej. Podobnie, jak poprzednio, są tu o dwa równania za mało, aby rozwiązanie było jednoznaczne, trzeba dołączyć np. warunki brzegowe. Na przykład, biorąc $c_0 = c_N = 0$, dostaniemy naturalną funkcję sklejaną.

Niech j oznacza taką liczbę, że $|c_j| = \max_i |c_i|$. Jeśli $j = 0$ lub $j = N$, to bezpośrednio z założeń twierdzenia mamy $|c_i| \leqslant 3M_2$ dla $i = 0, \ldots, N$.
W przeciwnym razie mamy oczywistą nierówność

$$|c_j| \leqslant 2|c_j| - \left|\frac{h_{j-1}}{h_{j-1}+h_j}c_{j-1}\right| - \left|\frac{h_j}{h_{j-1}+h_j}c_{j+1}\right|.$$

Z drugiej strony, z równania (◇)

$$2c_j = 6f[u_{j-1}, u_j, u_{j+1}] - \frac{h_{j-1}}{h_{j-1}+h_j}c_{j-1} - \frac{h_j}{h_{j-1}+h_j}c_{j+1}$$

wynika nierówność

$$2|c_j| \leqslant 6|f[u_{j-1}, u_j, u_{j+1}]| + \left|\frac{h_{j-1}}{h_{j-1}+h_j}c_{j-1}\right| + \left|\frac{h_j}{h_{j-1}+h_j}c_{j+1}\right|,$$

czyli

$$2|c_j| - \left|\frac{h_{j-1}}{h_{j-1}+h_j}c_{j-1}\right| - \left|\frac{h_j}{h_{j-1}+h_j}c_{j+1}\right| \leqslant 6|f[u_{j-1}, u_j, u_{j+1}]|,$$

skąd wynika, że

$$|c_j| \leqslant 6f[u_{j-1}, u_j, u_{j+1}].$$

Dla funkcji f klasy C^2 istnieje $\xi \in [u_{j-1}, u_{j+1}] \subset [a, b]$, takie że $f[u_{j-1}, u_j, u_{j+1}] = f''(\xi)/2$, zatem $|c_j| \leqslant 3M_2$, a stąd $|c_i| \leqslant 3M_2$ dla $i = 0, \ldots, N$.

Funkcja błędu $e(x) = f(x) - s(x)$ jest klasy C^2 i ma wartość 0 w każdym węźle u_i. Ponadto, ponieważ funkcja $|s''|$ wartość maksymalną przyjmuje w którymś węźle, gdzie nie przekracza $3M_2$, możemy oszacować $|e''(x)| \leqslant 4M_2$ dla każdego $x \in [a, b]$. Niech $x \in (u_i, u_{i+1})$. Na podstawie wzoru (**) (str. 7.5) możemy napisać

$$e[u_i, x, u_{i+1}] = \frac{e(x)}{(x-u_i)(x-u_{i+1})}, \quad \text{czyli}$$
$$e(x) = e[u_i, x, u_{i+1}](x-u_i)(x-u_{i+1}) = \frac{e''(\xi_i)}{2}(x-u_i)(x-u_{i+1}).$$

dla pewnego $\xi_i \in [u_i, u_{i+1}]$. Mamy też $|(x-u_i)(x-u_{i+1})| \leqslant \frac{1}{4}h_i^2$, skąd wynika, że

$$|e(x)| \leqslant \frac{1}{2}M_2h_i^2.$$

Należy jeszcze oszacować błąd aproksymacji pochodnej funkcji f przez s', czyli pochodną funkcji e. Jeśli $u_0 \leqslant x_1 < x_2 \leqslant u_N$, to

$$|e'(x_2) - e'(x_1)| = \left|\int_{x_1}^{x_2} e''(x)\,dx\right| \leqslant 4M_2(x_2 - x_1). \tag{$\circ$}$$

Przypuśćmy, że $e'(x) > 2M_2h_i$ dla pewnego $x \in [u_i, u_{i+1}]$. Z tego przypuszczenia i z nierówności ($\circ$) wynika, że jeśli $y \leqslant x$, to $e'(y) > 2M_2h_i - 4M_2(x-y)$, a jeśli $y \geqslant x$, to $e'(y) > 2M_2h_i - 4M_2(y-x)$. Niech $g_0 = x - u_i$, $g_1 = u_{i+1} - x$. Liczby g_0 i g_1 są nieujemne i $g_0 + g_1 = h_i$. Stąd wyciągamy wniosek

$$\begin{aligned}
&0 = e(u_{i+1}) - e(u_i) = \int_{u_i}^{u_{i+1}} e'(y)\,dy > \\
&\quad \int_{u_i}^{x} \big(2M_2h_i - 4M_2(x-y)\big)\,dy + \int_{x}^{u_{i+1}} \big(2M_2h_i - 4M_2(y-x)\big)\,dy = \\
&\quad 2M_2\big((g_0+g_1)^2 - g_0^2 - g_1^2\big) = 4M_2g_0g_1 \geqslant 0,
\end{aligned}$$

czyli $0 > 0$, oczywiście fałszywy. Taki sam wniosek, a dokładniej nierówność $0 < 0$, wynika z przypuszczenia, że istnieje takie $x \in [u_i, u_{i+1}]$, że $e'(x) < -2M_2h_i$, i wyciągnięciem wniosku z tych wniosków kończymy dowód. □

Udowodnione wyżej twierdzenie można stosować *nie tylko* do naturalnych kubicznych funkcji sklejanych. Wynika z niego, że do osiągnięcia dowolnie małego błędu wystarczy wybranie dostatecznie gęstego ciągu węzłów w przedziale $[a, b]$, a ponadto można w ten sposób również dowolnie zmniejszyć błąd aproksymacji pochodnej funkcji f.

Aproksymacja średniokwadratowa

Niech ρ oznacza funkcję określoną w przedziale A (który może być otwarty lub domknięty, ograniczony lub nieograniczony). Zakładamy, że funkcja ρ jest nieujemna, zbiór jej miejsc zerowych w A jest miary zero (np. pusty) i dla każdego wielomianu w całka z iloczynu $w\rho$ w zbiorze A jest skończona. Funkcję ρ nazywamy funkcją wagową albo wagą. Dla takiej funkcji wzór

$$\langle f, g\rangle_\rho = \int_A f(x)g(x)\rho(x)\,dx$$

określa iloczyn skalarny, a funkcjonał

$$\|f\|_\rho = \sqrt{\langle f, f\rangle_\rho} = \sqrt{\int_A f(x)^2\rho(x)\,dx}$$

jest normą. Zadanie aproksymacji średniokwadratowej często jest uogólniane w ten sposób, że dla danej funkcji f należy znaleźć w ustalonej przestrzeni V

(której wymiar jest skończony) funkcję g, taką że wyrażenie $\|f-g\|_\rho$ jest najmniejsze.

Rozwiązaniem zadania jest wektor (funkcja) $g^\star$, która jest rzutem prostopadłym wektora (funkcji) f na przestrzeń V; zadanie aproksymacji średniokwadratowej jest w istocie uogólnieniem liniowego zadania najmniejszych kwadratów. Mając bazę $\{p_0,\ldots,p_n\}$ przestrzeni V, wystarczy znaleźć współczynniki $x_0,\ldots,x_n$ wektora $g^\star=\sum_{j=0}^n x_j p_j$ w tej bazie. Wektor $f-g^\star$ jest prostopadły do wszystkich elementów bazy przestrzeni V. Na podstawie tego warunku możemy wyprowadzić układ równań normalnych

$$\sum_{j=0}^n \langle p_i,p_j\rangle_\rho x_j=\langle p_i,f\rangle_\rho, \quad \text{dla } i=0,\ldots,n.$$

Macierz $A=[\langle p_i,p_j\rangle_\rho]_{i,j}$ tego układu równań jest symetryczna i dodatnio określona. Zatem rozwiązanie zadania przez rozwiązanie układu równań liniowych z macierzą A wydaje się proste.

Przykład. Jeśli aproksymujemy funkcję wielomianem stopnia co najwyżej n i przyjmujemy iloczyn skalarny

$$\langle f,g\rangle=\int_0^1 f(x)g(x)\,dx,$$

to w razie użycia bazy potęgowej otrzymamy układ równań liniowych z macierzą Hilberta $(n+1)\times(n+1)$ (o współczynnikach $a_{ij}=1/(i+j+1)$); już dla niewielkich n ta macierz ma ogromny wskaźnik uwarunkowania (np. dla $n=10$ $\operatorname{cond}_2 A\approx 5\cdot 10^{14}$).

Wielomiany ortogonalne

Zadanie aproksymacji średniokwadratowej jest znacznie łatwiejsze do rozwiązania, jeśli dysponujemy bazą ortogonalną przestrzeni V, tj. układem niezerowych wektorów (funkcji) $p_0,\ldots,p_n$, takich że $\mathrm{lin}\{p_0,\ldots,p_n\}=V$ oraz $\langle p_i,p_j\rangle_\rho=0$ dla $i\neq j$. Dla takiej bazy macierz układu równań normalnych jest diagonalna. Mając dowolną bazę przestrzeni V, możemy znaleźć bazę ortogonalną za pomocą ortogonalizacji Grama-Schmidta. Jeśli $V=\mathbb{R}[x]_n$, tj. elementami przestrzeni V są wszystkie wielomiany stopnia co najwyżej n, to za pomocą ortogonalizacji bazy potęgowej możemy znaleźć bazę ortogonalną $p_0,\ldots,p_n$, w której dla każdego k stopień wielomianu p_k jest równy k. Bazę taką możemy również znaleźć za pomocą odpowiedniej formuły trójczłonowej; wcześniej otrzymaliśmy tym sposobem wielomiany Czebyszewa.

<u>Twierdzenie.</u> *Dla ustalonego przedziału* $A \subset \mathbb{R}$ *i funkcji wagowej* ρ *wielomiany* p_k*, tworzące układ ortogonalny dla* $k = 0, 1, \ldots$ *i takie, że dla każdego* k *stopień wielomianu* p_k *jest równy* k*, wyrażają się wzorem*

$$p_k(x) = (\alpha_k x + \beta_k)p_{k-1}(x) + \gamma_k p_{k-2}(x) \quad \textit{dla } k = 1, 2, \ldots,$$

dla pewnych liczb $\alpha_k \neq 0$ *(konstruując bazę, można je wybrać dowolnie), oraz*

$$\beta_k = -\frac{\alpha_k \langle x p_{k-1}, p_{k-1}\rangle_\rho}{\|p_{k-1}\|_\rho^2}, \qquad \gamma_k = -\frac{\alpha_k \langle x p_{k-1}, p_{k-2}\rangle_\rho}{\|p_{k-2}\|_\rho^2},$$

przy czym $p_{-1}(x) \stackrel{\text{def}}{=} 0$*,* $p_0(x) \stackrel{\text{def}}{=} a_0 \neq 0$*.*

<u>Dowód</u>. Dla każdego k stopień wielomianu p_k jest równy k, zatem jego współczynnik wiodący $a_k \neq 0$. Niech $\alpha_k = a_k/a_{k-1}$. Niech $w_k = p_k - \alpha_k x p_{k-1}$. Wielomian w_k jest stopnia mniejszego niż k. Dla iloczynu skalarnego określonego za pomocą całki z wagą zachodzi równość $\langle xf, g\rangle_\rho = \langle f, xg\rangle_\rho$, zatem dla $j < k-2$

$$\langle w_k, p_j\rangle_\rho = \langle p_k - \alpha_k x p_{k-1}, p_j\rangle_\rho = \langle p_k, p_j\rangle_\rho - \alpha_k\langle p_{k-1}, x p_j\rangle_\rho = 0.$$

Wyrażając wielomian w_k w bazie $p_0, \ldots, p_{k-1}$, otrzymamy

$$\langle w_k, p_j\rangle_\rho = \langle \sum_{i=0}^{k-1} b_{ki} p_i, p_j\rangle_\rho = \sum_{i=0}^{k-1} b_{ki}\langle p_i, p_j\rangle_\rho = b_{kj}\langle p_j, p_j\rangle_\rho.$$

Stąd $b_{kj} = 0$ dla $j < k-2$, a zatem mamy

$$p_k = \alpha_k x p_{k-1} + \beta_k p_{k-1} + \gamma_k p_{k-2},$$

dla $\beta_k = b_{k,k-1}$, $\gamma_k = b_{k,k-2}$. Możemy obliczyć

$$0 = \langle p_k, p_{k-1}\rangle_\rho = \alpha_k\langle x p_{k-1}, p_{k-1}\rangle_\rho + \beta_k\langle p_{k-1}, p_{k-1}\rangle_\rho + \gamma_k \underbrace{\langle p_{k-2}, p_{k-1}\rangle_\rho}_{=0},$$

$$0 = \langle p_k, p_{k-2}\rangle_\rho = \alpha_k\langle x p_{k-1}, p_{k-2}\rangle_\rho + \beta_k \underbrace{\langle p_{k-1}, p_{k-2}\rangle_\rho}_{=0} + \gamma_k\langle p_{k-2}, p_{k-2}\rangle_\rho$$

skąd otrzymujemy podane wyrażenia na β_k i γ_k. □

Oczywiście, można wybrać liczby a_0 i α_k tak, aby dostać bazę ortonormalną, ale nie zawsze się tak robi. Ważną własnością wielomianów ortogonalnych (dowód jest prostym ćwiczeniem) jest to, że wszystkie ich miejsca zerowe są rzeczywiste, jednokrotne i położone wewnątrz przedziału A.

Wielomiany ortogonalne są znane dla wielu różnych przedziałów i wag; największe znaczenie praktyczne mają

- <u>wielomiany Legendre'a</u>: $A = (-1, 1)$, $\rho(x) = 1$,

$$P_0(x) = 1, \quad P_1(x) = x, \quad P_k(x) = \frac{2k-1}{k} x P_{k-1}(x) - \frac{k-1}{k} P_{k-2}(x),$$

- <u>wielomiany Czebyszewa</u>: $A = (-1, 1)$, $\rho(x) = (1 - x^2)^{-1/2}$,

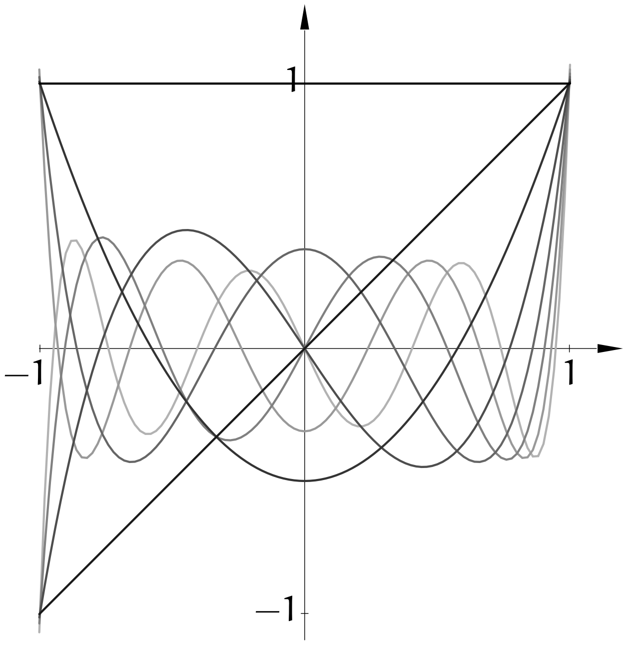
$$T_0(x) = 1, \quad T_1(x) = x, \quad T_k(x) = 2x T_{k-1}(x) - T_{k-2}(x),$$

- <u>wielomiany Hermite'a</u>[28]: $A = \mathbb{R}$, $\rho(x) = e^{-x^2}$,

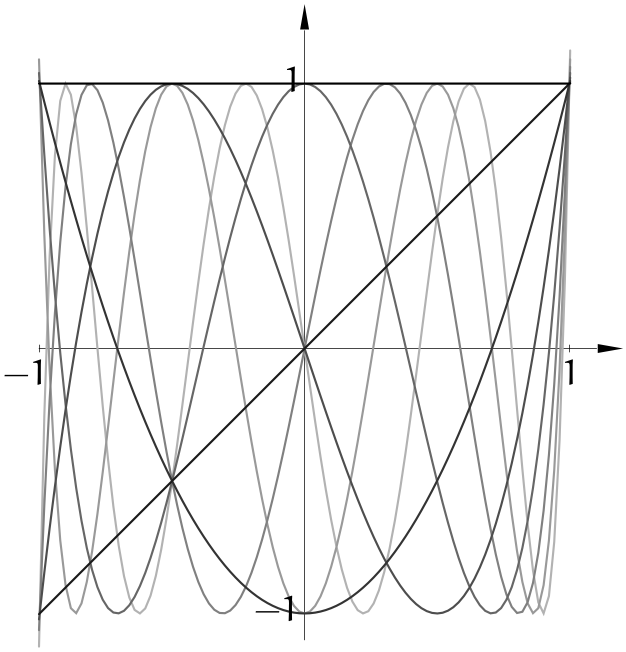
$$H_0(x) = 1, \quad H_1(x) = 2x, \quad H_k(x) = 2x H_{k-1}(x) - (2k-2) H_{k-2}(x),$$

- <u>wielomiany Laguerre'a</u>: $A = (0, +\infty)$, $\rho(x) = e^{-x}$,

$$L_0(x) = 1, \quad L_1(x) = 1 - x, \quad L_k(x) = \frac{2k-1-x}{k} L_{k-1}(x) - \frac{k-1}{k} L_{k-2}(x).$$



Wielomiany Legendre'a

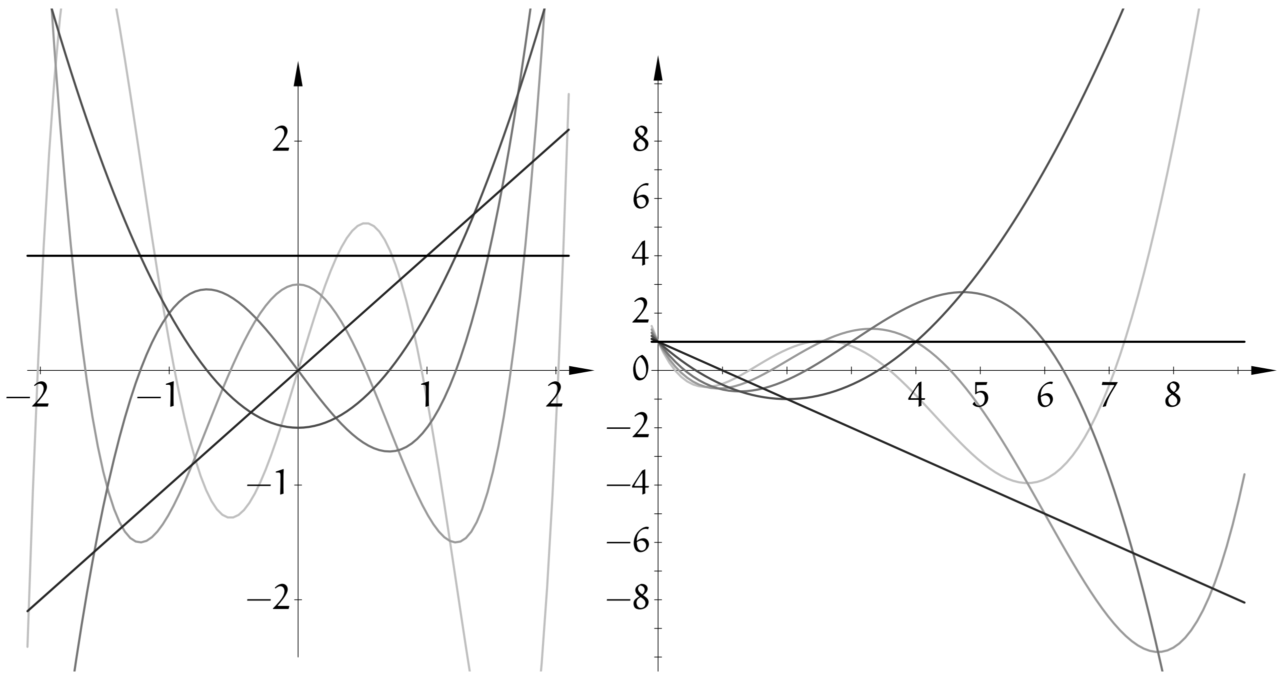


Wielomiany Czebyszewa

Tak więc jeśli mamy bazę przestrzeni $\mathbb{R}[x]_n$ ortogonalną w sensie iloczynu skalarnego $\langle \cdot, \cdot \rangle_\rho$, to zadanie znalezienia wielomianu $g_n^\star$ stopnia co najwyżej n, najlepiej przybliżającego funkcję f, sprowadza się do obliczenia współczynników wielomianu $g_n^\star$ w tej bazie:

$$x_i = \frac{\langle f, p_i \rangle_\rho}{\|p_i\|_\rho^2}.$$

[28]Nie należy mieszać wielomianów ortogonalnych Hermite'a z zadaniem interpolacyjnym Hermite'a.



Wielomiany Hermite'a
(tu są wykresy funkcji $2^{-k}H_k$)

Wielomiany Laguerre'a

Mamy przy tym, na podstawie twierdzenia Pitagorasa, wyrażenie opisujące błąd:

$$\|f - g_n^\star\|_\rho^2 = \|f\|_\rho^2 - \sum_{i=0}^{n} x_i^2 \|p_i\|_\rho^2,$$

dzięki czemu, jeśli błąd jest za duży, możemy zwiększać n, obliczając tylko kolejne współczynniki x_i (ale <u>uwaga:</u> są takie funkcje f, dla których błąd nie maleje do zera, gdy $n \to \infty$ — trzeba uważać). Podstawą rozwiązywania zadań aproksymacji średniokwadratowej jest obliczanie całek, co można robić analitycznie (jeśli umiemy, a nie zawsze tak jest) lub numerycznie. Numerycznym całkowaniem zajmiemy się wkrótce.

Zadania i problemy

1. Algorytm Clenshawa. Niech $h(x) = \sum_{i=0}^{n} a_i T_i(x)$. Dla $n > 1$ możemy obliczyć

$$\begin{aligned} h(x) = {} & a_n T_n(x) + a_{n-1} T_{n-1}(x) + a_{n-2} T_{n-1}(x) + \cdots = \\ & a_n(2x T_{n-1}(x) - T_{n-2}(x)) + a_{n-1} T_{n-1}(x) + a_{n-2} T_{n-1}(x) + \cdots = \\ & (2x a_n + a_{n-1}) T_{n-1}(x) + (a_{n-2} - a_n) T_{n-2}(x) + \cdots = \\ & \tilde{a}_{n-1} T_{n-1}(x) + \tilde{a}_{n-2} T_{n-2}(x) + \cdots . \end{aligned}$$

   Wykonując ten rachunek rekurencyjnie, możemy otrzymać liczby $\hat{a}_1$ i $\hat{a}_0$, takie że $h(x) = \hat{a}_1 T_1(x) + \hat{a}_0 T_0(x) = \hat{a}_1 x + \hat{a}_0$. Napisz procedurę obliczającą $h(x)$, opartą na tym pomyśle.

2. Znajdź najmniejsze n, takie że wielomian interpolacyjny h_n stopnia n z węzłami Czebyszewa w przedziale $[0, \pi/2]$ przybliża jednostajnie funkcję $f(x) = \sin x$ w tym przedziale z błędem nie większym niż 10^{-4}.

   Co można powiedzieć o błędzie względnym aproksymacji w tym przypadku?

   Wskazówka: w oszacowaniach wygodnie jest korzystać z nierówności $\pi^2 < 10$.

3. Znajdź wielomian stopnia co najwyżej 1, który przybliża jednostajnie funkcję $f(x) = e^x$ w przedziale $[0, 1]$ z najmniejszym możliwym błędem.

4. Znajdź wielomian $q^*(x)$ stopnia co najwyżej 7, który najlepiej przybliża w normie maksimum na przedziale $[-1, 1]$ wielomian $f(x) = x^8$.

5. Znajdź wielomiany stopnia 0, 1 i 2, będące rozwiązaniami aproksymacji średniokwadratowej dla funkcji $f(x) = e^x$ w przedziale $[0, 1]$ z wagą $\rho(x) = 1$. W rachunkach użyj bazy potęgowej.

6. Udowodnij, że wielomiany Czebyszewa stanowią układ ortogonalny dla iloczynu skalarnego określonego za pomocą całki w przedziale $(-1, 1)$ z wagą $(1 - x^2)^{-1/2}$.

7. Znajdź wielomiany ortogonalne Legendre'a stopnia 0, 1, 2 i 3, za pomocą ortogonalizacji bazy potęgowej.

8. Udowodnij, że jeśli p_k jest wielomianem k-tego stopnia, należącym do rodziny ortogonalnej w sensie iloczynu skalarnego określonego za pomocą całki w spójnym przedziale A z funkcją wagową ρ, to wszystkie jego miejsca zerowe są rzeczywiste, jednokrotne i leżą wewnątrz przedziału A.

   Wskazówka: przy założeniu, że liczba j punktów w przedziale A, w otoczeniu których wielomian p_k zmienia znak jest mniejsza niż k, określ wielomian stopnia j, którego to są miejsca zerowe, i zbadaj iloczyn skalarny tego wielomianu i p_k.

9. Spróbuj udowodnić (odpowiednio przeformułowane) twierdzenie na str. 10.10 ze słabszymi założeniami: funkcja f jest tylko klasy $C^1[a, b]$, ale jej pochodna spełnia warunek Lipschitza.

10. Równanie ($\Diamond$) wyprowadzone w dowodzie twierdzenia o aproksymacyjnych własnościach kubicznych funkcji sklejanych można wyprowadzić w inny sposób, za pomocą różnic dzielonych.

    Niech p_i oznacza wielomian trzeciego stopnia opisujący funkcję sklejaną w przedziale $[u_i, u_{i+1}]$. Dla $k \in \{i, i+1\}$ oznaczamy $a_k = p_i(u_k)$, $b_k = p_i'(u_k)$, $c_k = p_i''(u_k)$. Niech $d_i = p_i'''(x) = \text{const}$; liczba $\frac{1}{6}d_i$ jest różnicą dzieloną trzeciego rzędu wielomianu p_i na dowolnych węzłach. Piszemy schemat różnic dzielonych:

$$
\begin{array}{l|lclclcl}
u_i & a_i & & & & & & \\
 & & \searrow & & & & & \\
u_i & a_i & \to & b_i & & & & \\
 & & \searrow & & \searrow & & & \\
u_i & a_i & \to & b_i & \to & \frac{1}{2}c_i & & \\
 & & \searrow & & \searrow & & \searrow & \\
u_{i+1} & a_{i+1} & \to & p_i[u_i, u_{i+1}] & \to & p_i[u_i, u_i, u_{i+1}] & \to & \frac{1}{6}d_i \\
 & & \searrow & & \searrow & & \searrow & \\
u_{i+1} & a_{i+1} & \to & b_{i+1} & \to & p_i[u_i, u_{i+1}, u_{i+1}] & \to & \frac{1}{6}d_i \\
 & & \searrow & & \searrow & & \searrow & \\
u_{i+1} & a_{i+1} & \to & b_{i+1} & \to & \frac{1}{2}c_{i+1} & \to & \frac{1}{6}d_i
\end{array}
$$

    Mamy stąd dwa równania:

$$
\begin{aligned}
p_i[u_i, u_i, u_{i+1}] - \frac{1}{2}c_i &= p_i[u_i, u_{i+1}, u_{i+1}] - p_i[u_i, u_i, u_{i+1}], \\
\frac{1}{2}c_{i+1} - p_i[u_i, u_{i+1}, u_{i+1}] &= p_i[u_i, u_{i+1}, u_{i+1}] - p_i[u_i, u_i, u_{i+1}].
\end{aligned}
$$

    Po rozwinięciu liczników różnic dzielonych i pomnożeniu równań przez mianownik $h_i = u_{i+1} - u_i$ otrzymamy z nich układ równań

$$
\begin{aligned}
b_{i+1} - 3p_i[u_i, u_{i+1}] + 2b_i + \frac{h_i}{2}c_i &= 0, \\
-2b_{i+1} + 3p_i[u_i, u_{i+1}] - b_i + \frac{h_i}{2}c_{i+1} &= 0.
\end{aligned}
$$

    Biorąc ich odpowiednią kombinację liniową, po prostych przekształceniach otrzymamy

$$
b_i = p_i[u_i, u_{i+1}] - \frac{h_i}{6}c_{i+1} - \frac{h_i}{3}c_i.
$$

    Zastępując w powyższym rachunku wielomian p_i przez p_{i-1}, który ma w węźle u_i tę samą wartość i pierwsze dwie pochodne, oraz węzeł u_{i+1} przez u_{i-1} (uwaga: $u_{i-1} - u_i = -h_{i-1}$), dostaniemy

$$
b_i = p_{i-1}[u_i, u_{i-1}] + \frac{h_{i-1}}{6}c_{i-1} + \frac{h_{i-1}}{3}c_i.
$$

    Wystarczy napisać równość wyrażeń po prawych stronach i uporządkować ją.

11. Udowodnione na wykładzie twierdzenie o aproksymacyjnych własnościach kubicznych funkcji sklejanych można jeszcze „wzmocnić”:

Twierdzenie. *Niech* $f \in C^2[a,b]$ *i niech* s *oznacza kubiczną funkcję sklejaną z węzłami* $u_0 = a < u_1 < \cdots < u_{N-1} < u_N = b$, *taką że* $s(u_i) = f(u_i)$ *dla* $i = 0, \dots, N$. *Jeśli pochodna drugiego rzędu funkcji* f *spełnia warunek Lipschitza ze stałą* L_2, *a ponadto* $|f''(a) - s''(a)| \leqslant 3L_2h$ *i* $|f''(b) - s''(b)| \leqslant 3L_2h$, *gdzie* $h = \max_i(u_{i+1} - u_i)$, *to dla każdego* $x \in [a,b]$ *są spełnione nierówności*

$$|f(x) - s(x)| \leqslant \frac{7}{16}L_2h^3, \quad |f'(x) - s'(x)| \leqslant \frac{7}{4}L_2h^2, \quad |f''(x) - s''(x)| \leqslant \frac{7}{2}L_2h.$$

Szkic dowodu. Przyjmiemy takie same oznaczenia, jak w dowodzie twierdzenia na wykładzie. Ponadto oznaczymy $d_i = s''(u_i) - f''(u_i)$. Wstawiając $d_i + f''(u_i)$ w miejsce $c_i = s''(u_i)$ do równania ($\diamond$), dostaniemy dla $i = 1, \dots, N-1$

$$\frac{h_{i-1}}{h_{i-1}+h_i}d_{i-1} + 2d_i + \frac{h_i}{h_{i-1}+h_i}d_{i+1} =$$
$$6f[u_{i-1},u_i,u_{i+1}] - \frac{h_{i-1}}{h_{i-1}+h_i}f''(u_{i-1}) - 2f''(u_i) - \frac{h_i}{h_{i-1}+h_i}f''(u_{i+1}).$$

Niech q_i oznacza wyrażenie po prawej stronie tej równości.
Istnieje $\xi_i \in [u_{i-1}, u_{i+1}]$, takie że $6f[u_{i-1},u_i,u_{i+1}] = 3f''(\xi_i)$. Możemy zatem napisać

$$q_i = \frac{h_{i-1}}{h_{i-1}+h_i}\big(f''(\xi_i) - f''(u_{i-1})\big) + 2\big(f''(\xi_i) - f''(u_i)\big) +$$
$$\frac{h_i}{h_{i-1}+h_i}\big(f''(\xi_i) - f''(u_{i+1})\big).$$

Ze spełnienia warunku Lipschitza przez f'' wynika, że

$$|f''(\xi_i) - f''(u_{i-1})| \leqslant L_2|\xi_i - u_{i-1}|, \quad |f''(\xi_i) - f''(u_i)| \leqslant L_2\max(h_{i-1},h_i),$$
$$|f''(\xi_i) - f''(u_{i+1})| \leqslant L_2|\xi_i - u_{i+1}|.$$

Możemy oszacować

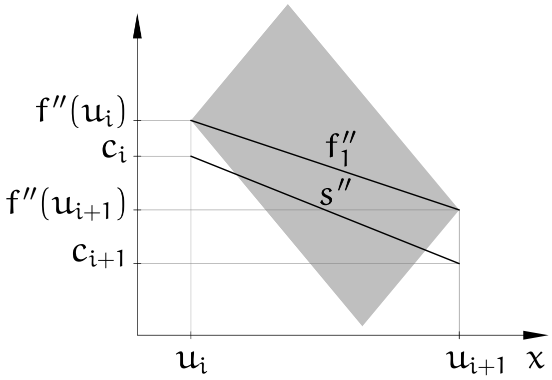
$$\left|\frac{h_{i-1}}{h_{i-1}+h_i}\big(f''(\xi_i) - f''(u_{i-1})\big) + \frac{h_i}{h_{i-1}+h_i}\big(f''(\xi_i) - f''(u_{i+1})\big)\right| \leqslant$$
$$\frac{\max(h_{i-1},h_i)}{h_{i-1}+h_i}L_2\big(|\xi_i - u_{i-1}| + |\xi_i - u_{i+1}|\big) = L_2\max(h_{i-1},h_i)$$

i na tej podstawie

$$|q_i| \leqslant 3L_2\max(h_{i-1},h_i) \leqslant 3L_2h.$$

Wybieramy takie j, że $|d_j| \geqslant |d_i|$ dla $i = 0, \ldots, N$, i wykonując ten sam chwyt, co na wykładzie, znajdujemy oszacowanie $|d_i| \leqslant |d_j| \leqslant 3L_2h$ dla każdego i.

Teraz przedstawiamy f'' w postaci sumy $f_1'' + f_2''$, której składnik f_1'' jest ciągłą funkcją sklejaną stopnia 1 (z węzłami $u_0, \ldots, u_N$), zaś $f_2''(u_i) = 0$ dla $i = 0, \ldots, N$. Niech $x \in [u_i, u_{i+1}]$. Szacujemy



$$|f''(x) - s''(x)| \leqslant |f_2''(x)| + |f_1''(x) - s''(x)|.$$

Składnik $|f_2''(x)|$ możemy oszacować tak (uzasadnienie można dostrzec na rysunku):

$$|f_2''(x)| \leqslant \frac{L_2}{2}h_i \leqslant \frac{L_2}{2}h,$$

natomiast $|f_1''(x) - s''(x)| \leqslant \max(|d_i|, |d_{i+1}|) \leqslant |d_j| \leqslant 3L_2h$, ponieważ obie funkcje, f_1'' i s'', są sklejane pierwszego stopnia z węzłami $u_0, \ldots, u_N$. A więc istotnie, $|f''(x) - s''(x)| \leqslant \frac{7}{2}L_2h$ dla każdego $x \in [a, b]$.

Aby udowodnić pozostałe nierówności stanowiące tezę twierdzenia, wystarczy powtórzyć końcowe rachunki z dowodu na wykładzie, wykorzystując nowe oszacowanie pochodnej drugiego rzędu funkcji błędu, $e(x) = f(x) - s(x)$, tzn. $|e''(x)| \leqslant \frac{7}{2}L_2h$ (zamiast $|e''(x)| \leqslant 4M_2$; stała M_2 w dowodzonym teraz twierdzeniu jest nieistotna). □

Jeśli zatem funkcja f spełnia założenia, to można, wybierając dostatecznie gęsty ciąg węzłów (wyznaczający dostatecznie małe h), otrzymać interpolacyjną kubiczną funkcję sklejaną, której pochodne rzędu 1 i 2 przybliżają pochodne funkcji f z dowolnie małym błędem — ale trzeba poświęcić należytą uwagę warunkom brzegowym. W tym przypadku naturalne funkcje sklejane na ogół nie są dobrym rozwiązaniem.

Numeryczne obliczanie całek

Def. Niech f oznacza pewną funkcję określoną w przedziale $[a,b]$. Kwadratura jest to kombinacja liniowa wartości funkcji f w pewnych punktach $x_i \in [a,b]$, zwanych węzłami kwadratury:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i).$$

Liczby A_i są nazywane współczynnikami kwadratury.
Ogólniejsza definicja określa kwadraturę jako kombinację liniową wartości funkcji f i jej pochodnych w węzłach kwadratury.

Kwadratura jest zatem funkcjonałem liniowym na przestrzeni funkcji określonych w przedziale $[a,b]$, podobnie jak całka oznaczona[29]:

$$I(f) = \int_a^b f(x)\rho(x)\,dx.$$

W odróżnieniu od całki, mogąc obliczać wartości funkcji f w dowolnych punktach przedziału $[a,b]$, można obliczyć wartość kwadratury za pomocą skończenie wielu działań arytmetycznych. Numeryczne obliczanie całek polega na obliczaniu kwadratur. Ważne jest zapewnienie dostatecznej dokładności, tj. dostatecznie małego błędu aproksymacji całki przez kwadraturę. Temu celowi służy wybór węzłów i współczynników kwadratury. Jak zwykle, skuteczność wyboru zależy od własności funkcji, które mamy zamiar całkować.

Kwadratury interpolacyjne

Kwadratura interpolacyjna jest całką z wielomianu interpolacyjnego Lagrange'a lub Hermite'a funkcji f z węzłami w przedziale $[a,b]$. Jeśli jest to wielomian interpolacyjny Lagrange'a (tj. węzły są jednokrotne, obliczamy w nich tylko wartości funkcji f), to kwadratura ma współczynniki

$$A_i = \int_a^b \prod_{j\in\{0,\ldots,n-1\}\setminus\{i\}} \frac{x - x_j}{x_i - x_j}\rho(x)\,dx.$$

Wśród kwadratur interpolacyjnych wyróżniamy kwadratury Newtona-Cotesa, których węzły dzielą przedział $[a,b]$ na części o równych długościach (kwadratury te określa się z wagą $\rho(x) = 1$), kwadratury Gaussa, których węzły są miejscami

[29] Obliczając całkę, w pewnych przypadkach wygodnie jest wyróżnić pewien czynnik w funkcji podcałkowej, $\rho(x)$, i traktować go jako wagę, choć zwykle przyjmuje się wagę $\rho(x) = 1$. Funkcje wagowe, które nie są stałe, przydają się zwłaszcza podczas obliczania całek w przedziale nieograniczonym lub całek z funkcji nieograniczonych. Zakładamy, że funkcja ρ spełnia warunki podane w wykładzie o aproksymacji, i w szczególności jest nieujemna.

zerowymi wielomianów ortogonalnych, a także inne kwadratury, dobierane specjalnie do zastosowań.

Błąd kwadratury jest to oczywiście różnica $I(f) - Q(f)$, która zależy od funkcji f. Błąd kwadratury interpolacyjnej opartej na n węzłach można oszacować, obliczając całkę z wyrażenia opisującego resztę interpolacji:

$$|I(f) - Q(f)| \leqslant \frac{M_n}{n!} \int_a^b |p_n(x)|\rho(x)\,dx,$$

ale to oszacowanie jest poprawne, jeśli funkcja f jest klasy $C^n[a,b]$, i możemy go użyć bezpośrednio, jeśli umiemy znaleźć stałą M_n, taką że $\|f^{(n)}\|_\infty \leqslant M_n$.

Def. Rząd kwadratury jest to liczba r, taka że kwadratura ma tę samą wartość co całka dla każdego wielomianu stopnia mniejszego niż r oraz inną wartość niż całka dla pewnego wielomianu stopnia r.

Z definicji kwadratury interpolacyjnej natychmiast wynika, że jej rząd jest nie mniejszy niż liczba węzłów. Rząd żadnej kwadratury opartej na n węzłach nie może być większy niż $2n$, ponieważ jeśli p_n jest wielomianem stopnia n, którego miejscami zerowymi są wszystkie węzły, to mamy $Q(p_n^2) = 0$ oraz $I(p_n^2) > 0$.

Możemy wybrać pewien ciąg kwadratur $Q_1, Q_2, \ldots$, np. kwadratur Newtona-Cotesa coraz wyższych rzędów, i zbadać zbieżność ciągu liczb $Q_1(f), Q_2(f), \ldots$ dla funkcji f spełniającej określone warunki (np. funkcji ciągłej). Chciałoby się, aby ten ciąg miał granicę, równą $I(f)$; jeśli ją ma, to istotna jest szybkość zbieżności do tej granicy. Korzystając m.in. z twierdzenia Weierstrassa, można udwodnić

Twierdzenie. *Ciąg* $Q_1(f), Q_2(f), \ldots$ *jest zbieżny do granicy* $I(f)$ *dla dowolnej funkcji ciągłej* f *wtedy i tylko wtedy, gdy jest zbieżny dla każdego wielomianu i istnieje stała* K, *taka że suma wartości bezwzględnych współczyników każdej kwadratury w rozpatrywanym ciągu jest mniejsza niż* K.

Pierwszy warunek podany w twierdzeniu jest spełniony przez każdy ciąg kwadratur interpolacyjnych coraz wyższych rzędów, natomiast aby spełnić drugi warunek, wystarczy zapewnić, że współczynniki każdej kwadratury są nieujemne. Niestety, ciąg kwadratur Newtona-Cotesa tego warunku nie spełnia, co więcej, sumy wartości bezwzględnych współczynników tych kwadratur rosną nieograniczenie. Praktycznie użyteczne kwadratury Newtona-Cotesa mają tylko kilka (mniej niż 8) węzłów. Zbadamy dwie najprostsze z nich.

<u>Kwadratura trapezów</u> oparta jest na dwóch węzłach, będących końcami przedziału $[a, b]$:

$$T(f) = \frac{b-a}{2}\big(f(a) + f(b)\big).$$

Łatwo jest sprawdzić, że rząd tej kwadratury jest równy 2. Jeśli funkcja f jest klasy $C^2[a, b]$, to $p_2(x) = (x-a)(x-b)$ i mamy oszacowanie błędu

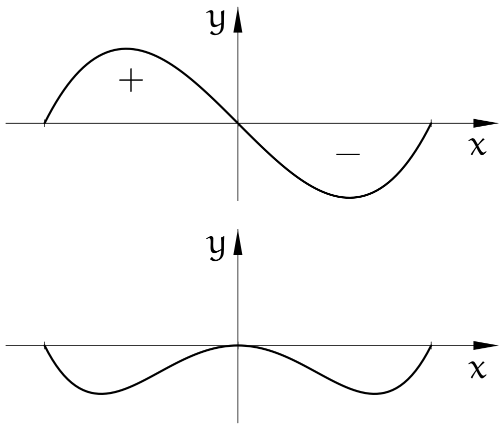
$$|I(f) - T(f)| \leqslant \frac{M_2}{2}\int_a^b |p_2(x)|\,dx = \frac{M_2}{12}(b-a)^3,$$

ze stałą M_2, taką że $|f''(x)| \leqslant M_2$ dla każdego $x \in [a, b]$.

<u>Kwadratura Simpsona</u> oparta jest na trzech węzłach: końcach i środku przedziału $[a, b]$; oznaczymy $c = (a+b)/2$:

$$S(f) = \frac{b-a}{6}\big(f(a) + 4f(c) + f(b)\big).$$

Okazuje się, że rząd kwadratury Simpsona jest równy 4. Jest tak dlatego, że to jest kwadratura interpolacyjna, której środkowy węzeł jest dwukrotny, ale współczynnik, przez który należałoby pomnożyć $f'(c)$, jest równy 0. Inne wyjaśnienie tego zjawiska jest na rysunku obok; jako kwadratura interpolacyjna z trzema węzłami, kwadratura Simpsona jest dokładna dla wszystkich wielomianów stopnia 2 i mniejszych. Dla pokazanego na rysunku wielomianu stopnia 3 kwadratura Simpsona też jest dokładna, ale stąd wynika, że jest dokładna dla wszystkich wielomianów trzeciego stopnia. Jest oczywiste, że kwadratura Simpsona nie jest równa całce z pokazanego niżej wielomianu stopnia 4, ale to oznacza, że nierówność zachodzi dla każdego wielomianu, którego stopień jest równy 4. Błąd kwadratury Simpsona możemy oszacować na dwa sposoby:



$$|I(f) - S(f)| \leqslant \frac{M_3}{6}\int_a^b |(x-a)(x-c)(x-b)|\,dx = \frac{M_3}{192}(b-a)^4,$$
$$|I(f) - S(f)| \leqslant \frac{M_4}{24}\int_a^b |(x-a)(x-c)^2(x-b)|\,dx = \frac{M_4}{2880}(b-a)^5,$$

gdzie M_3 i M_4 to oszacowania wartości bezwzględnych pochodnych trzeciego i czwartego rzędu funkcji f w przedziale $[a, b]$. Oczywiście, każdego z tych oszacowań możemy używać pod warunkiem, że odpowiednia pochodna funkcji f jest ciągła.

Zamiana zmiennych

Jeśli $f(u) = g(x)$ dla $x = su + t$, gdzie $s > 0$ i t są ustalonymi liczbami, oraz $c = sa + t$, $d = sb + t$, to

$$\int_a^b f(u)\rho(u)\,du = \frac{1}{s}\int_c^d g(x)\rho((x-t)/s)\,dx, \quad \text{oraz}$$
$$Q_1(f) = \sum_{i=0}^{n-1} A_i f(u_i) = \frac{1}{s}\sum_{i=0}^{n-1} sA_i g(su_i + t) = \frac{1}{s}Q_2(g).$$

W ten sposób, mając dowolną kwadraturę Q_1:

$$Q_1(f) = \sum_{i=0}^{n-1} A_i f(u_i) \approx \int_a^b f(u)\rho(u)\,du,$$

możemy otrzymać nową kwadraturę Q_2:

$$Q_2(g) = \sum_{i=0}^{n-1} B_i g(x_i) \approx \int_c^d g(x)\rho((x-t)/s)\,dx,$$

z węzłami $x_i = su_i + t$ i współczynnikami $B_i = sA_i$.

Kwadratury Q_1 i Q_2 mają ten sam rząd. Ponadto, mając oszacowanie błędu kwadratury Q_1, podobne do podanych wcześniej oszacowań dla kwadratur trapezów i Simpsona, można podać oszacowanie błędu kwadratury Q_2. Mianowicie, jeśli funkcje f i g są klasy C^k w swoich przedziałach całkowania[30] i błąd kwadratury Q_1 ma górne oszacowanie o postaci

$$C(b-a)^{k+1} \max_{u\in[a,b]} |f^{(k)}(u)|,$$

to błąd kwadratury Q_2 jest nie większy niż

$$C(c-d)^{k+1} \max_{x\in[c,d]} |g^{(k)}(x)|,$$

z tą samą stałą C.

Kwadratury Gaussa

Niech $A \subset \mathbb{R}$ oznacza (ograniczony lub nieograniczony) przedział całkowania, niech ρ oznacza funkcję wagową i niech $p_0, p_1, \ldots$ będzie ciągiem wielomianów

[30] Tu rozważamy k w ogólności inne niż liczba węzłów n. Dla kwadratury interpolacyjnej rzędu $r \geqslant n$ można wskazać takie oszacowania dla $k = \{n, \ldots, r\}$, przykładem jest kwadratura Simpsona opisana wcześniej i kwadratury Gaussa, opisane dalej.

ortogonalnych w sensie iloczynu skalarnego

$$\langle f, g\rangle_\rho \stackrel{\text{def}}{=} \int_A f(x)g(x)\rho(x)\,dx.$$

Ustalmy liczbę n i określmy kwadraturę interpolacyjną Q z węzłami, które są miejscami zerowymi $x_0, \ldots, x_{n-1}$ wielomianu p_n; możemy to zrobić, bo miejsca zerowe tego wielomianu są jednokrotne i znajdują się w przedziale A. Dowolny wielomian w stopnia mniejszego niż $2n$ możemy przedstawić w postaci

$$w(x) = p_n(x)a(x) + r(x),$$

gdzie a i r to iloraz i reszta z dzielenia wielomianu w przez p_n; stopnie wielomianów a i r są mniejsze niż n. Dzięki temu zachodzą równości

$$\begin{aligned} I(w) = \int_A w(x)\rho(x)\,dx &= \int_A \big(p_n(x)a(x) + r(x)\big)\rho(x)\,dx = \\ &\underbrace{\langle p_n, a\rangle_\rho}_{=0} + \int_A r(x)\rho(x)\,dx = Q(r) = Q(w), \end{aligned}$$

ponieważ wartości wielomianów w i r we wszystkich węzłach kwadratury są jednakowe. Skonstruowana w ten sposób kwadratura jest zatem rzędu $2n$.

Kwadratury interpolacyjne, których węzły są miejscami zerowymi wielomianów ortogonalnych (odpowiadających danemu przedziałowi i funkcji wagowej) są nazywane <u>kwadraturami Gaussa</u>; nazwisko rodziny, do której odpowiedni wielomian należy, jest dołączane do nazwiska Gauss, i w ten sposób mówi się np. o <u>kwadraturach Gaussa-Legendre'a</u>:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-1}^{1} f(x)\,dx,$$

<u>kwadraturach Gaussa-Czebyszewa</u>:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-1}^{1} \frac{f(x)}{\sqrt{1-x^2}}\,dx,$$

<u>kwadraturach Gaussa-Hermite'a</u>:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{-\infty}^{\infty} f(x)e^{-x^2}\,dx$$

i <u>kwadraturach Gaussa-Laguerre'a</u>:

$$Q(f) = \sum_{i=0}^{n-1} A_i f(x_i) \approx \int_{0}^{\infty} f(x)e^{-x}\,dx.$$

Konstruując kwadraturę Gaussa, na ogół trzeba jej węzły znaleźć, rozwiązując numerycznie równanie $p_n(x) = 0$. Współczynniki kwadratury Gaussa można obliczyć tak, jak współczynniki dowolnej kwadratury interpolacyjnej, lub na podstawie wzoru

$$A_i = \frac{1}{\sum_{k=0}^{n-1} \tilde{p}_k(x_i)^2},$$

w którym występują wielomiany ortonormalne $\tilde{p}_k(x) = p_k(x)/\|p_k\|_\rho$. Wygodnie jest użyć w tym obliczeniu formuły trójczłonowej. Zatem współczynniki każdej kwadratury Gaussa są dodatnie i z podanego wcześniej twierdzenia wynika, że dla dowolnej funkcji ciągłej ciąg kwadratur Gaussa coraz wyższych rzędów zbiega do całki z tej funkcji (z odpowiednią wagą).

Największe znaczenie praktyczne mają kwadratury Gaussa-Legendre'a, ponieważ najczęściej oblicza się całki w skończonym przedziale, z wagą $\rho(x) = 1$. Najprostsza kwadratura Gaussa-Legendre'a jest iloczynem długości przedziału całkowania i wartości funkcji w środku tego przedziału. Jest to więc kwadratura rzędu 2, oparta na jedym węźle.

Niech p_n oznacza wielomian ortogonalny Legendre'a stopnia n, wyskalowany tak, aby jego współczynnik wiodący był równy 1. Błąd aproksymacji jednostajnej funkcji f klasy $C^{2n}[-1, 1]$ przez wielomian interpolacyjny Hermite'a h_{2n-1} stopnia $2n - 1$, oparty na węzłach kwadratury Gaussa-Legendre'a (czyli miejscach zerowych wielomianu p_n), z których każdy liczymy dwukrotnie[31], ma oszacowanie

$$\max_{x\in[-1,1]} |f(x) - h_{2n-1}(x)| \leqslant \frac{M_{2n}}{(2n)!} p_n(x)^2,$$

gdzie $M_{2n} = \max_{x\in[-1,1]} |f^{(2n)}(x)|$. Niech

$$C_n = \int_{-1}^{1} p_n(x)^2 \,dx.$$

Po dokonaniu zamiany zmiennych, możemy oszacować błąd kwadratury Gaussa-Legendre'a rzędu $2n$ dla przedziału $[a, b]$:

$$|I(f) - Q(f)| \leqslant C_n \frac{M_{2n}}{(2n)!} \left(\frac{b-a}{2}\right)^{2n+1},$$

[31] Kwadratura Gaussa jest w istocie całką z wielomianu interpolacyjnego Hermite'a, którego wszystkie węzły mają krotność 2. Oprócz wartości funkcji podcałkowej należałoby więc uwzględnić wartości jej pochodnej w węzłach, ale wybór węzłów zapewnia, że wszystkie współczynniki, przez które trzeba pomnożyć wartości pochodnej, są równe 0 — zobacz też opis kwadratury Simpsona.

przy czym teraz M_{2n} oznacza oszacowanie pochodnej rzędu $2n$ funkcji f w przedziale $[a, b]$.

Kwadratury złożone

Tak jak w aproksymacji jednostajnej funkcji, skutecznym sposobem zmniejszenia błędu aproksymacji całki przez kwadraturę jest podzielenie przedziału całkowania na krótsze podprzedziały i obliczenie sumy kwadratur interpolacyjnych dla tych podprzedziałów. W ten sposób otrzymuje się kwadratury złożone. Błąd takiej kwadratury jest sumą błędów kwadratur dla podprzedziałów, przy czym błędy te mogą mieć różne znaki, a zatem mogą się znosić. Oszacowania błędów kwadratur złożonych zwykle są sumami oszacowań błędów w podprzedziałach, przez co często bywają pesymistyczne.

Dodatkową korzyścią z zastosowania kwadratury złożonej jest możliwość podziału przedziału całkowania w punktach nieciągłości funkcji podcałkowej lub jej pochodnych (jeśli punktów tych jest skończenie wiele i je znamy). Wtedy w każdym podprzedziale funkcja podcałkowa ma wyższą klasę ciągłości, co umożliwia stosowanie kwadratur odpowiednio wyższego rzędu. Ponadto, po dokonaniu podziału można stosować w podprzedziałach różne kwadratury, dostosowane do zachowania funkcji podcałkowej w tych podprzedziałach. Kolejna możliwość to adaptacja — dla konkretnej funkcji można znaleźć oszacowania błędów w poszczególnych podprzedziałach, i na tej podstawie podejmować decyzję o dalszym (rekurencyjnym) podziale niektórych z nich.

Kwadratury w podprzedziałach konstruujemy za pomocą opisanej wcześniej zamiany zmiennych. Zobaczmy przykłady kwadratur z podziałem przedziału $[a, b]$ na N części o tej samej długości $h = (b - a)/N$.

Złożona kwadratura trapezów powstaje w ten sposób, że w każdym z podprzedziałów przedziału $[a, b]$ stosujemy kwadraturę trapezów. W ten sposób otrzymamy liczbę

$$T_h(f) = h\Big(\frac{1}{2}f(x_0) + \sum_{i=1}^{N-1} f(x_i) + \frac{1}{2}f(x_N)\Big),$$

gdzie $x_i = a + ih$. Jeśli funkcja f jest klasy $C^2[a, b]$ i $|f''(x)| \leqslant M_2$ dla każdego $x \in [a, b]$, to wartość bezwzględna lokalnego błędu kwadratury trapezów w przedziale $[x_i, x_{i+1}]$ nie przekracza $\frac{M_2}{12}h^3$, a zatem suma tych błędów ma oszacowanie

$$|I(f) - T_h(f)| \leqslant \frac{M_2}{12}(b-a)h^2.$$

Złożoną kwadraturę Simpsona otrzymujemy analogicznie. Oznaczmy $x_i = a + ih/2$ dla $i = 0, \ldots, 2N$. Suma kwadratur Simpsona w N podprzedziałach o długości h jest równa

$$S_h(f) = \frac{h}{6}\Big(f(x_0) + 4f(x_1) + \sum_{i=1}^{N-1}\big(2f(x_{2i}) + 4f(x_{2i+1})\big) + f(x_{2N})\Big),$$

zaś dla funkcji f odpowiednio klasy $C^3[a,b]$ i $C^4[a,b]$ błąd ma oszacowania

$$|I(f) - S_h(f)| \leqslant \frac{M_3}{192}(b-a)h^3,$$
$$|I(f) - S_h(f)| \leqslant \frac{M_4}{2880}(b-a)h^4.$$

Konstruowanie złożonych kwadratur Gaussa jest utrudnione, jeśli funkcja wagowa nie jest stała, dlatego powyższe podejście stosuje się tylko do kwadratur Gaussa-Legendre'a. Jeśli funkcja f jest klasy $C^{2n}[a,b]$, to możemy w każdym przedziale o długości h użyć kwadratury Gaussa-Legendre'a opartej na n węzłach i wtedy dostaniemy oszacowanie błędu o postaci

$$|I(f) - Q_h(f)| \leqslant C_n M_{2n}(b-a)h^{2n},$$

w którym stała C_n zależy tylko od rzędu kwadratury. Jak widać, dla $h \to 0$ błąd bardzo szybko dąży do zera. Jeśli funkcja f nie ma ciągłych pochodnych aż tak wysokiego rzędu, to błąd nadal dąży do zera, choć wolniej.

Ekstrapolacja Richardsona i metoda Romberga

Niech f oznacza funkcję klasy $C^{2n+2}[a,b]$. Dowodzi się, że błąd złożonej kwadratury trapezów, z przedziałem $[a,b]$ podzielonym na podprzedziały o jednakowej długości h, można wyrazić wzorem

$$I(f) - T_h(f) = c_1h^2 + c_2h^4 + \cdots + c_nh^{2n} + O(h^{2n+2}),$$

zwanym wzorem sumacyjnym Eulera-Maclaurina. Współczynniki $c_1, \ldots, c_n$ zależą od pochodnych funkcji f w przedziale $[a,b]$, ale nie zależą od długości podprzedziałów.

Możemy ten wzór przepisać dla złożonej kwadratury trapezów z dwukrotnie drobniejszym podziałem przedziału całkowania:

$$I(f) - T_{h/2}(f) = \frac{c_1}{4}h^2 + \frac{c_2}{16}h^4 + \cdots + \frac{c_n}{4^n}h^{2n} + O(h^{2n+2}),$$

Jeśli strony powyższego wzoru pomnożymy przez $4/3$ i odejmiemy od nich strony wzoru dla kwadratury z podprzedziałami o długości h pomnożone przez $1/3$, to otrzymamy równość

$$I(f) - \left(\frac{4}{3}T_{h/2}(f) - \frac{1}{3}T_h(f)\right) = d_2h^4 + \cdots + d_nh^{2n} + O(h^{2n+2}).$$

Kombinacja liniowa[32] $T_h^{(1)}(f) = 4/3T_{h/2}(f) - 1/3T_h(f)$ jest kwadraturą, której dominujący składnik błędu jest rzędu h^4, zatem znacznie szybciej maleje podczas zmniejszania h. Opisany sposób wyeliminowania dominującego składnika błędu (który można stosować także w innych przypadkach, gdy błąd jest opisany za pomocą szeregu potęgowego) jest nazywany ekstrapolacją Richardsona.

Ekstrapolację Richardsona możemy iterować. Mając kwadratury $T_h^{(j)}$ i $T_{h/2}^{(j)}$, których dominujące składniki błędów są proporcjonalne do h^{2j+2}, określamy kwadraturę

$$T_h^{(j+1)}(f) = \frac{2^{2j+2}}{2^{2j+2}-1}T_{h/2}^{(j)}(f) - \frac{1}{2^{2j+2}-1}T_h^{(j)}(f),$$

której błąd ma dominujący składnik błędu h^{2j+4}. Oparta na tym pomyśle metoda numerycznego całkowania jest nazywana metodą Romberga. Podprogram obliczający całkę, dla ustalonego h, oblicza kwadratury $T_h(f)$ i $T_{h/2}(f)$ i oblicza kwadraturę $T_h^{(1)}(f)$. Wyrażenie $|T_h(f) - T_{h/2}(f)|$ może być przyjęte za oszacowanie błędu, co jest analogią do przyrostowego kryterium stopu w metodach numerycznych rozwiązywania równań nieliniowych. Jeśli to oszacowanie jest zbyt duże, to obliczana jest kwadratura $T_{h/4}(f)$, a następnie $T_{h/2}^{(1)}(f)$ i $T_h^{(2)}(f)$ itd. Obliczenie przebiega zgodnie ze schematem

$$\begin{array}{ccccccc}
T_h(f) & & & & & & \\
 & \searrow & & & & & \\
T_{h/2}(f) & \to & T_h^{(1)}(f) & & & & \\
 & \searrow & & \searrow & & & \\
T_{h/4}(f) & \to & T_{h/2}^{(1)}(f) & \to & T_h^{(2)}(f) & & \\
\vdots & & \vdots & & \vdots & \ddots & \\
 & \searrow & & \searrow & & \searrow & \\
T_{h/2^k}(f) & \to & T_{h/2^{k-1}}^{(1)}(f) & \to & \ldots & \to & T_h^{(k)}(f)
\end{array}$$

Za oszacowanie błędu każdej kwadratury otrzymanej przez ekstrapolację możemy przyjąć różnicę kwadratur, na podstawie których została ona obliczona. Zauważmy, że po każdym zmniejszeniu długości podprzedziałów dla kwadratury trapezów wartości funkcji podcałkowej wystarczy tylko obliczyć tylko w nowych węzłach i nie ma potrzeby przechowywania wartości funkcji f w tablicy.

[32] Co ciekawe, to jest złożona kwadratura Simpsona, S_h. Natomiast dalej opisane kwadratury $T_h^{(2)}, T_h^{(3)}, \ldots$ *nie są* złożonymi kwadraturami interpolacyjnymi.

Całkowanie funkcji wielu zmiennych

Znane z analizy twierdzenie Fubiniego umożliwia sprowadzenie zadania obliczenia całki z funkcji f określonej w wielowymiarowym obszarze A do obliczenia całek jednowymiarowych. Analogicznie można postępować z kwadraturami. Jest to szczególnie proste, gdy obszar A jest kostką. Powiedzmy, że jest to prostokąt: $A = [a, b] \times [c, d]$. Mając kwadratury przybliżające całki w przedziałach $[a, b]$ i $[c, d]$, odpowiednio z węzłami $x_0, \ldots, x_{n-1}$ i $y_0, \ldots, y_{m-1}$ oraz współczynnikami $A_0, \ldots, A_{n-1}$ i $B_0, \ldots, B_{m-1}$, możemy obliczyć

$$Q(f) = \sum_{i=0}^{n-1} A_i \sum_{j=0}^{m-1} B_j f(x_i, y_j) \approx \int_a^b \left(\int_c^d f(x, y)\,dy \right) dx.$$

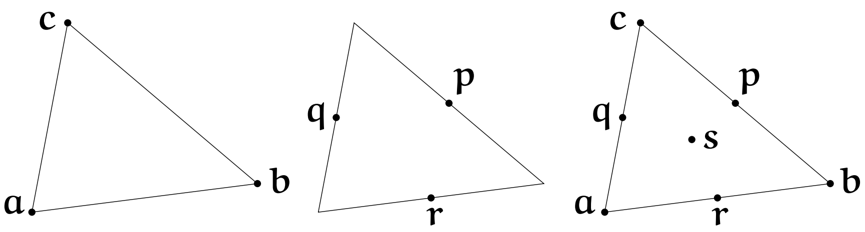
Twierdzenie Fubiniego umożliwia obliczanie całek w obszarach dwu- lub więcej wymiarowych, ale bezpośrednie przełożenie go na kwadratury jest kłopotliwe, jeśli obszar całkowania A nie jest kostką. Problem bierze się stąd, że nawet jeśli funkcja $f(x, y)$ jest gładka w obszarze A, to funkcja określona wzorem

$$g(x) = \int_{A \cap l_x} f(x, y)\,dy,$$

w którym $l_x = \{\,(x, y)\colon y \in \mathbb{R}\,\}$, może mieć osobliwości (np. nieograniczone pochodne) — można to zobaczyć na przykładzie funkcji f stałej w obszarze A, który jest kołem. Takie osobliwości nie pozwalają szacować błędów na podstawie oszacowań błędów kwadratur dla funkcji jednej zmiennej. Dlatego całki wielowymiarowe na ogół trzeba przekształcić przed zastosowaniem kwadratury.

Jeśli obszar całkowania jest kołem (lub kulą d-wymiarową), to często dokonuje się zamiany zmiennych, przekształcającej ten obszar na prostokąt (odpowiednio: kostkę d-wymiarową), przez wprowadzenie współrzędnych biegunowych. Jakobian tego przekształcenia też ma osobliwość, którą trzeba uwzględnić w oszacowaniach błędu (ale ta osobliwość może się znosić z ewentualną osobliwością funkcji f w środku kuli — to w każdym razie należy zbadać).

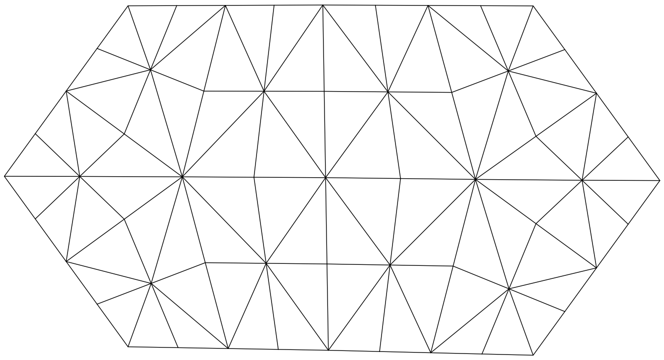
Inne podejście polega na przybliżeniu obszaru całkowania przez pewien obszar $\tilde{A}$, który jest np. wielokątem (wielościanem d-wymiarowym) zawartym w A. Całkę z funkcji f w obszarze A zastępujemy przez kwadraturę przybliżającą całkę w obszarze $\tilde{A}$ — oczywiście, należy zadbać o to, żeby oba błędy aproksymacji były dostatecznie małe. Wielokąt możemy podzielić na tak zwane elementy, np. na trójkąty; w ten sposób dla obszaru $\tilde{A}$ określamy kwadraturę złożoną, która jest sumą kwadratur w poszczególnych elementach.



Przykładowe kwadratury określone wzorami

$$
\begin{aligned}
Q_1(f) &= \frac{T}{3}\big(f(\mathbf{a}) + f(\mathbf{b}) + f(\mathbf{c})\big),\\
Q_2(f) &= \frac{T}{3}\big(f(\mathbf{p}) + f(\mathbf{q}) + f(\mathbf{r})\big),\\
Q_3(f) &= \frac{T}{60}\Big(3\big(f(\mathbf{a}) + f(\mathbf{b}) + f(\mathbf{c})\big) + 8\big(f(\mathbf{p}) + f(\mathbf{q}) + f(\mathbf{r})\big) + 27f(\mathbf{s})\Big),
\end{aligned}
$$

w których T oznacza pole trójkąta o wierzchołkach $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$, środkach boków $\mathbf{p}$, $\mathbf{q}$, $\mathbf{r}$ i środku ciężkości $\mathbf{s}$, są dokładne (tzn. równe całce z funkcji f po tym trójkącie), jeśli funkcja f jest odpowiednio wielomianem stopnia 1, 2 i 3. Błąd kwadratury złożonej można zmniejszyć przez „rozdrobnienie” podziału na elementy.
Zauważmy, że dopuszczenie elementów o mniejszej średnicy umożliwia także zmniejszenie błędu przybliżenia krzywoliniowego obszaru A przez wielokąt, który jest sumą elementów. Warto wspomnieć, że w pewnych przypadkach stosuje się elementy krzywoliniowe, które umożliwiają lepszą aproksymację takich obszarów.



Postępowanie z całkami w obszarach trójwymiarowych może być podobne; wielościan $\tilde{A}$ można podzielić na czworościany i obliczyć kwadratury interpolacyjne w tych czworościanach oraz ich sumę, czyli kwadraturę złożoną. Natomiast numeryczne całkowanie jest bardzo kłopotliwe, jeśli *wymiar* obszaru A jest duży. Istnieją zadania praktyczne (biorące się m.in. z fizyki i ekonomii), w których wymiar d obszaru całkowania jest rzędu kilkuset. Obliczenie całki w kostce d-wymiarowej za pomocą kwadratury otrzymanej analogicznie, jak dla prostokąta, jest niewykonalne. Nawet gdyby w przedziale zmienności każdej zmiennej wybrać tylko dwa węzły, liczba punktów, w których trzeba by obliczyć wartości funkcji podcałkowej, byłaby równa 2^d. Zjawisko wykładniczego wzrostu złożoności obliczeniowej zadania ze wzrostem wymiaru dziedziny funkcji nosi nazwę przekleństwa wymiaru (ang. *dimensionality curse*).

Znanych jest kilka sposobów obliczania przybliżonych wartości całek wielowymiarowych za pomocą wartości funkcji obliczonych w znacznie mniejszej liczbie punktów. Sposób najprostszy i jednocześnie skuteczny dla najszerszej klasy takich zadań wynalazł Ulam w 1946 r. Sposób ten jest znany pod nazwą metody Monte Carlo. Obszar A uznajemy za przestrzeń zadrzeń elementarnych i określamy w nim jednostajny rozkład prawdopodobieństwa. Wtedy funkcja f jest zmienną losową. Iloczyn wartości oczekiwanej tej zmiennej losowej i miary $|A|$ obszaru A jest poszukiwaną całką, $\int_A f$. Dla n niezależnych losowań punktów $\boldsymbol{x}_i \in A$ możemy określić nową zmienną losową wzorem

$$Q(f) = |A|\frac{1}{n}\sum_{i=0}^{n-1} f(\boldsymbol{x}_i).$$

Jest to właśnie kwadratura Monte Carlo; jej wartość oczekiwana też jest równa poszukiwanej całce. Jeśli zmienna losowa f ma wariancję σ^2, to wariancja σ_n^2 kwadratury Monte Carlo jest równa $|A|\sigma^2/n$. Zatem odchylenie standardowe σ_n zmiennej losowej $Q(f)$ jest proporcjonalne do $n^{-1/2}$ i w szczególności nie zależy od wymiaru d obszaru A. Dla dostatecznie dużego n możemy oczekiwać, że błąd jest bardzo mały — z dużym prawdopodobieństwem, ale nie z całkowitą pewnością. Stosując kwadratury Monte Carlo, też często dokonuje się rozmaitych przekształceń (np. zamiany zmiennych) w celu przekształcenia zbioru całkowania na kostkę lub otrzymania funkcji podcałkowej o mniejszej wariancji.

Zadania i problemy

1. Dla kwadratury interpolacyjnej opisanej wzorem

$$Q(f) = A_0 f(a) + A_1 f(c),$$

przybliżającej całkę $\int_a^b f(x)\,dx$, znajdź węzeł c i współczyniki A_0, A_1 tak, aby rząd tej kwadratury był jak największy. Podaj rząd i oszacowanie błędu tej kwadratury.

Wskazówka. Zacznij od przypadku $a = 0$, $b = 1$, następnie uogólnij otrzymane wyniki na przypadek dowolnego przedziału $[a, b]$.

2. Znajdź współczynniki kwadratury

$$Q(f) = A_0 f(a) + B_0 f'(a) + A_1 f(b) + B_1 f'(b),$$

przybliżającej całkę $\int_a^b f(x)\,dx$.
Jaki jest rząd tej kwadratury? Podaj oszacowanie jej błędu.

Wskazówka. Rozwiąż zadanie dla przypadku $a = 0$, $b = 1$, korzystając z wielomianów H_{00}, H_{01}, H_{10} i H_{11}, użytych do konstrukcji interpolacyjnych kubicznych funkcji sklejanych. Następnie uogólnij wynik.

3. Oblicz współczynniki kwadratury Gaussa-Czebyszewa rzędu $2n$.

Wskazówka. Dokonaj zamiany zmiennych, podstawiając $x = \cos t$.

4. Jaka kwadratura wydaje się najodpowiedniejsza do obliczania tzw. zupełnych całek eliptycznych pierwszego i drugiego rodzaju, tj. funkcji określonych wzorami

$$K(k) = \int_0^1 \frac{1}{\sqrt{(1-k^2x^2)(1-x^2)}}\,dx \qquad \text{i} \qquad E(k) = \int_0^1 \sqrt{\frac{1-k^2x^2}{1-x^2}}\,dx$$

dla $k \in [0, 1)$, i jak jej użyć?

5. Przypuśćmy, że należy obliczyć całkę z funkcji f w dwuwymiarowym obszarze A, który leży między wykresami dwóch funkcji ciągłych, c i d, określonych w przedziale $[a, b]$ i takich że $c(x) < d(x)$ dla każdego $x \in [a, b]$. Zaprojektuj algorytm obliczania tej całki za pomocą pewnej kwadratury Q, której węzły i współczynniki są podane w tablicach. Kwadratura Q przybliża całkę z funkcji jednej zmiennej w przedziale $[0, 1]$.

6. Znajdź współczynniki kwadratury złożonej przybliżającej całkę $\int_a^b f(x)\,dx$. Przedział $[a, b]$ jest dzielony na odcinki o długości $h = (b - a)/N$. W każdym z nich jest konstruowana kwadratura interpolacyjna oparta na dwóch węzłach dwukrotnych — końcach odcinka. Znajdź oszacowanie błędu tej kwadratury dla funkcji klasy $C^4[a, b]$ w zależności od N.