

Te materiały można znaleźć pod adresem:

<http://www.is.umk.pl/~kg/zajecia/MetNum2.pdf>

Bibliografia

- D. Kincaid, W. Cheney. Analiza numeryczna. WNT 2006.
- J. Stoer, R. Bulirsch. Wstęp do analizy numerycznej. PWN 1987.
- G. Dahlquist, A. Björck. Metody numeryczne. PWN 1983.
- A. Ralston. Wstęp do analizy numerycznej. PWN 1971.
- B. P. Demidowicz, I. A. Maron, E. Z. Szuwałowa. Metody Numeryczne. PWN 1965.
- Z. Fortuna, B. Macukow, J. Wąsowski. Metody Numeryczne. WNT 1993.
- W. H. Press i in. Numerical Recipes in C (the Art of Scientific Computing).

Plan

1. Metody numeryczne. Błędy.	3
2. Interpolacja.	27
3. Całkowanie numeryczne.	68
4. Aproksymacja.	101
5. Rozwiązywanie równań nieliniowych.	131
6. Rozwiązywanie układów równań liniowych.	164
7. Metody ortogonalizacji.	208
8. Wyznaczanie wartości i wektorów własnych macierzy.	216

Metody numeryczne

- Cel – oszacowanie dokładności wyników obliczeń.
- Błędy ograniczają dokładność.
- Rodzaje *błędów*:
 - *w danych wejściowych* (przyrządów pomiarowych itp.),
 - *zaokrągleń* (ograniczenie dokładności obliczeń),
 - *aproksymacji* (uproszczenie problemu, szacowanie a nie wyliczanie),

Przykład 1:

$$e = 1 + \frac{1}{1!} + \frac{1}{2!} + \frac{1}{3!} + \dots \quad (1)$$

można szacować licząc skończone sumy (powstaje *błąd obcięcia*).

Przykład 2:

całkę oznaczoną liczymy jako skończoną sumę pól obszarów (*błąd dyskretyzacji*).

Reprezentacja liczb

- Systemy liczenia: dwójkowy (binarny), ósemkowy, dziesiętny, szesnastkowy.

system	cyfry	przykład	dziesiętnie
dwójkowy	0 1	101	$1 * 2^2 + 1 * 2^0 =$ $1 * 4 + 1 = 5$
ósemkowy	0 1 2 3 4 5 6 7	270	$2 * 8^2 + 7 * 8^1 =$ $2 * 64 + 7 * 8 = 184$
dziesiętny	0 1 2 3 4 5 6 7 8 9	309	$3 * 10^2 + 9 * 10^0 =$ $3 * 100 + 9 = 309$
szesnastkowy	0 1 2 3 4 5 6 7 8 9 A B C D E F	10A	$1 * 16^2 + 10 * 16^0 =$ $1 * 256 + 10 = 266$

Reprezentacja liczb - c.d.

- Reprezentacja stałopozycyjna
 - w praktyce tylko dla liczb całkowitych
- Reprezentacja zmiennopozycyjna (zmiennoprzecinkowa)
 - mantysa i cecha, liczba = mantysa * 2^{cecha}
 $18.5 = 0.185 * 10^2 = 0.100101 * 2^{101}$
 - rozdzielczość i skończoność komputerowych liczb rzeczywistych
 - typy 4-bajtowe (3 bajty mantysy i 1 bajt cechy), 6-bajtowe, 8-bajtowe (podwójnej precyzji)
 - należy być ostrożnym przy dodawaniu małej liczby zmiennoprzecinkowej do dużej (mała może zniknąć)
 - wartości, które nie są poprawnymi liczbami (NaN - not a number)
 - standard IEEE 754 – dobry opis pod <http://research.microsoft.com/~hollasch/cgindex/coding/ieeefloat.html>

Standard IEEE 754

Reprezentacja bitowa

precyzja	znak	wykładnik	część ułamkowa	przesunięcie
pojedyncza	1 [31]	8 [30-23]	23 [22-00]	127
podwójna	1 [63]	11 [62-52]	52 [51-00]	1023

- reprezentacja binarna – podstawa równa 2
- znak – 0 - liczba dodatnia, 1 - liczba ujemna
- wykładnik i przesunięcie – rzeczywista cecha = wykładnik - przesunięcie
- mantysa
 - *postać znormalizowana liczby* – w mantysie przecinek po pierwszej niezerowej cyfrze ($1 \leq m < 10$)
Przykład: $1.25 * 10^2$, nie $0.125 * 10^3$ ani $125 * 10^0$
 - w systemie dwójkowym niezerowa znaczy jedynka
 - pierwszy bit mantysy zawsze równy 1 – ukryty, reszta to część ułamkowa

Standard IEEE 754 – przykłady

23.75

- postać binarna: $10111.11 = 2^4 + 2^2 + 2^1 + 2^0 + 2^{-1} + 2^{-2}$
- po normalizacji: $1.011111 * 2^4 = 1.011111 * 2^{100}$
- znak 0
 wykładnik z przesunięciem $4 + 127 = 131 = 10000011$
 część ułamkowa 011111
- ostatecznie mamy $0\ 10000011\ 011111000000000000000000$
 czyli $01000001\ 10111110\ 00000000\ 00000000$
 szesnastkowo $0x41BE0000$

-0.7

- postać binarna: $-0.1011001100110011\dots$
- znak = 1, wykładnik = -1 , $127 - 1 = 126 = 01111110$
- ostatecznie mamy $1\ 01111110\ 01100110011001100110011$
 czyli $10111111\ 00110011\ 00110011\ 00110011$
 szesnastkowo $0xBF333333$

Standard IEEE 754 – c.d.

Wartości specjalne – wykładniki z samych jedynek bądź z samych zer

znak	wykładnik (w)	cz. ułamkowa (u)	wartość
0	00..00	00..00	+0
0	00..00	00..01 – 11..11	dodatnia zdenorm. $0.u * 2^{-p+1}$
0	00..01 – 11..10	00..00 – 11..11	dodatnia znorm. $1.u * 2^{w-p}$
0	11..11	00..00	+Infinity (nieskończoność)
0	11..11	00..01 – 01..11	SNaN (Signalling Not A Number)
0	11..11	10..00 – 11..11	QNaN (Quiet Not A Number)
1	00..00	00..00	-0
1	00..00	00..01 – 11..11	ujemna zdenorm. $-0.u * 2^{-p+1}$
1	00..01 – 11..10	00..00 – 11..11	ujemna znorm. $-1.u * 2^{w-p}$
1	11..11	00..00	-Infinity (nieskończoność)
1	11..11	00..01 – 01..11	SNaN
1	11..11	10..00 – 11..11	QNaN

Standard IEEE 754 – c.d.

Operacje specjalne:

Operacja	Wynik
$n / \pm\text{Infinity}$	0
$\pm\text{Infinity} \times \pm\text{Infinity}$	$\pm\text{Infinity}$
$\pm\text{nonzero} / 0$	$\pm\text{Infinity}$
$\text{Infinity} + \text{Infinity}$	Infinity

Operacja	Wynik
$\pm 0 / \pm 0$	NaN
$\text{Infinity} - \text{Infinity}$	NaN
$\pm\text{Infinity} / \pm\text{Infinity}$	NaN
$\pm\text{Infinity} \times 0$	NaN

Zakresy dwójkowo:

precyzja	zdenormalizowane	znormalizowane
pojedyncza	$\pm 2^{-149} \dots (1 - 2^{-23}) * 2^{-126}$	$\pm 2^{-126} \dots (2 - 2^{-23}) * 2^{127}$
podwójna	$\pm 2^{-1074} \dots (1 - 2^{-52}) * 2^{-1022}$	$\pm 2^{-1022} \dots (2 - 2^{-52}) * 2^{1023}$

Zakresy dziesiętne:

precyzja	dziesiętne
pojedyncza	$\pm \sim 10^{-44.85} \dots \sim 10^{38.53}$
podwójna	$\pm \sim 10^{-323.3} \dots \sim 10^{308.3}$

Błąd bezwzględny i względny

Niech $\tilde{x}$ będzie oszacowaniem wartości x (dokładnej)

- Błąd bezwzględny

$$\Delta x = |\tilde{x} - x| \quad (2)$$

- Błąd względny

$$\delta x = \left| \frac{\tilde{x} - x}{x} \right| \quad (3)$$

Liczby maszynowe

- zbiór *liczb maszynowych* A - zbiór liczb reprezentowalnych w danym formacie
- aproksymacja liczby x liczbą maszynową $rd(x)$:

$$\forall_{g \in A} |x - rd(x)| \leq |x - g| \quad (4)$$

Błędy zaokrągleń

- zaokrąglenie (o ile jest liczbą maszynową) jest dobrą aproksymacją – w systemie o k cyfrach znaczących (gdzie precyzja mantysy to k cyfr) mantysę $m = \alpha_1.\alpha_2\alpha_3 \dots$ zaokrąglamy do

$$\tilde{m} = \begin{cases} \alpha_1.\alpha_2 \dots \alpha_k & \text{jeżeli } \alpha_{k+1} < \frac{B}{2} \\ \alpha_1.\alpha_2 \dots \alpha_k + B^{-k+1} & \text{jeżeli } \alpha_{k+1} \geq \frac{B}{2} \end{cases} \quad (5)$$

a liczbę znormalizowaną $x = m * B^c$ do $\tilde{rd}(x) = \text{sgn}(x)\tilde{m} * B^c$.

Błąd względny przybliżenia:

$$\left| \frac{\tilde{rd}(x) - x}{x} \right| \leq \frac{\frac{B}{2} * B^{-k}}{|m|} \leq \frac{1}{2} * B^{-k+1} \quad (6)$$

- dokładność maszynowa $\text{eps} \stackrel{\text{def}}{=} \frac{1}{2} * B^{-k+1}$
 Zatem: $\tilde{rd}(x) = x(1 + \epsilon)$, gdzie $|\epsilon| \leq \text{eps}$
 Dla systemów binarnych $\text{eps} = 2^{-k}$

Błędy zaokrągleń – c.d.

- nadmiar i niedomiar cechy – nieprawidłowość – zatrzymanie obliczeń, wyjątek – na szczęście bardzo rzadko się zdarzają
- wyniki działań arytmetycznych nie muszą być liczbami maszynowymi – mamy więc *działania zmiennopozycyjne* aproksymujące właściwe działania, np:

$$x +^* y \stackrel{\text{def}}{=} rd(x + y)$$

$$x -^* y \stackrel{\text{def}}{=} rd(x - y)$$

$$x \times^* y \stackrel{\text{def}}{=} rd(x \times y)$$

$$x / ^* y \stackrel{\text{def}}{=} rd(x / y)$$

Wówczas:

$$x +^* y = (x + y)(1 + \epsilon_1)$$

$$x -^* y = (x - y)(1 + \epsilon_2)$$

$$x \times^* y = (x \times y)(1 + \epsilon_3)$$

$$x / ^* y = (x / y)(1 + \epsilon_4),$$

gdzie $|\epsilon_i| \leq eps$

Problemy arytmetyki zmiennopozycyjnej

- Jeżeli $|y| < \frac{eps}{B}|x|$, to $x +^* y = x$
- Działania zmiennopozycyjne nie muszą być łączne lub rozdzielne.

Przykład:

$$a = 2.3371258_{10} - 4, \quad b = 3.3678429_{10}2, \quad c = -3.3677811_{10}2$$

$$a + b + c = 6.41371258_{10} - 3$$

$$a +^* (b +^* c) = 2.3371258_{10} - 4 +^* 6.1800000_{10} - 3 = 6.4137126_{10} - 3$$

$$(a +^* b) +^* c = 3.3678452_{10}2 +^* -3.3677811_{10}2 = 6.4100000_{10} - 3$$

Oznaczenia:

- $fl(\text{wyrażenie})$ – wartość wyrażenia w arytmetyce zmiennopozycyjnej
- *działania elementarne* – działania arytmetyczne (podstawowe plus pewne dodatkowe takie jak pierwiastek, sinus itp.)

Pojęcie algorytmu

Zadanie: Mamy daną funkcję $\phi : D \rightarrow R^m$, gdzie $D \subseteq R^n$, $n, m \in N$. Szukamy metody wyznaczania wartości $y = \phi(x)$.

Odwzorowaniem elementarnym nazywamy odwzorowanie $\psi : D \subseteq R^k \rightarrow R^l$ takie, że jego funkcje składowe ψ_i są działaniami elementarnymi.

Algorytmem realizującym funkcję $\phi : D \rightarrow R^m$ ($D \subseteq R^n$, $n, m \in N$) nazywamy taki ciąg odwzorowań $\phi^{(1)}, \dots, \phi^{(r)}$, (gdzie $\phi^{(i)} : D_i \rightarrow D_{i+1}$, $D_i \subseteq R^{n_i}$, $D_1 = D$, $n_{r+1} = m$), że

$$\phi = \phi^{(1)} \circ \dots \circ \phi^{(r)} \quad (7)$$

Przykład:

Zadanie $\phi(a, b, c) = a + b + c$ może być realizowane algorytmem $\{\phi^{(1)}, \phi^{(2)}\}$, gdzie

$$\phi^{(1)}(a, b, c) = \begin{bmatrix} a + b \\ c \end{bmatrix}, \quad \phi^{(2)}(d, e) = d + e. \quad (8)$$

Propagacja błędów

Prześledźmy błędy zaokrągleń dla przykładu sumowania $(a + b) + c$:

$$\begin{aligned} fl((a + b) + c) &= fl(fl(a + b) + c) = \\ &= ((a + b)(1 + \epsilon_1) + c)(1 + \epsilon_2) = \\ &= (a + b + c) \left[1 + \frac{a + b}{a + b + c} \epsilon_1 (1 + \epsilon_2) + \epsilon_2 \right] \end{aligned}$$

$\frac{a+b}{a+b+c}$ jest współczynnikiem wzmocnienia błędu ϵ_1 – jego duża wartość powoduje duży błąd względny

Ponieważ we wcześniejszym przykładzie:

$$\frac{a + b}{a + b + c} \approx 5 * 10^4, \quad \frac{b + c}{a + b + c} \approx 0.97 \quad (9)$$

to wybrać należy metodę $fl(a + (b + c))$.

Zadanie: Analizujemy „naturalny” algorytm znajdujący przybliżenie b_n rozwiązania β_n równania liniowego

$$c - a_1 b_1 - \cdots - a_{n-1} b_{n-1} - a_n \beta_n = 0, \quad (10)$$

dla danych $c, a_1, \dots, a_n, b_1, \dots, b_{n-1}, a_n \neq 0$. Oszacować residuum

$$r = c - a_1 b_1 - \cdots - a_n b_n. \quad (11)$$

Algorytm: Wyznaczamy

$$b_n = fl \left(\frac{c - a_1 b_1 - \cdots - a_{n-1} b_{n-1}}{a_n} \right) \quad (12)$$

poprzez:

$$s_0 = c,$$

$$s_j = fl(s_{j-1} - a_j b_j) = [s_{j-1} - a_j b_j (1 + \mu_j)] (1 + \alpha_j), \quad j = 1, 2, \dots, n-1,$$

$$b_n = fl \left(\frac{s_{n-1}}{a_n} \right) = (1 + \delta) \frac{s_{n-1}}{a_n}. \quad (13)$$

Oszacowanie 1: Z (13) wynikają (drugie równanie dla $j = 1, 2, \dots, n-1$):

$$s_0 - c = 0,$$

$$s_j - (s_{j-1} - a_j b_j) = s_j - \left(\frac{s_j}{1 + \alpha_j} + a_j b_j \mu_j \right) = s_j \frac{\alpha_j}{1 + \alpha_j} - a_j b_j \mu_j, \quad (14)$$

$$a_n b_n - s_{n-1} = \delta s_{n-1}.$$

Dodając stronami dostajemy:

$$r = c - \sum_{i=1}^n a_i b_i = \sum_{j=1}^{n-1} \left(-s_j \frac{\alpha_j}{1 + \alpha_j} + a_j b_j \mu_j \right) - \delta s_{n-1}. \quad (15)$$

Stąd oszacowanie:

$$|r| \leq \frac{eps}{1 - eps} \left(\delta' |s_{n-1}| + \sum_{j=1}^{n-1} (|s_j| + |a_j b_j|) \right), \quad \delta' = \begin{cases} 0 & \text{jeżeli } a_n = 1 \\ 1 & \text{w przec. przyp.} \end{cases} \quad (16)$$

Oszacowanie 2: Z (13):

$$b_n = \left[c \prod_{k=1}^{n-1} (1 + \alpha_k) - \sum_{j=1}^{n-1} a_j b_j (1 + \mu_j) \prod_{k=j}^{n-1} (1 + \alpha_k) \right] \frac{1 + \delta}{a_n}. \quad (17)$$

Stąd

$$c = \sum_{j=1}^{n-1} a_j b_j (1 + \mu_j) \prod_{k=1}^{j-1} (1 + \alpha_k)^{-1} + a_n b_n (1 + \delta)^{-1} \prod_{k=1}^{n-1} (1 + \alpha_k)^{-1} \quad (18)$$

Łatwo pokazać indukcyjnie (ze względu na m), że z

$$1 + \sigma = \prod_{k=1}^m (1 + \sigma_k)^{\pm 1}, \quad |\sigma_k| \leq \text{eps}, \quad m \cdot \text{eps} < 1 \quad (19)$$

wynika

$$|\sigma| \leq \frac{m \cdot \text{eps}}{1 - m \cdot \text{eps}}. \quad (20)$$

Zatem istnieją wielkości ε_j takie, że

$$c = \sum_{j=1}^{n-1} a_j b_j (1 + j\varepsilon_j) + a_n b_n (1 + (n-1 + \delta')\varepsilon_n),$$

$$|\varepsilon_j| \leq \frac{eps}{1 - n \cdot eps}, \quad \delta' = \begin{cases} 0 & \text{jeżeli } a_n = 1 \\ 1 & \text{w przec. przyp.} \end{cases}$$
(21)

Ostatecznie dla $r = c - a_1 b_1 - \dots - a_n b_n$ dostajemy:

$$|r| \leq \frac{eps}{1 - n \cdot eps} \left[\sum_{j=1}^{n-1} j |a_j b_j| + (n-1 + \delta') |a_n b_n| \right].$$
(22)

Wskaźniki uwarunkowania

Niech $\phi : D \rightarrow R^m$ dla otwartego zbioru $D \subseteq R^n$, oraz $\phi = \begin{bmatrix} \phi_1 \\ \vdots \\ \phi_m \end{bmatrix}$ przy czym

ϕ_i mają ciągłe pochodne na D . Przy założeniu pomiarów niedokładnych $\tilde{x}$ wielkości x otrzymujemy $\tilde{y} = \phi(\tilde{x})$ jako przybliżenie $y = \phi(x)$. Rozwinięcie w szereg Taylora z pominięciem wyrazów wyższych rzędów daje:

$$\Delta y_i = \tilde{y}_i - y_i = \phi_i(\tilde{x}) - \phi_i(x) \doteq \sum_{j=1}^n (\tilde{x}_j - x_j) \frac{\delta \phi_i(x)}{\delta x_j} = \sum_{j=1}^n \frac{\delta \phi_i(x)}{\delta x_j} \Delta x_j \quad (23)$$

W związku z tym błąd względny:

$$\frac{\Delta y_i}{y_i} \doteq \sum_{j=1}^n \frac{x_j}{\phi_i(x)} \frac{\delta \phi_i(x)}{\delta x_j} \frac{\Delta x_j}{x_j} \quad (24)$$

Współczynniki $\frac{x_j}{\phi_i(x)} \frac{\delta \phi_i(x)}{\delta x_j}$ nazywamy *wskaźnikami uwarunkowania zadania*.

Zadania źle i dobrze uwarunkowane

Zadanie nazywamy *źle uwarunkowanym* jeżeli przynajmniej jeden ze wskaźników uwarunkowania ma dużą wartość bezwzględną. W przeciwnym razie zadanie nazywamy *dobrze uwarunkowanym*.

- tylko dla niezerowych x_j i y_i
- mn wskaźników
- inne definicje, np: c jest wskaźnikiem uwarunkowania zadania ϕ jeśli

$$\frac{\|\phi(\tilde{x}) - \phi(x)\|}{\|\phi(x)\|} \leq c \frac{\|\tilde{x} - x\|}{\|x\|} \quad (25)$$

Błędy nieuniknione i numeryczna stabilność

Błędem nieuniknionym związanym z $y = \phi(x)$ nazywamy optymalny poziom błędu wyniku y czyli błąd wynikający z zaokrągleń danych wejściowych (niezależny od algorytmu). Analiza błędów w kontekście algorytmów jako złożenia odwzorowań elementarnych prowadzi do oszacowania:

$$\Delta^{(0)}y = [|D_{\phi(x)}| \cdot |x| + |y|]eps. \quad (26)$$

Jeśli błędy zaokrągleń algorytmu są w przybliżeniu równe $\Delta^{(0)}y$, to mówimy, że są one *nieszkodliwe*, a algorytm nazywamy *stabilnym (poprawnym) numerycznie* lub mówimy, że *zachowuje się dobrze*.

Algorytm jest *numerycznie użyteczniejszy* od innego (wyznaczającego tę samą wielkość) dla danego zbioru, gdy całkowity błąd zaokrąglenia jest mniejszy w pierwszym niż w drugim.

Uwaga: Stabilność numeryczna i większa użyteczność numeryczna są własnościami lokalnymi (zależnymi od wartości wejścia).

Arytmetyka przedziałowa

- Cel: wyznaczanie górnych ograniczeń błędu bezwzględnego
- Uwzględnia błędy zaokrągleń i błędy danych
- Obliczenia na przedziałach możliwych wartości liczby zamiast na liczbach: dla $a \in R$ mamy przedział $\tilde{a} = [a', a'']$, gdzie $a', a'' \in A$.
- Działania na przedziałach: dla $\odot \in \{\oplus, \ominus, \otimes, \oslash\}$ wynik

$$\tilde{c} = \tilde{a} \odot \tilde{b} \supseteq \{x \cdot y \mid x \in \tilde{a}, y \in \tilde{b}\} \quad (27)$$

Przykłady:

$$[a', a''] \oplus [b', b''] = [\max\{x \in A \mid x \leq a' + b'\}, \min\{x \in A \mid x \geq a'' + b''\}]$$

$$[a', a''] \otimes [b', b''] = [\max\{x \in A \mid x \leq a' \times b'\}, \min\{x \in A \mid x \geq a'' \times b''\}]$$

Uwaga: Oszacowania przedziałowe są prawdziwe, ale dość pesymistyczne

Przykład: Schemat Hornera a wzory skróconego mnożenia dla

$$x^3 - 3x^2 + 3x = (x - 1)^3 + 1 \quad (28)$$

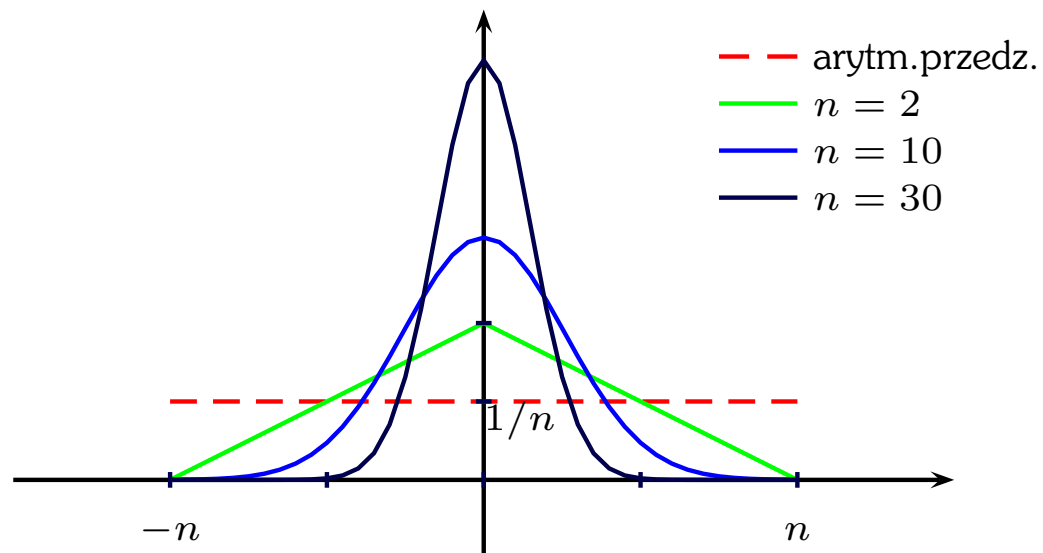
Arytmetyka przedziałowa a statystyka

Pesymizm arytmetyki przedziałowej widać już na przykładzie sumy n liczb.

Niech $n = 2$ i $\phi(x, y) = z = x + y$.

Założmy, że $\tilde{x} = \tilde{y} = [-1, 1]$.

Wówczas $\tilde{z} = [-2, 2]$. Ale rozkład prawdopodobieństwa $\tilde{z}$ nie jest jednostajny.



Dla $n = 30$, arytmetyka przedziałowa daje zakres około dwukrotnie szerszy niż przedział istotnie niezerowego prawdopodobieństwa.

Statystyczna teoria błędów

- *Błędem standardowym* wartości przybliżonej nazywamy odchylenie standardowe tej wartości traktowanej jako zmienna losowa.
- *Błędem prawdopodobnym* nazywamy taką liczbę x , że z prawdopodobieństwem $\frac{1}{2}$ mamy $|\epsilon| < x$.
 - Dla rozkładu normalnego mamy więc $x = 0.6745\sigma$, gdzie σ jest odchyleniem standardowym tego rozkładu.
 - W poprzednim przykładzie mamy $\sigma = \sqrt{\frac{n}{3}}$, a więc:

$$n = 1 \quad \Rightarrow \quad \sigma \approx 0.577$$

$$n = 10 \quad \Rightarrow \quad \sigma \approx 1.82 = 0.182n$$

$$n = 30 \quad \Rightarrow \quad \sigma \approx 3.16 \approx 0.1n$$

Twierdzenie 1 (o błędzie standardowym) *Jeżeli błędy $\Delta x_1, \dots, \Delta x_n$ są niezależnymi zmiennymi losowymi o wartości oczekiwanej równej zero i odchyleniach standardowych odpowiednio $\epsilon_1, \dots, \epsilon_n$, to błąd standardowy dla $y = y(x_1, \dots, x_n)$ dany jest wzorem:*

$$\epsilon \approx \sqrt{\sum_{i=1}^n \left(\frac{\delta y}{\delta x_i} \right)^2 \epsilon_i^2}. \quad (29)$$

Uwaga: Prosta suma wielu zmiennych jest przypadkiem szczególnym powyższego twierdzenia (pochodne cząstkowe są równe 1).

Zadanie interpolacji

Definicja 1 Niech $\Phi(x; a_0, \dots, a_n)$ będzie funkcją jednej zmiennej z $n + 1$ parametrami. Wartości parametrów określają funkcje rodziny Φ . Zadanie interpolacyjne dla Φ przy zadanych $m + 1$ parach liczb $(x_i, f_i), i = 0, \dots, m$, gdzie $x_i \neq x_j$ dla $i \neq j$, polega na wyznaczeniu parametrów a_i tak, aby

$$\Phi(x_i; a_0, \dots, a_n) = f_i, \quad i = 0, \dots, m. \quad (30)$$

Pary (x_i, f_i) nazywamy **punktami węzłowymi**, a wartości x_i i f_i odpowiednio **odciętą** i **rzędną** węzła.

Zadania interpolacji

- i. wielomianowa

$$\Phi(x; a_0, \dots, a_n) = a_0 + a_1x + \dots + a_nx^n \quad (31)$$

- i. trygonometryczna

$$\Phi(x; a_0, \dots, a_n) = a_0 + a_1e^{xi} + \dots + a_ne^{nxi} \quad (32)$$

- i. funkcjami sklejanymi
- i. wymierna

$$\Phi(x; a_0, \dots, a_k, b_0, \dots, b_l) = \frac{a_0 + a_1x + \dots + a_kx^k}{b_0 + b_1x + \dots + b_lx^l} \quad (33)$$

- i. eksponencjalna

$$\Phi(x; a_0, \dots, a_n; \lambda_0, \dots, \lambda_n) = a_0e^{\lambda_0x} + a_1e^{\lambda_1x} + \dots + a_ne^{\lambda_nx} \quad (34)$$

Wzór interpolacyjny Lagrange'a

Oznaczenie: Π_n – zbiór wszystkich wielomianów stopnia nie większego niż n (postaci $P(x) = a_0 + a_1x + \dots + a_nx^n$)

Twierdzenie 2 Dla $n+1$ dowolnych punktów węzłowych (x_i, f_i) , $i = 0, \dots, n$, takich, że $x_i \neq x_j$ dla $i \neq j$, istnieje tylko jeden wielomian $p \in \Pi_n$ spełniający

$$P(x_i) = f_i, \quad i = 0, \dots, n \quad (35)$$

Dowód:

Jednoznaczności: Niech $P_1, P_2 \in \Pi_n$ spełniają warunki (35). Wówczas $P = P_1 - P_2 \in \Pi_n$ jest wielomianem o $n + 1$ różnych zerach, a więc jest tożsamościowo równy zeru. □

Istnienia: Takim wielomianem jest:

$$P(x) = \sum_{i=0}^n f_i L_i(x) \quad (36)$$

gdzie L_i są tak zwanymi **wielomianami Lagrange'a** zdefiniowanymi jako:

$$L_i(x) = \prod_{\substack{j=0 \\ (j \neq i)}}^n \frac{x - x_j}{x_i - x_j}, \quad (37)$$

i spełniającymi następujące warunki (delty Kroneckera):

$$L_i(x_j) = \delta_{ij} = \begin{cases} 1, & \text{gdy } i = j \\ 0, & \text{gdy } i \neq j \end{cases}. \quad (38)$$

Uwagi:

- Nieduża przydatność w praktyce ($n^2 - 1$ mnożeń, 1 dzielenie).
- Metoda przydatna wówczas, gdy rozważamy wiele zadań interpolacji dla węzłów o tych samych odciętych.

Algorytm Neville'a

Idea: Rozwiązanie w krokach - od pojedynczych węzłów do całego ich zbioru.

Oznaczenie: Dla danego zbioru punktów węzłowych (x_i, f_i) , $i = 0, \dots, n$ przez $P_{i_0, \dots, i_k}$ oznaczamy wielomian ze zbioru Π_k taki, że

$$P_{i_0 \dots i_k}(x_{i_j}) = f_{i_j}, \quad j = 0, \dots, k. \quad (39)$$

Twierdzenie 3 Wielomiany typu (39) mają następujące własności:

$$1^\circ \quad P_i(x) = f_i$$

$$2^\circ \quad P_{i_0 \dots i_k}(x) = \frac{(x - x_{i_0})P_{i_1 \dots i_k}(x) - (x - x_{i_k})P_{i_0 \dots i_{k-1}}(x)}{x_{i_k} - x_{i_0}}$$

Dowód: Własność 1° jest oczywista.

Uwzględniając definicję (39) łatwo zauważyć, że prawa strona własności 2° jest dla $x = x_{i_0}$ równa f_{i_0} oraz dla $x = x_{i_k}$ równa f_{i_k} .

Dla pozostałych x_{i_j} (dla $j = 1, \dots, k-1$) otrzymujemy z prawej strony

$$\frac{(x_{i_j} - x_0)f_{i_j} - (x_{i_j} - x_{i_k})f_{i_k}}{x_{i_k} - x_{i_0}} = f_{i_j},$$

co ostatecznie dowodzi własności 2°.

□

Tabela ilustrująca algorytm Neville'a:

	$k = 0$	1	2	3
x_0	$f_0 = P_0(x)$			
		$P_{01}(x)$		
x_1	$f_1 = P_1(x)$		$P_{012}(x)$	
		$P_{12}(x)$		$P_{0123}(x)$
x_2	$f_2 = P_2(x)$		$P_{123}(x)$	
		$P_{23}(x)$		
x_3	$f_3 = P_3(x)$			

Uwagi:

- sprawnie wyznacza wartości funkcji dla pojedynczych punktów
- gorzej z wyznaczeniem samego wielomianu interpolacyjnego

Algorytm Neville'a – wersja 2

Oznaczenie:

$$T_{i+k} \stackrel{\text{def}}{=} P_{i \dots i+k} \quad (40)$$

Tablica ilustrująca algorytm:

x_0	$f_0 = T_{00}(x)$			
x_1	$f_1 = T_{10}(x)$	$T_{11}(x)$		
x_2	$f_2 = T_{20}(x)$	$T_{21}(x)$	$T_{22}(x)$	
x_3	$f_3 = T_{30}(x)$	$T_{31}(x)$	$T_{32}(x)$	$T_{33}(x)$

$$1^\circ \quad T_{i0}(x) = f_i$$

$$2^\circ \quad T_{ij}(x) = \frac{(x - x_{i-j})T_{i,j-1}(x) - (x - x_i)T_{i-1,j-1}(x)}{x_i - x_{i-j}} = T_{i,j-1} + \frac{T_{i,j-1} - T_{i-1,j-1}}{\frac{x - x_{i-j}}{x - x_i} - 1}$$

Wzór interpolacyjny Newtona

Wielomian interpolacyjny $P \in \Pi_n$, $P(x_i) = f_i$ dla $i = 0, \dots, n$ można przedstawić w postaci:

$$P_{0\dots n}(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)(x - x_1) + \dots + a_n(x - x_0) \cdots (x - x_{n-1}). \quad (41)$$

Wartości $P(x)$ można wyznaczać na wzór schematu Hornera ($n - 1$ mnożeń).

Współczynniki a_i można wyznaczyć wprost z:

$$f_0 = P(x_0) = a_0$$

$$f_1 = P(x_1) = a_0 + a_1(x_1 - x_0)$$

$$f_2 = P(x_2) = a_0 + a_1(x_2 - x_0) + a_2(x_2 - x_0)(x_2 - x_1)$$

...

wykonując n dzielení i $n(n - 1)$ mnożeń.

Ale można też za pomocą $n(n+1)/2$ dzielení. Zauważmy, że dla $P_{i_0 \dots i_k}$ oraz $P_{i_0 \dots i_{k-1}}$ istnieje jednoznacznie współczynnik $f_{i_0 \dots i_k}$ (**iloraz różnicowy k-tego rzędu**) taki, że:

$$P_{i_0 \dots i_k}(x) - P_{i_0 \dots i_{k-1}}(x) = f_{i_0 \dots i_k}(x - x_{i_0}) \cdots (x - x_{i_{k-1}}). \quad (42)$$

Mamy więc następującą postać Newtona częściowego wielomianu $P_{i_0 \dots i_k}(x)$:

$$f_{i_0} + f_{i_0 i_1}(x - x_0) + f_{i_0 i_1 i_2}(x - x_0)(x - x_1) + \cdots + f_{i_0 \dots i_k}(x - x_0) \cdots (x - x_{i_{k-1}}). \quad (43)$$

Twierdzenie 4 *Ilorazy różnicowe $f_{i_0 \dots i_k}$ są niezależne od permutacji indeksów.*

Dowód: $f_{i_0 \dots i_k}$ jest współczynnikiem przy najwyższej potędze wielomianu $P_{i_0 \dots i_k}$, który jest jednoznacznie wyznaczony przez węzły, które interpoluje. $\square$

Korzystając z własności 2° z twierdzenia 3 oraz z (42) mamy:

$$\begin{aligned}
 f_{i_0 \dots i_k} &\stackrel{(42)}{=} \frac{P_{i_0 \dots i_k}(x) - P_{i_0 \dots i_{k-1}}(x)}{(x - x_{i_0}) \cdots (x - x_{i_{k-1}})} \stackrel{2^\circ}{=} \\
 &\frac{\frac{(x - x_{i_0})P_{i_1 \dots i_k}(x) - (x - x_{i_k})P_{i_0 \dots i_{k-1}}(x)}{x_{i_k} - x_{i_0}} - P_{i_0 \dots i_{k-1}}(x)}{(x - x_{i_0}) \cdots (x - x_{i_{k-1}})} = \\
 &\frac{(x - x_{i_0})P_{i_1 \dots i_k}(x) - (x - x_{i_k})P_{i_0 \dots i_{k-1}}(x) - (x_{i_k} - x_{i_0})P_{i_0 \dots i_{k-1}}(x)}{(x_{i_k} - x_{i_0})(x - x_{i_0}) \cdots (x - x_{i_{k-1}})} = \\
 &\frac{\cancel{(x - x_{i_0})}(P_{i_1 \dots i_k}(x) - P_{i_0 \dots i_{k-1}}(x))}{(x_{i_k} - x_{i_0})\cancel{(x - x_{i_0})} \cdots (x - x_{i_{k-1}})} = \\
 &\frac{(P_{i_1 \dots i_k}(x) - P_{i_1 \dots i_{k-1}}(x) + P_{i_1 \dots i_{k-1}}(x) - P_{i_0 \dots i_{k-1}}(x))}{(x_{i_k} - x_{i_0})(x - x_{i_1}) \cdots (x - x_{i_{k-1}})} \stackrel{(42)}{=} \frac{f_{i_1 \dots i_k} - f_{i_0 \dots i_{k-1}}}{x_{i_k} - x_{i_0}}
 \end{aligned}$$

Schemat ilorazów różnicowych (tablica zestawiająca ilorazy różnicowe liczone wg wzoru z poprzedniej strony):

	$k = 0$	1	2	3
x_0	f_0			
		f_{01}		
x_1	f_1		f_{012}	
		f_{12}		f_{0123}
x_2	f_2		f_{123}	
		f_{23}		
x_3	f_3			

Wielomian interpolacyjny:

$$P(x) = f_0 + f_{01}(x - x_0) + f_{012}(x - x_0)(x - x_1) + \cdots + f_{0\dots n}(x - x_0) \cdots (x - x_{n-1}) \quad (44)$$

Uwagi:

- Dla zminimalizowania błędu przy obliczaniu wartości wielomianu interpolacyjnego Newtona można ustalić odpowiednią kolejność punktów węzłowych. Dla danego x dobieramy kolejność $i_0, \dots, i_n$ tak, aby $|x - x_{i_k}| \geq |x - x_{i_{k-1}}|$ dla $k = 1, \dots, n$.
- Jeżeli wszystkie odcięte x_i węzłów są uporządkowane tak, że $x_i < x_{i+1}$, to ze schematu ilorazów różnicowych można odczytać wszystkie reprezentacje Newtona. Dla danego x wybór indeksów następuje tak, by każdy następny był sąsiedni do któregoś z poprzednich.

Przykład: Zbiór węzłów: $\{(0, 1), (1, 2), (3, 10)\}$.

$x_0 = 0$	1	$P_{210}(x) = 10 + 4(x - 3) + 1(x - 3)(x - 1)$
	1	$P_{210}(x) = (1(x - 1) + 4)(x - 3) + 10$
$x_1 = 1$	2	$P_{210}(2) = 5$
	4	$P_{120}(x) = 2 + 4(x - 1) + 1(x - 1)(x - 3)$
$x_2 = 3$	10	$P_{120}(x) = (1(x - 3) + 4)(x - 1) + 2$

Reszta w interpolacji wielomianowej

Twierdzenie 5 *Jeżeli funkcja f jest $n + 1$ -krotnie różniczkowalna, to dla każdego argumentu $\bar{x}$ istnieje liczba ξ w najmniejszym przedziale $I[x_0, \dots, x_n, \bar{x}]$ zawierającym $\bar{x}$ i odcięte wszystkich węzłów, spełniająca*

$$f(\bar{x}) - P_{01\dots n}(\bar{x}) = \frac{\omega(\bar{x})f^{(n+1)}(\xi)}{(n+1)!}, \quad (45)$$

gdzie

$$\omega(x) = (x - x_0)(x - x_1) \cdots (x - x_n). \quad (46)$$

Uwagi:

- Interpolacja wielomianowa nie nadaje się do ekstrapolacji
- Brak zbieżności do wartości funkcji interpolowanej przy liczbie punktów węzłowych rosnącej do nieskończoności i malejącej do zera średnicy podziałów.

Interpolacja wymierna

Zadanie interpolacji węzłów (x_i, f_i) funkcją wymierną:

$$\Phi^{\mu,\nu}(x) = \frac{P^{\mu,\nu}(x)}{Q^{\mu,\nu}(x)} = \frac{a_0 + a_1x + \cdots + a_\mu x^\mu}{b_0 + b_1x + \cdots + b_\nu x^\nu}. \quad (47)$$

- Parę (μ, ν) nazywamy **stopniem zadania interpolacji wymiernej**.
- $\mu + \nu + 2$ parametry minus czynnik wspólny
- $\mu + \nu + 1$ węzłów $(x_i, f_i), i = 0, \dots, \mu + \nu$

Rozwiązanie zadania musi być rozwiązaniem układu równań $S^{\mu,\nu}$:

$$P^{\mu,\nu}(x_i) - f_i Q^{\mu,\nu}(x_i) = 0 \quad (48)$$

czyli

$$a_0 + a_1x_i + \cdots + a_\mu x_i^\mu - f_i(b_0 + b_1x_i + \cdots + b_\nu x_i^\nu) = 0 \quad (49)$$

Przykład: Zbiór punktów węzłowych $\{(0, 1), (1, 2), (2, 2)\}$ dla $\mu = \nu = 1$ prowadzi do układu równań:

$$\begin{aligned}a_0 - b_0 &= 0 \\a_0 + a_1 - 2(b_0 + b_1) &= 0 \\a_0 + 2a_1 - 2(b_0 + 2b_1) &= 0\end{aligned}$$

Rozwiązanie układu $a_0 = b_0 = 0, a_1 = 2, b_1 = 1$ nie spełnia warunków zadania – funkcja wymierna $\Phi^{1,1}(x) = \frac{2x}{x}$ nie jest określona w punkcie $x = 0$. Zatem rozwiązanie zadania nie istnieje.

Definicja 2 Wyrażenia wymierne:

$$\Phi_1(x) = \frac{P_1(x)}{Q_1(x)} \quad \Phi_2(x) = \frac{P_2(x)}{Q_2(x)} \quad Q_1 \not\equiv 0, \quad Q_2 \not\equiv 0$$

nazywamy **równoważnymi** gdy przedstawiają tę samą funkcję wymierną tj. wtedy, gdy

$$P_1(x)Q_2(x) = P_2(x)Q_1(x)$$

Wyrażenie wymierne nazywamy **relatywnie pierwszym** jeśli jego licznik i mianownik są relatywnie pierwsze, czyli nie są podzielne przez ten sam wielomian dodatniego stopnia.

Twierdzenie 6 Układ $S^{\mu,\nu}$ ma zawsze rozwiązanie nietrywialne. Dla każdego takiego rozwiązania $Q^{\mu,\nu} \not\equiv 0$.

Dowód: $S^{\mu,\nu}$ jest jednorodnym układem $\mu + \nu + 1$ równań z $\mu + \nu + 2$ niewiadomymi, więc ma rozwiązania nietrywialne $(a_0, \dots, a_\mu, b_0, \dots, b_\nu)$. Gdyby $Q^{\mu,\nu} \equiv 0$, czyli $\forall_{i=0}^\nu b_i = 0$, to mielibyśmy

$$P^{\mu,\nu}(x_i) = 0, \quad \text{dla } i = 0, \dots, \mu + \nu.$$

Ponieważ $P^{\mu,\nu}$ jest stopnia $\nu < \mu + \nu + 1$, więc mielibyśmy $P^{\mu,\nu} \equiv 0$, co jest sprzeczne z założeniem. □

Twierdzenie 7 (Jednoznaczność rozwiązania nietrywialnego) *Jeżeli Φ_1 oraz Φ_2 są dwoma nietrywialnymi rozwiązaniami układu $S^{\mu,\nu}$, to są one równoważne.*

Dowód: Jeśli $\Phi_1(x) = \frac{P_1(x)}{Q_1(x)}$ oraz $\Phi_2(x) = \frac{P_2(x)}{Q_2(x)}$, to

$$P(x) = P_1(x)Q_2(x) - P_2(x)Q_1(x) \quad (50)$$

ma $\mu + \nu + 1$ różnych zer (x_i , dla $i = 0, \dots, \mu + \nu + 1$). Ponieważ stopień P jest nie większy niż $\mu + \nu$, to $P \equiv 0$, czyli Φ_1 i Φ_2 są równoważne. $\square$

Wniosek 8 (z twierdzeń 6 i 7) *Dla każdego zadania interpolacji wymiernej $A^{\mu,\nu}$ istnieje jednoznaczna funkcja wymierna $\Phi^{\mu,\nu}$, która rozwiązuje odpowiadający mu układ równań $S^{\mu,\nu}$. Albo funkcja $\Phi^{\mu,\nu}$ jest rozwiązaniem zadania $A^{\mu,\nu}$, albo zadanie to nie jest rozwiązalne.*

Oznaczenia:

- **Punkty nieosiągalne** zadania $A^{\mu,\nu}$ to węzły, które nie są spełnione przez funkcję wymierną $\Phi^{\mu,\nu}$.
- $\tilde{\Phi}^{\mu,\nu} \stackrel{\text{ozn}}{=} \text{wyrażenie wymierne relatywnie pierwsze, równoważne } \Phi^{\mu,\nu}$.

Twierdzenie 9 Zadanie $A^{\mu,\nu}$ jest rozwiązalne, wtedy i tylko wtedy, gdy dla każdego rozwiązania $\Phi^{\mu,\nu}$ układu $S^{\mu,\nu}$, $\tilde{\Phi}^{\mu,\nu}$ rozwiązuje $S^{\mu,\nu}$.

Dowód:

$\Rightarrow$: Załóżmy, że $\tilde{\Phi}^{\mu,\nu}$ nie rozwiązuje $S^{\mu,\nu}$. Wówczas funkcja wymierna odpowiadająca wyrażeniu $\Phi^{\mu,\nu}$ nie rozwiązuje zadania $A^{\mu,\nu}$. Stąd i z wniosku 8, zadanie $A^{\mu,\nu}$ jest nierozwiązalne.

$\Leftarrow$: Niech $\tilde{\Phi}^{\mu,\nu}(x) = \frac{P(x)}{Q(x)}$. Rozważmy dwa przypadki:

1. $\forall_{i \in \{0, \dots, \mu+\nu\}} Q(x_i) \neq 0$. Wówczas $\tilde{\Phi}^{\mu,\nu}(x_i) = f_i$.
2. $\exists_{i \in \{0, \dots, \mu+\nu\}} Q(x_i) = 0$. Wówczas węzeł (x_i, f_i) byłby nieosiągalny, oraz mielibyśmy $P(x_i) = 0$. Zatem P i Q zawierałyby czynnik $x - x_i$ i nie byłyby relatywnie pierwsze – sprzeczność! □

Uwagi:

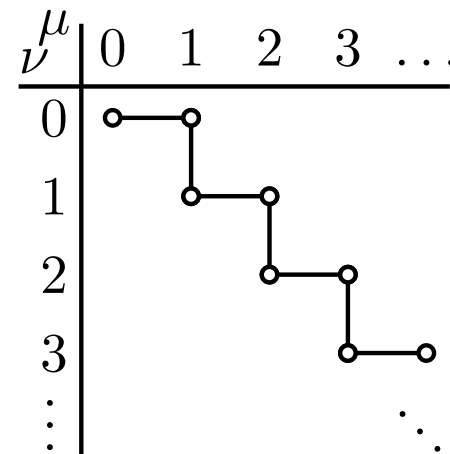
- Zadanie nierozwiązalne można zredukować do rozwiązalnego poprzez odpowiednie zmniejszenie liczby węzłów.
- Dla zadań niedegenerowalnych (tj. takich, że każdy podzbiór węzłów stanowi zadanie rozwiązalne) $A^{m,n}$ można rozpatrywać ciąg wyrażeń wymiernych $\Phi^{\mu,\nu}$ stopni (μ, ν) , gdzie $\mu \leq m, \nu \leq n$.
- W kolejnym kroku zwiększamy stopień licznika bądź mianownika.

Odwrotności ilorazów różnicowych

$$\begin{aligned}
 \phi(x_i, x_j) &\stackrel{\text{def}}{=} \frac{x_i - x_j}{f_i - f_j} \\
 \phi(x_i, x_j, x_k) &\stackrel{\text{def}}{=} \frac{x_j - x_k}{\phi(x_i, x_j) - \phi(x_i, x_k)} \\
 &\vdots \\
 \phi(x_i, \dots, x_l, x_m, x_n) &\stackrel{\text{def}}{=} \frac{x_m - x_n}{\phi(x_i, \dots, x_l, x_m) - \phi(x_i, \dots, x_l, x_n)}
 \end{aligned} \tag{51}$$

Uwaga: Nie są symetrycznymi funkcjami swoich argumentów.

Metoda wyznacza wyrażenia wymierne odpowiadające głównej przekątnej:



Założmy, że wyznaczamy funkcję wymierną

$$\Phi^{n,n} = \frac{P^n(x)}{Q^n(x)} \quad (52)$$

taką, że $\Phi^{n,n}(x_i) = f_i$ dla $i = 0, \dots, 2n$. Mamy:

$$\begin{aligned} \frac{P^n(x)}{Q^n(x)} &= f_0 + \frac{P^n(x)}{Q^n(x)} - \frac{P^n(x_0)}{Q^n(x_0)} = f_0 + \frac{P^n(x) - \frac{P^n(x_0)}{Q^n(x_0)}Q^n(x)}{Q^n(x)} = \\ &= f_0 + (x - x_0) \frac{P^{n-1}(x)}{Q^n(x)} = f_0 + \frac{x - x_0}{Q^n(x)/P^{n-1}(x)} \end{aligned} \quad (53)$$

Dla $i = 1, \dots, 2n$ zachodzi więc:

$$\frac{Q^n(x_i)}{P^{n-1}(x_i)} = \frac{x_i - x_0}{f_i - f_0} = \phi(x_0, x_i) \quad (54)$$

I dalej:

$$\begin{aligned} \frac{Q^n(x)}{P^{n-1}(x)} &= \phi(x_0, x_1) + \frac{Q^n(x)}{P^{n-1}(x)} - \frac{Q^n(x_1)}{P^{n-1}(x_1)} = \\ &= \phi(x_0, x_1) + \frac{x - x_1}{P^{n-1}(x)/Q^{n-1}(x)} \end{aligned} \quad (55)$$

Kontynuując, otrzymamy wyrażenie reprezentowane przez ułamek łańcuchowy:

$$\begin{aligned} \Phi^{n,n}(x) = f_0 + \frac{x - x_0}{\phi(x_0, x_1)} + \frac{x - x_1}{\phi(x_0, x_1, x_2)} + \cdots \\ \cdots + \frac{x - x_{2n-1}}{\phi(x_0, x_1, \dots, x_{2n})} \end{aligned} \quad (56)$$

Uwaga: Kolejne ułamki częściowe są wyrażeniami wymiernymi $\Phi^{\mu,\mu}(x)$ oraz $\Phi^{\mu+1,\mu}(x)$ dla $\mu = 0, 1, \dots, n-1$ spełniającymi odpowiednie podzbiory zbioru węzłów zadania:

$$\begin{aligned} \Phi^{0,0}(x) &= f_0 \\ \Phi^{1,0}(x) &= f_0 + \frac{x - x_0}{\phi(x_0, x_1)} \\ \Phi^{1,1}(x) &= f_0 + \frac{x - x_0}{\phi(x_0, x_1)} + \frac{x - x_1}{\phi(x_0, x_1, x_2)} \\ &\dots \end{aligned} \quad (57)$$

Algorytm: Realizacja wzoru 56 z pomocą tablicy odwrotności ilorazów:

0	x_0	f_0			
1	x_1	f_1	$\phi(x_0, x_1)$		
2	x_2	f_2	$\phi(x_0, x_2)$	$\phi(x_0, x_1, x_2)$	
3	x_3	f_3	$\phi(x_0, x_3)$	$\phi(x_0, x_1, x_3)$	$\phi(x_0, x_1, x_2, x_3)$
	$\dots$				

Odwrotne ilorazy różnicowe

$$\begin{aligned}
 \rho(x_i) &\stackrel{\text{def}}{=} f_i \\
 \rho(x_i, x_{i+1}) &\stackrel{\text{def}}{=} \frac{x_i - x_{i+1}}{f_i - f_{i+1}} \\
 &\vdots \\
 \rho(x_i, \dots, x_{i+k}) &\stackrel{\text{def}}{=} \frac{x_i - x_{i+k}}{\rho(x_i, \dots, x_{i+k-1}) - \rho(x_{i+1}, \dots, x_{i+k})} + \rho(x_{i+1}, \dots, x_{i+k-1})
 \end{aligned} \tag{58}$$

Uwaga: Są symetrycznymi funkcjami swoich argumentów.

Twierdzenie 10 Dla $p = 1, 2, \dots$, przyjmując $\rho(x_0, \dots, x_{p-2}) = 0$ dla $p = 1$ spełnione jest:

$$\phi(x_0, \dots, x_p) = \rho(x_0, \dots, x_p) - \rho(x_0, \dots, x_{p-2}) \tag{59}$$

Dowód: Indukcja. Dla $p = 1$ teza jest oczywista.

Zakładamy prawdziwość dla p . Wówczas:

$$\begin{aligned}\phi(x_0, \dots, x_{p+1}) &= \frac{x_p - x_{p+1}}{\phi(x_0, \dots, x_p) - \phi(x_0, \dots, x_{p-1}, x_{p+1})} \\ &= \frac{x_p - x_{p+1}}{\rho(x_0, \dots, x_p) - \rho(x_0, \dots, x_{p-1}, x_{p+1})}\end{aligned}\quad (60)$$

Z definicji ρ mamy:

$$\rho(x_{p+1}, x_0, \dots, x_p) - \rho(x_0, \dots, x_{p-1}) = \frac{x_{p+1} - x_p}{\rho(x_{p+1}, x_0, \dots, x_{p-1}) - \rho(x_0, \dots, x_p)} \quad (61)$$

Wobec symetryczności ρ mamy tezę. □

Ułamek łańcuchowy Thiele'a otrzymujemy zastępując odwrotności ilorazów różnicowych odwrotnymi ilorazami różnicowymi:

$$\begin{aligned}\Phi^{n,n}(x) &= f_0 + \frac{x - x_0}{\rho(x_0, x_1)} + \frac{x - x_1}{\rho(x_0, x_1, x_2) - \rho(x_0)} + \dots \\ &\quad \dots + \frac{x - x_{2n-1}}{\rho(x_0, \dots, x_{2n}) - \rho(x_0, \dots, x_{2n-2})}\end{aligned}\quad (62)$$

Uogólnienie algorytmu Neville'a

Oznaczenie:

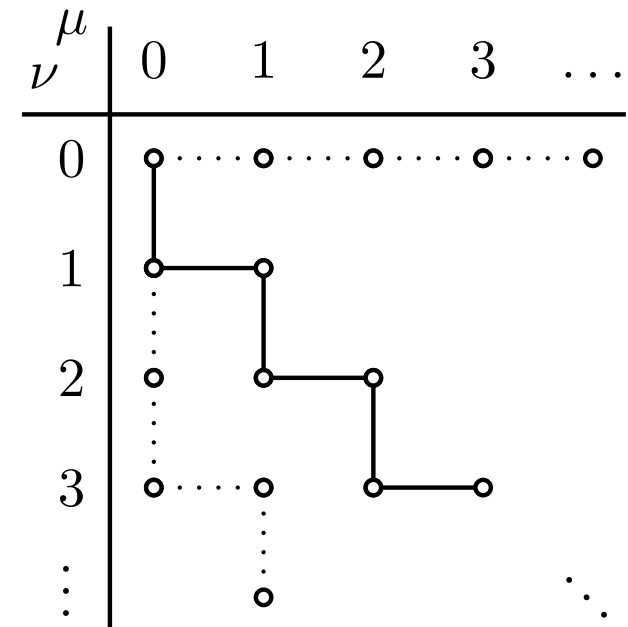
- $\Phi_s^{\mu,\nu}(x) \stackrel{\text{def}}{=} \frac{P_s^{\mu,\nu}(x)}{Q_s^{\mu,\nu}(x)}$ – wyrażenie wymierne spełniające warunki: $\Phi_s^{\mu,\nu}(x_i) = f_i$ dla $i = s, s+1, \dots, s+\mu+\nu$. $P_s^{\mu,\nu}(x)$ i $Q_s^{\mu,\nu}(x)$ są wielomianami stopni odpowiednio μ i ν .

Twierdzenie 11 Dla $\mu \geq 1$ oraz $\nu \geq 1$ zachodzą:

$$(a) \quad \Phi_s^{\mu,\nu}(x) = \Phi_{s+1}^{\mu-1,\nu}(x) + \frac{\Phi_{s+1}^{\mu-1,\nu}(x) - \Phi_s^{\mu-1,\nu}(x)}{\frac{x - x_s}{x - x_{s+\mu+\nu}} \left(1 - \frac{\Phi_{s+1}^{\mu-1,\nu}(x) - \Phi_s^{\mu-1,\nu}(x)}{\Phi_{s+1}^{\mu-1,\nu}(x) - \Phi_{s+1}^{\mu-1,\nu-1}(x)} \right) - 1}$$

$$(b) \quad \Phi_s^{\mu,\nu}(x) = \Phi_{s+1}^{\mu,\nu-1}(x) + \frac{\Phi_{s+1}^{\mu,\nu-1}(x) - \Phi_s^{\mu,\nu-1}(x)}{\frac{x - x_s}{x - x_{s+\mu+\nu}} \left(1 - \frac{\Phi_{s+1}^{\mu,\nu-1}(x) - \Phi_s^{\mu,\nu-1}(x)}{\Phi_{s+1}^{\mu,\nu-1}(x) - \Phi_{s+1}^{\mu-1,\nu-1}(x)} \right) - 1}$$

- Wzorów z twierdzenia **11** można użyć do wyznaczenia wartości wyrażeń wymiernych przy odpowiednio wzrastających stopniach.
- Różne ścieżki to różne metody i różne finalne funkcje wymierne.
- Dla $\nu = 0$ i wzrastającego μ mamy interpolację wielomianową Neville'a.
- $\mu = 0$ i wzrastające ν odpowiadają interpolacji wielomianowej dla węzłów $(x_i, 1/f_i)$.
- Doświadczenie pokazuje, że warto realizować ścieżkę oznaczoną linią ciągłą.
- W przypadku konkretnej ścieżki suma $\mu + \nu$ jednoznacznie wyznacza μ i ν , co pozwala uprościć oznaczenia.



Dla ścieżki oznaczonej ciągłą linią możemy wprowadzić:

$$T_{ik} \stackrel{\text{ozn}}{=} \Phi_s^{\mu, \nu}, \quad \text{gdzie } i = s + \mu + \nu, \quad k = \mu + \nu \quad (63)$$

Mamy wówczas następujący algorytm:

$$T_{i0} = f_i, \quad T_{i,-1} = 0$$

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\frac{x - x_{i-k}}{x - x_i} \left(1 - \frac{T_{i,k-1} - T_{i-1,k-1}}{T_{i,k-1} - T_{i-1,k-2}} \right) - 1} \quad (64)$$

Tabela ilustrująca algorytm:

$(\mu, \nu) =$	$(0, 0)$	$(0, 1)$	$(1, 1)$	$(1, 2)$
	$T_{00} = f_0$			
$T_{0,-1} = 0$		T_{11}		
	$T_{10} = f_1$			
$T_{1,-1} = 0$		T_{21}	T_{22}	T_{33}
	$T_{20} = f_2$		T_{32}	
$T_{2,-1} = 0$		T_{31}		
	$T_{30} = f_3$			

Interpolacja wielomianowa a wymierna – $\text{ctg}(2.5^\circ) = 22.9037655484$

x_1	$f_i = \text{ctg}(x_i)$				
1°	57.28996163				
		14.30939911			
2°	28.63625328		21.47137102		
		23.85869499		22.36661762	
3°	19.08113669		23.26186421		22.63519158
		21.47137190		23.08281486	
4°	14.30066626		22.18756808		
		18.60658719			
5°	11.43005230				
1°	57.28996163				
		22.90760673			
2°	28.63625328		22.90341624		
		22.90201805		22.90369573	
3°	19.08113669		22.90411487		22.90376552
		22.91041916		22.90384141	
4°	14.30066626		22.90201975		
		22.94418151			
5°	11.43005230				

Interpolacja funkcjami sklejanymi

Definicja 3 Niech $\Delta \stackrel{\text{def}}{=} \{x_0, x_1, \dots, x_n\}$, gdzie $a = x_0 < x_1 < \dots < x_n = b$ będzie podziałem przedziału $[a, b]$. **Funkcją sklejaną trzeciego stopnia** na Δ nazywamy każdą funkcję $S_\Delta : [a, b] \rightarrow R$ taką, że:

- $S_\Delta \in C^2[a, b]$ (dwukrotnie ciągle różniczkowalna na $[a, b]$)
- S_Δ na każdym podprzedziale $[x_i, x_{i+1}]$, $i = 0, \dots, n - 1$ pokrywa się z wielomianem trzeciego stopnia.

Oznaczenie: Dla zbioru liczb rzeczywistych $Y = \{y_0, \dots, y_n\}$ przez

$$S_{\Delta}(Y, \cdot) \tag{65}$$

rozumieć będziemy funkcję sklejaną S_{Δ} spełniającą warunki $S_{\Delta}(Y, x_i) = y_i$ dla $i = 0, \dots, n$.

Uwaga: S_{Δ} nie jest wyznaczona jednoznacznie (n funkcji stopnia 3 to $4n$ parametrów. $2n$ równań dla punktów węzłowych + $2(n - 1)$ równości pochodnych = $4n - 2$ równania). Pozostałe dwa stopnie swobody można wyeliminować np. jednym z następujących trzech warunków:

- (a) $S''_{\Delta}(Y, a) = S''_{\Delta}(Y, b) = 0$.
 - (b) $S_{\Delta}^{(k)}(Y, a) = S_{\Delta}^{(k)}(Y, b)$ dla $k = 0, 1, 2$. $S_{\Delta}(Y, \cdot)$ jest okresowa.
 - (c) $S'_{\Delta}(Y, a) = y'_0$, $S'_{\Delta}(Y, b) = y'_n$ dla danych $y'_0, y'_n \in R$.
- (66)

Interpolacja funkcjami sklejanymi

Niech $\Delta = \{x_0, x_1, \dots, x_n\}$ będzie ustalonym podziałem przedziału $[a, b]$ oraz $Y = \{y_0, y_1, \dots, y_n\} \subset R$

Oznaczenie:

$$h_{j+1} \stackrel{\text{ozn}}{=} x_{j+1} - x_j, \quad j = 0, 1, \dots, n-1 \quad (67)$$

Definicja 4 ***Momentami** funkcji $S_\Delta(Y, \cdot)$ nazywamy wartości jej drugich pochodnych w węzłach:*

$$M_j \stackrel{\text{def}}{=} S''_\Delta(Y, x_j), \quad j = 0, 1, \dots, n \quad (68)$$

Uwagi:

- Druga pochodna funkcji sklejaney to funkcja kawałkami liniowa (liniowa na przedziałach $[x_i, x_{i+1}]$).
- Funkcje sklepane są jednoznacznie wyznaczone przez swoje momenty.
- Momenty mogą być wyznaczone jako rozwiązanie układu równań liniowych.

Druga pochodna a momenty:

$$S''_{\Delta}(Y, x) = M_j \frac{x_{j+1} - x}{h_{j+1}} + M_{j+1} \frac{x - x_j}{h_{j+1}} \quad \text{dla } x \in [x_j, x_{j+1}] \quad (69)$$

Całkując otrzymujemy

$$S'_{\Delta}(Y, x) = -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + A_j \quad (70)$$

$$S_{\Delta}(Y, x) = M_j \frac{(x_{j+1} - x)^3}{6h_{j+1}} + M_{j+1} \frac{(x - x_j)^3}{6h_{j+1}} + A_j(x - x_j) + B_j \quad (71)$$

dla $x \in [x_j, x_{j+1}]$, $j = 0, 1, \dots, n-1$, gdzie A_j oraz B_j oznaczają stałe całkowania.

Ze spełniania odpowiednich punktów węzłowych otrzymujemy:

$$M_j \frac{h_{j+1}^2}{6} + B_j = y_j \quad (72)$$

$$M_{j+1} \frac{h_{j+1}^2}{6} + A_j h_{j+1} + B_j = y_{j+1} \quad (73)$$

Stąd:

$$B_j = y_j - M_j \frac{h_{j+1}^2}{6} \quad (74)$$

$$A_j = \frac{y_{j+1} - M_{j+1} \frac{h_{j+1}^2}{6} - B_j}{h_{j+1}} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6} (M_{j+1} - M_j) \quad (75)$$

Mamy więc dla $x \in [x_j, x_{j+1}]$:

$$S_\Delta(Y, x) = \alpha_j + \beta_j(x - x_j) + \gamma_j(x - x_j)^2 + \delta_j(x - x_j)^3 \quad (76)$$

gdzie:

$$\alpha_j = S_\Delta(Y, x_j) = y_j, \quad \gamma_j = \frac{S''_\Delta(Y, x_j)}{2} = \frac{M_j}{2} \quad (77)$$

$$\beta_j = S'_\Delta(Y, x_j) = -\frac{M_j h_{j+1}}{2} + A_j = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6} (M_{j+1} + 2M_j) \quad (78)$$

$$\delta_j = \frac{S'''_\Delta(Y, x_j^+)}{6} = \frac{M_{j+1} - M_j}{6h_{j+1}} \quad (79)$$

Wyznaczanie momentów

Założenie ciągłości $S'_\Delta(Y, \cdot)$ w węzłach wyznacza $n - 1$ równań dla momentów. Podstawiając A_j z (75) do (70) otrzymujemy:

$$S'_\Delta(Y, x) = -M_j \frac{(x_{j+1} - x)^2}{2h_{j+1}} + M_{j+1} \frac{(x - x_j)^2}{2h_{j+1}} + \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{6} (M_{j+1} - M_j) \quad (80)$$

W szczególności dla $j = 1, 2, \dots, n - 1$ mamy:

$$S'_\Delta(Y, x_j^-) = \frac{y_j - y_{j-1}}{h_j} + \frac{h_j}{3} M_j + \frac{h_j}{6} M_{j-1} \quad (81)$$

$$S'_\Delta(Y, x_j^+) = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{h_{j+1}}{3} M_j - \frac{h_{j+1}}{6} M_{j+1} \quad (82)$$

A z równości granic:

$$\frac{h_j}{6} M_{j-1} + \frac{h_j + h_{j+1}}{3} M_j + \frac{h_{j+1}}{6} M_{j+1} = \frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \quad (83)$$

czyli mamy $n - 1$ równań dla $n + 1$ momentów.

Wyznaczanie momentów - c.d.

Pozostałe 2 równania z dodatkowych warunków (66).

Przypadek (66.a):

$$S''_{\Delta}(Y, a) = M_0 = 0 \quad S''_{\Delta}(Y, b) = M_n = 0 \quad (84)$$

Przypadek (66.b): funkcja okresowa, czyli $h_{n+1} = h_1, M_{n+1} = M_1, y_{n+1} = y_1$:

$$S''_{\Delta}(Y, a) = S''_{\Delta}(Y, b) \equiv M_0 = M_n \quad (85)$$

$$S'_{\Delta}(Y, a) = S'_{\Delta}(Y, b) \equiv \frac{h_n}{6} M_{n-1} + \frac{h_n + h_1}{3} M_n + \frac{h_1}{6} M_1 = \frac{y_1 - y_n}{h_1} - \frac{y_n - y_{n-1}}{h_n} \quad (86)$$

Przypadek (66.c):

$$S'_{\Delta}(Y, a) = y'_0 \equiv \frac{h_1}{3} M_0 + \frac{h_1}{6} M_1 = \frac{y_1 - y_0}{h_1} - y'_0 \quad (87)$$

$$S'_{\Delta}(Y, b) = y'_n \equiv \frac{h_n}{6}M_{n-1} + \frac{h_n}{3}M_n = y'_n - \frac{y_n - y_{n-1}}{h_n} \quad (88)$$

Ogólna postać dla równań (83), (86), (87), (88):

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, 2, \dots, n-1, \quad (89)$$

gdzie:

$$\begin{aligned} \lambda_j &= \frac{h_{j+1}}{h_j + h_{j+1}}, & \mu_j &= 1 - \lambda_j = \frac{h_j}{h_j + h_{j+1}}, \\ d_j &= \frac{6}{h_j + h_{j+1}} \left(\frac{y_{j+1} - y_j}{h_{j+1}} - \frac{y_j - y_{j-1}}{h_j} \right) \end{aligned} \quad (90)$$

Dodatkowo:

Przypadek (66.a):

$$\lambda_0 = 0, \quad d_0 = 0, \quad \mu_n = 0, \quad d_n = 0 \quad (91)$$

Przypadek (66.c):

$$\begin{aligned}\lambda_0 &= 1, & d_0 &= \frac{6}{h_1} \left(\frac{y_1 - y_0}{h_1} - y'_0 \right) \\ \mu_n &= 1, & d_n &= \frac{6}{h_n} \left(y'_n - \frac{y_n - y_{n-1}}{h_n} \right)\end{aligned}\tag{92}$$

Zatem dla przypadków (66.a) i (66.c) otrzymujemy układ równań:

$$\begin{bmatrix} 2 & \lambda_0 & & & 0 \\ \mu_1 & 2 & \lambda_1 & & \\ & & \ddots & & \\ & & \mu_{n-1} & 2 & \lambda_{n-1} \\ & & & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_0 \\ M_1 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}\tag{93}$$

Przypadek (66.b):

$$\begin{aligned}\lambda_n &= \frac{h_1}{h_n + h_1}, \quad \mu_n = 1 - \lambda_n = \frac{h_n}{h_n + h_1} \\ d_n &= \frac{6}{h_n + h_1} \left(\frac{y_1 - y_n}{h_1} - \frac{y_n - y_{n-1}}{h_n} \right)\end{aligned}\tag{94}$$

Otrzymujemy następujący układ równań:

$$\begin{bmatrix} 2 & \lambda_1 & & & \mu_1 \\ \mu_2 & 2 & \lambda_2 & & \\ & & \ddots & & \\ & & & \mu_{n-1} & 2 & \lambda_{n-1} \\ \lambda_n & & & \mu_n & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{bmatrix}\tag{95}$$

oraz $M_0 = M_n$

Interpolacja funkcjami sklejanymi - podsumowanie

Twierdzenie 12 Układy równań liniowych 93 i 95 są nieosobliwe dla dowolnego podziału Δ przedziału $[a, b]$.

Schemat dowodu: Najpierw dowodzimy własności:

$$Az = w \Rightarrow \max_i |z_i| \leq \max_i |w_i|.$$

Następnie pokazujemy, że osobliwość macierzy przeczy tej własności. □

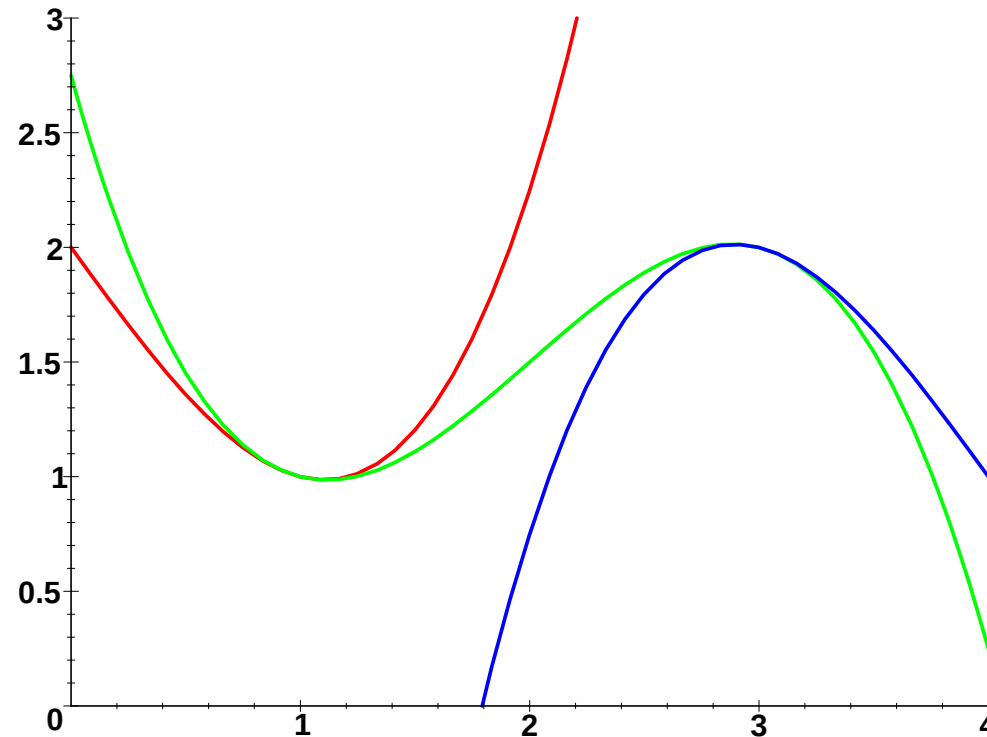
Uwagi:

- Układy te rozwiązujemy metodą eliminacji Gaussa (uproszczoną, bo macierze trójdzielne).
- Bardzo ważna własność! W przeciwieństwie do wielomianów, interpolowane funkcje sklepane zbiegają (przy pewnych słabych założeniach) do funkcji interpolowanej.

Interpolacja funkcjami sklejanymi - przykład

Interpolujemy funkcjami sklejanymi węzły: $(0, 2)$, $(1, 1)$, $(3, 2)$, $(4, 1)$ przyjmując wariant 66.a) tzn. $M_0 = M_3 = 0$. Rozwiązanie:

$$M = \begin{bmatrix} 0 \\ \frac{9}{4} \\ -\frac{9}{4} \\ 0 \end{bmatrix}$$



$$f(x) = \begin{cases} 2 - \frac{11}{8}x + \frac{3}{8}x^3 & \text{jeżeli } x \in [0, 1] \\ 1 - \frac{1}{4}(x-1) + \frac{9}{8}(x-1)^2 - \frac{3}{8}(x-1)^3 & \text{jeżeli } x \in [1, 3] \\ 2 - \frac{1}{4}(x-3) - \frac{9}{8}(x-3)^2 + \frac{3}{8}(x-3)^3 & \text{jeżeli } x \in [3, 4] \end{cases}$$

Metody całkowania

Definicja

$$\int_a^b f(x)dx = F(b) - F(a), \quad \text{gdzie } F'(x) = f(x) \quad (96)$$

Całkowanie symboliczne i liczenie wartości wprost z definicji często jest nierealne

- trudność wyznaczenia funkcji pierwotnej
- brak możliwości wyznaczenia funkcji pierwotnej (np. nie znamy do końca samej funkcji $f(x)$)
- wyznaczanie całek za pomocą dyskretyzacji

Kwadratury Newtona-Cotesa

- Zastępujemy funkcję podcałkową wielomianem interpolacyjnym.
- Przyjmijmy równomierny podział przedziału całkowania $[a.b]$:

$$x_i = a + ih, \quad i = 0, \dots, n, \quad \text{gdzie } h = \frac{b-a}{n}, n > 0, n \in \mathbb{N} \quad (97)$$

Niech P_n będzie wielomianem interpolacyjnym stopnia $\leq n$ takim, że:

$$P_n(x_i) = f_i = f(x_i) \quad \text{dla } i = 0, \dots, n \quad (98)$$

Ze wzoru interpolacyjnego Lagrange'a mamy:

$$P_n(x) = \sum_{i=0}^n f_i L_i(x), \quad \text{gdzie } L_i(x) = \prod_{\substack{k=0 \\ k \neq i}}^n \frac{x - x_k}{x_i - x_k} \quad (99)$$

- Wprowadzamy zmienną t taką, że $x = a + ht$. Mamy więc:

$$L_i(x) = \varphi_i(t) = \prod_{\substack{k=0 \\ k \neq i}}^n \frac{t - k}{i - k} \quad (100)$$

- Całkujemy:

$$\int_a^b P_n(x) dx = \sum_{i=0}^n f_i \int_a^b L_i(x) dx = h \sum_{i=0}^n f_i \int_0^n \varphi_i(t) dt = h \sum_{i=0}^n f_i \alpha_i \quad (101)$$

gdzie wagi

$$\alpha_i = \int_0^n \varphi_i(t) dt \quad (102)$$

nie zależą od całkowanej funkcji ani od przedziału całkowania.

- Dla $n = 2$ otrzymujemy tzw. *kwadraturę Simpsona*:

$$\begin{aligned}
 \alpha_0 &= \int_0^2 \frac{(t-1)(t-2)}{(0-1)(0-2)} dt = \frac{1}{2} \int_0^2 (t^2 - 3t + 2) dt = \\
 &= \frac{1}{2} \left[\frac{1}{3}t^3 - \frac{3}{2}t^2 + 2t \right]_0^2 = \frac{1}{2} \left[\frac{8}{3} - 6 + 4 \right] = \frac{1}{3} \\
 \alpha_1 &= \int_0^2 \frac{(t-0)(t-2)}{(1-0)(1-2)} dt = - \int_0^2 (t^2 - 2t) dt = \frac{4}{3} \\
 \alpha_2 &= \int_0^2 \frac{(t-0)(t-1)}{(2-0)(2-1)} dt = \frac{1}{2} \int_0^2 (t^2 - t) dt = \frac{1}{3}
 \end{aligned} \tag{103}$$

Ostatecznie:

$$\int_a^b f(x) dx \approx \int_a^b P_n(x) dx = \frac{b-a}{2} \frac{1}{3} (f_0 + 4f_1 + f_2) \tag{104}$$

- w ogólnej postaci: *kwadratury Newtona-Cotesa*

$$\int_a^b P_n(x)dx = h \sum_{i=0}^n f_i \alpha_i, \quad f_i = f(a + ih), \quad h = \frac{b-a}{n} \quad (105)$$

– Zachodzi: $\sum_{i=0}^n \alpha_i = n$

- Niech s będzie wspólnym mianownikiem wag α_i . Wówczas $\sigma_i \stackrel{\text{ozn}}{=} s\alpha_i$ są całkowite oraz:

$$\int_a^b P_n(x)dx = h \sum_{i=0}^n f_i \alpha_i = \frac{b-a}{ns} \sum_{i=0}^n \sigma_i f_i \quad (106)$$

- *reszta kwadratury* tj. błąd aproksymacji:

$$\int_a^b P_n(x)dx - \int_a^b f(x)dx = h^{p+1} K f^{(p)}(\xi), \quad \text{gdzie } \xi \in (a, b) \quad (107)$$

Zestawienie parametrów kwadratur Newtona-Cotesa:

n	σ_i							ns	Reszta	Kwadratura
1	1	1						2	$h^3 \frac{1}{12} f^{(2)}(\xi)$	Trapezów
2	1	4	1					6	$h^5 \frac{1}{90} f^{(4)}(\xi)$	Simpsona
3	1	3	3	1				8	$h^5 \frac{3}{80} f^{(4)}(\xi)$	3/8
4	7	32	12	32	7			90	$h^7 \frac{8}{945} f^{(6)}(\xi)$	Milne'a
5	19	75	50	50	75	19		288	$h^7 \frac{275}{12096} f^{(6)}(\xi)$	—
6	41	216	27	272	27	216	41	840	$h^9 \frac{9}{1400} f^{(8)}(\xi)$	Weddle'a

Dla $n > 6$ pojawiają się ujemne współczynniki co stanowi niebezpieczeństwo dużych błędów numerycznych.

Uwagi:

- rząd kwadratury $\stackrel{\text{def}}{=} r$ gdy całkowanie wielomianów stopnia $\leq r - 1$ jest bezbłędne
- kwadratura trapezów jest rzędu 2, Simpsona rzędu 4

Kwadratury złożone

- Całkujemy w $[x_i, x_{i+1}]$ dla $i = 0, \dots, n-1$, gdzie $x_i = a + ih$, $h = \frac{b-a}{n}$
- Złożona kwadratura trapezów

$$T(h) \stackrel{\text{def}}{=} \sum_{i=0}^{n-1} \frac{1}{2} h [f(x_i) + f(x_{i+1})] = h \left[\frac{f(a) + f(b)}{2} + \sum_{i=1}^{n-2} f(a + ih) \right] \quad (108)$$

Dla $f \in C^2[a, b]$ reszta kwadratury w przedziale $[x_i, x_{i+1}]$ wynosi

$$\frac{1}{12} h^3 f^{(2)}(\xi_i), \quad \text{gdzie } \xi_i \in [x_i, x_{i+1}] \quad (109)$$

Sumując reszty otrzymujemy:

$$\begin{aligned} T(h) - \int_a^b f(x) dx &= \frac{h^3}{12} \sum_{i=0}^{n-1} f^{(2)}(\xi_i) = \\ &= \frac{b-a}{12} h^2 \frac{1}{n} \sum_{i=0}^{n-1} f^{(2)}(\xi_i) = \frac{b-a}{12} h^2 f^{(2)}(\xi), \quad \text{gdzie } \xi \in (a, b) \end{aligned} \quad (110)$$

- Złożona kwadratura Simpsona (n parzyste)

$$S(h) \stackrel{\text{def}}{=} \frac{1}{3}h[f(a) + 4f(a+h) + 2f(a+2h) + 4f(a+3h) + \dots \\ \dots + 2f(b-2h) + 4f(b-h) + f(b)] \quad (111)$$

Reszta kwadratury:

$$S(h) - \int_a^b f(x)dx = \frac{h^5}{90} \sum_{i=0}^{n/2-1} f^{(4)}(\xi_i) = \\ = \frac{h^4}{90} \frac{b-a}{2} \frac{2}{n} \sum_{i=0}^{n/2-1} f^{(4)}(\xi_i) = \frac{b-a}{180} h^4 f^{(4)}(\xi), \quad \text{gdzie } \xi \in (a, b) \quad (112)$$

Uwagi:

- rząd kwadratury złożonej pozostaje taki jak składanej kwadratury
- reszty zbiegają do zera z potęgą h (mimo sumowania się błędów składanych kwadratur)

Złożona kwadratura Simpsona w C++:

```
1 typedef double FunCalc(double);
2
3 double SimpsonZ(FunCalc *f, double a, double b, int n)
4 {
5     if (n%2 || n<4) throw "Niepoprawne n";
6     double h = (b-a)/n;
7     double parts[3];
8     parts[0] = parts[1] = 0;
9     parts[2] = f(a) + f(b);
10    double z = b-h;
11    for (int i=n, k=0; --i; z-=h, k=!k)
12        parts[k] += f(z);
13    return (parts[2]+2*parts[1]+4*parts[0])*h/3;
14 }
```

Kwadratury z interpolacji Hermite'a

Interpolacja Hermite'a

- zagadnienie interpolacji węzłów $(x_i, y_i^{(k)})$, $k = 0, \dots, n_i - 1$, $i = 0, \dots, m$ wielomianem P stopnia $n = \sum_{i=0}^m n_i - 1$ spełniającego $P^{(k)}(x_i) = y_i^{(k)}$.
- rozwiązanie jest jednoznaczne
- uogólnienie wielomianów Lagrange'a:

$$P(x) = \sum_{i=0}^m \sum_{k=0}^{n_i-1} y_i^{(k)} L_{ik}(x), \quad \text{gdzie } L_{ik}^{(\sigma)}(x_j) = \begin{cases} 1 & \text{gdy } i = j \wedge k = \sigma \\ 0 & \text{w przec. przyp.} \end{cases} \quad (113)$$

Kwadratura dla węzłów $f(0)$, $f(1)$, $f'(0)$, $f'(1)$

Z uogólnionego wzoru Lagrange'a dostajemy:

$$P(t) = f(0)[(t-1)^2 + 2t(t-1)^2] + f(1)[t^2 - 2t^2(t-1)] + f'(0)t(t-1)^2 + f'(1)t^2(t-1) \quad (114)$$

Po scałkowaniu:

$$\int_0^1 P(t)dt = \frac{1}{2}(f(0) + f(1)) + \frac{1}{12}(f'(0) - f'(1)) \quad (115)$$

Po przekształceniu zmiennych:

$$\int_a^b f(x)dx \approx M(h) = \frac{1}{2}h(f(a) + f(b)) + \frac{1}{12}h^2(f'(a) - f'(b)) \quad (116)$$

gdzie $h = b - a$.

Reszta kwadratury:

$$M(h) - \int_a^b f(x)dx = -\frac{1}{720}h^5 f^{(4)}(\xi), \quad \xi \in (a, b) \quad (117)$$

Wzór sumacyjny Eulera-Maclaurina

Twierdzenie 13 Niech $g \in C^{2m+2}[0, 1]$, $m \in N$. Wówczas:

$$\int_0^1 g(t) dt = \frac{1}{2}(g(0) + g(1)) + \sum_{l=1}^m \frac{B_{2l}}{(2l)!} (g^{(2l-1)}(0) - g^{(2l-1)}(1)) - \frac{B_{2m+2}}{(2m+2)!} g^{(2m+2)}(\xi), \quad \xi \in (0, 1) \quad (118)$$

gdzie B_k to liczby Bernoulliego ($B_2 = \frac{1}{6}$, $B_4 = -\frac{1}{30}$, $B_6 = \frac{1}{42}$, $B_8 = -\frac{1}{30}$, $\dots$).

Schemat dowodu

- Wielomiany i liczby Bernoulliego – definicja i własności.
- Stosowne rozpisanie całki $\int_0^1 g(t) dt$.

Wielomiany i liczby Bernoulliego

Definicja 5 *Wielomianami Bernoulliego* nazywamy rodzinę wielomianów B_i , $i = 0, 1, \dots$ spełniającą następujące warunki:

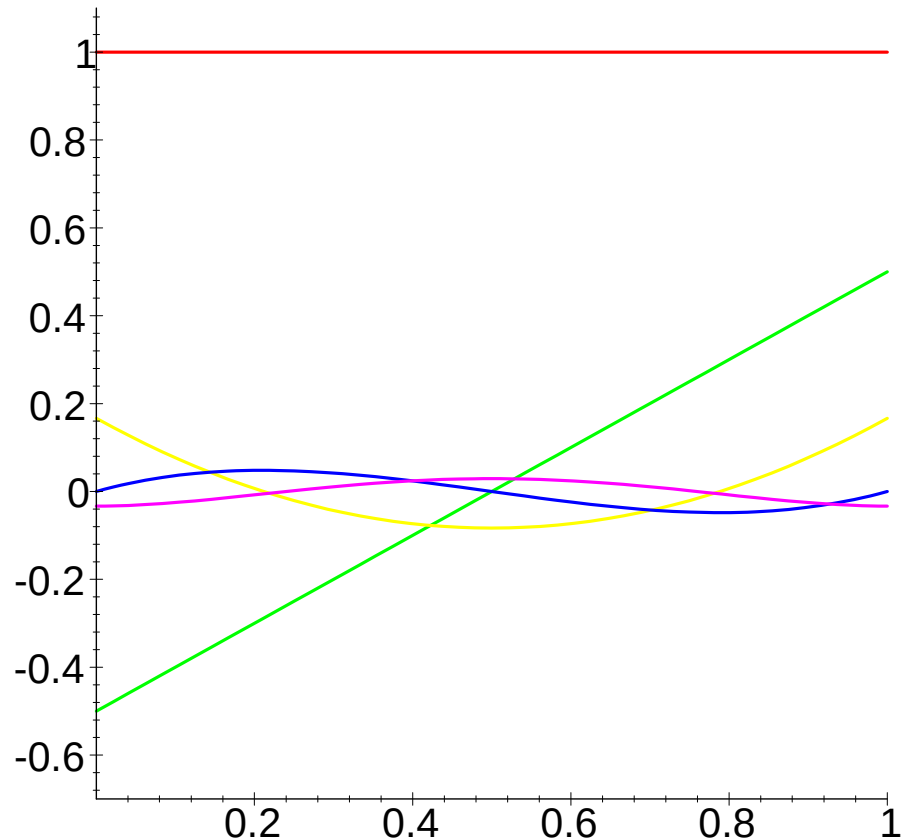
1. $B_1(x) = x - \frac{1}{2}$
2. $B'_{k+1}(x) = (k+1)B_k(x)$, $k = 0, 1, \dots$
3. $B_{2l+1}(0) = B_{2l+1}(1) = 0$, $l = 1, 2, \dots$

Przykłady:

$$\begin{aligned} B_0(x) &= 1, & B_1(x) &= x - \frac{1}{2}, & B_2(x) &= x^2 - x + \frac{1}{6}, \\ B_3(x) &= x^3 - \frac{3}{2}x^2 + \frac{1}{2}x, & B_4(x) &= x^4 - 2x^3 + x^2 - \frac{1}{30} \end{aligned} \quad (119)$$

Definicja 6 *Liczby Bernoulliego* to wartości $B_k = B_k(0)$, $k = 0, 1, \dots$

Wielomiany Bernoulliego



Własności:

- Wielomian B_k jest stopnia k .
- $(-1)^k B_k(1-x) = B_k(x)$
- $B_k(0) = B_k(1) = B_k$ dla $k \neq 1$

Wykorzystanie wielomianów Bernoulliego dla dowodu wzoru (118)

Całkowanie przez części:

$$\begin{aligned}
 \int_0^1 g(t) dt &= B_1(t)g(t)\Big|_0^1 - \int_0^1 B_1(t)g'(t) dt \\
 \int_0^1 B_1(t)g'(t) dt &= \frac{1}{2}B_2(t)g'(t)\Big|_0^1 - \frac{1}{2} \int_0^1 B_2(t)g''(t) dt \\
 &\vdots \\
 \int_0^1 B_{k-1}(t)g^{(k-1)}(t) dt &= \frac{1}{k}B_k(t)g^{(k-1)}(t)\Big|_0^1 - \frac{1}{k} \int_0^1 B_k(t)g^{(k)}(t) dt
 \end{aligned} \tag{120}$$

Wzór sumacyjny Eulera-Maclaurina

Rozszerzając wzór (118) na $n \in N$ przedziałów otrzymujemy dla $g \in C^{2m+2}[0, n]$:

$$\int_0^n g(t)dt = \frac{g(0) + g(n)}{2} + \sum_{i=1}^{n-1} g(i) + \sum_{l=1}^m \frac{B_{2l}}{(2l)!} (g^{(2l-1)}(0) - g^{(2l-1)}(n)) - \frac{B_{2m+2}}{(2m+2)!} n g^{(2m+2)}(\xi), \quad \xi \in (0, n) \quad (121)$$

W ogólnym przypadku przedziału $[a, b]$ i jego równomiernego podziału na n części ($h = \frac{b-a}{n}$):

$$\int_a^b f(x)dx = T(h) + \sum_{l=1}^m h^{2l} \frac{B_{2l}}{(2l)!} (f^{(2l-1)}(a) - f^{(2l-1)}(b)) - h^{2m+2} \frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\xi), \quad \xi \in (a, b), \quad (122)$$

gdzie $T(h)$ oznacza złożoną kwadraturę trapezów.

Kwadratury ekstrapolacyjne

Rozwinięcie (122) złożonego wzoru trapezów $T(h)$ dla funkcji f w zależności od długości kroku h :

$$T(h) = \tau_0 + \tau_1 h^2 + \cdots + \tau_m h^{2m} + \alpha_{m+1}(h) h^{2m+2}, \quad (123)$$

gdzie

$$\tau_0 = \int_a^b f(t) dt, \quad \tau_k = \frac{B_{2k}}{(2k)!} (f^{(2k-1)}(b) - f^{(2k-1)}(a)), \quad k = 1, \dots, m \quad (124)$$

oraz

$$\alpha_{m+1}(h) = \frac{B_{2m+2}}{(2m+2)!} (b-a) f^{(2m+2)}(\xi(h)), \quad \xi(h) \in (a, b) \quad (125)$$

Kwadratura Romberga

Jeśli f ma wszystkie pochodne w $[a, b]$, to w przejściu do granicy $m \rightarrow \infty$ otrzymujemy szereg potęgowy:

$$\tau_0 + \tau_1 h^2 + \tau_2 h^4 + \dots \quad (126)$$

Reszta rozwinięcia maleje z malejącym h . Rozwinięcie można traktować jak wielomian zmiennej h^2 przyjmujący wartość całki τ_0 w $h = 0$.

Dla ciągu długości kroków $\{h_i = \frac{b-a}{n_i} : i = 0, 1, \dots, m\}$, gdzie $n_0 = 1$ oraz $\{n_i : i = 0, 1, \dots, m\}$ jest ciągiem ściśle rosnącym, wyznaczamy kwadratury trapezów:

$$T_{i0} = T(h_i), \quad i = 0, 1, \dots, m, \quad (127)$$

które traktujemy jako węzły zadania interpolacji.

Jeśli $\tilde{T}_{mm}(h)$ jest wielomianem zmiennej h^2 interpolującym te węzły, to wartość ekstrapolowana $\tilde{T}_{mm}(0)$ jest przybliżeniem szukanej całki τ_0 .

Interpolacja Neville'a dla kwadratury Romberga

Cel: Wyznaczyć $\tilde{T}_{mm}(0)$

Algorytm: Dla indeksów i, k takich, że $1 \leq k \leq i \leq m$ definiujemy $\tilde{T}_{ik}(h)$ jako wielomian zmiennej h^2 stopnia co najwyżej k spełniający:

$$\tilde{T}_{ik}(h_j) = T(h_j) \quad \text{dla } j = i - k, \dots, i. \quad (128)$$

Wartość $\tilde{T}_{mm}(0)$ wyznaczamy algorytmem Neville'a:

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left[\frac{h_{i-k}}{h_i}\right]^2 - 1}. \quad (129)$$

Wówczas:

$$T_{ik} = \tilde{T}_{ik}(0) \quad (130)$$

Tablica algorytmu Neville'a dla kwadratury Romberga:

$$\begin{array}{c|l}
 h_0^2 & T(h_0) = T_{00} \\
 h_1^2 & T(h_1) = T_{10} \quad \begin{array}{c} T_{11} \end{array} \\
 h_2^2 & T(h_2) = T_{20} \quad \begin{array}{c} T_{21} \end{array} \quad \begin{array}{c} T_{22} \end{array} \\
 h_3^2 & T(h_3) = T_{30} \quad \begin{array}{c} T_{31} \end{array} \quad \begin{array}{c} T_{32} \end{array} \quad \begin{array}{c} T_{33} \end{array}
 \end{array}$$

Uwagi:

- Niektóre kwadratury dla $i = k$ to kwadratury Newtona-Cotesa:
 - $h_1 = \frac{h_0}{2} \Rightarrow T_{11}$ jest kwadraturą Simpsona
 - dodatkowo $h_2 = \frac{h_1}{2} \Rightarrow T_{22}$ jest kwadraturą Milne'a
- oryginalna metoda Romberga: $h_i = \frac{h_{i-1}}{2}$, $i = 1, 2, \dots$
- ciąg Bulirsch'a: $h_{2i-1} = \frac{h_{2i-2}}{2}$, $h_{2i} = \frac{h_{2i-2}}{3}$, $i = 1, 2, \dots$

Kwadratury Romberga – przykład

Całkujemy $f(x) = x^3$ w przedziale $[0, 2]$.

Przyjmujemy $h_1 = \frac{h_0}{2}$.

Wówczas:

$$\begin{array}{l|l} h_0 = 2 & T_{00} = 8 \\ h_1 = 1 & T_{10} = 5 \end{array} \quad T_{11} = 5 + \frac{5-8}{\left(\frac{2}{1}\right)^2 - 1} = 4$$

Interpolacja wymierna dla kwadratury Romberga

Dla interpolacji wielomianowej

- trudno określić dokładność,
- często warunek stopu to: $|T_{m-1,m-1} - T_{m,m-1}| \leq \varepsilon s$, gdzie s jest oszacowaniem wartości całki $\int_a^b |f(t)| dt$,

W interpolacji wymiernej mamy:

$$T_{i0} = T(h_i) \quad (131)$$

$$T_{ik} = T_{i,k-1} + \frac{T_{i,k-1} - T_{i-1,k-1}}{\left[\frac{h_{i-k}}{h_i}\right]^2 \left[1 - \frac{T_{i,k-1} - T_{i-1,k-1}}{T_{i,k-1} - T_{i-1,k-2}}\right] - 1}, \quad 1 \leq k \leq i \leq m \quad (132)$$

- można oszacować reszty

Kwadratury Gaussa

Zadanie: Wyznaczanie całek postaci

$$I(f) = \int_a^b \omega(x) f(x) dx, \quad (133)$$

gdzie $\omega(x)$ jest funkcją wagową na (skończonym lub nieskończonym) przedziale $[a, b]$, spełniającą następujące warunki:

- $\omega(x) \geq 0$
- ω jest funkcją mierzalną na przedziale $[a, b]$
- dla $k = 0, 1, \dots$ momenty $\mu_k = \int_a^b x^k \omega(x) dx$ istnieją i są skończone
- dla wielomianów $s(x)$ nieujemnych na $[a, b]$ z równości $\int_a^b \omega(x) s(x) dx = 0$ wynika, że $s \equiv 0$.

Rozpatrywane kwadratury:

$$\tilde{I}(f) = \sum_{i=1}^n w_i f(x_i). \quad (134)$$

- W przeciwieństwie do kwadratur Newtona-Cotesa tutaj węzły nie muszą być równoodległe.
- Jak dobrać odcięte x_i oraz w_i , żeby zmaksymalizować rząd kwadratury?

Podstawowe definicje:

Definicja 7 Zbiorem **unormowanych wielomianów rzeczywistych** stopnia j nazywamy

$$\overline{\Pi}_j \stackrel{\text{def}}{=} \{p : p(x) = x^j + a_1 x^{j-1} + \dots + a_j\} \quad (135)$$

Definicja 8 **Iloczynem skalarnym** funkcji f i g nazywamy

$$(f, g) \stackrel{\text{def}}{=} \int_a^b \omega(x) f(x) g(x) dx. \quad (136)$$

Definicja 9 Przez $L^2[a, b]$ oznaczać będziemy przestrzeń wszystkich funkcji określonych na przedziale $[a, b]$, dla których całka

$$(f, f) = \int_a^b \omega(x) (f(x))^2 dx \quad (137)$$

jest dobrze określona i skończona.

Twierdzenie 14 (Wielomiany ortogonalne) Istnieją wielomiany $p_j \in \overline{\Pi}_j, j = 0, 1, \dots$ spełniające

$$(p_i, p_k) = 0 \quad \text{dla } i \neq k. \quad (138)$$

Wielomiany te są wyznaczone jednoznacznie wzorami:

$$p_0(x) \equiv 1 \quad (139)$$

$$p_{i+1}(x) \equiv (x - \delta_{i+1})p_i(x) - \gamma_{i+1}^2 p_{i-1}(x), \quad (140)$$

gdzie $p_{-1}(x) = 0$ oraz

$$\delta_{i+1} = \frac{(xp_i, p_i)}{(p_i, p_i)} \quad \text{dla } i \geq 0 \quad (141)$$

$$\gamma_{i+1}^2 = \begin{cases} 0 & \text{dla } i = 0 \\ \frac{(p_i, p_i)}{(p_{i-1}, p_{i-1})} & \text{dla } i \geq 1 \end{cases} \quad (142)$$

Dowód: Ortogonalizacja Grama-Schmidta:

Z definicji $\overline{\Pi}_0$ mamy $p_0(x) \equiv 1$.

Założmy prawdziwość tezy dla wszystkich $j \leq i$. Dowolny wielomian $p_{i+1} \in \overline{\Pi}_{i+1}$

ma jednoznaczne przedstawienie

$$p_{i+1}(x) = (x - \delta_{i+1})p_i(x) + c_{i-1}p_{i-1}(x) + \cdots + c_0p_0(x). \quad (143)$$

Jeśli $p_{i+1} \in \overline{\Pi}_{i+1}$, to

- dla $j = i$ mamy:

$$(p_{i+1}, p_j) = (xp_i, p_i) - \delta_{i+1}(p_i, p_i) = 0 \quad (144)$$

- dla $j < i$ mamy:

$$(p_{i+1}, p_j) = (xp_i, p_j) + c_j(p_j, p_j) = 0 \quad (145)$$

Z założeń dla $\omega(x)$ przy podstawieniu $s = p_k^2$ wynika, że $(p_k, p_k) \neq 0$, więc δ_{i+1} i c_j są wyznaczone jednoznacznie.

Z założenia indukcyjnego dla $j < i$ mamy:

$$p_{j+1}(x) = (x - \delta_{j+1})p_j(x) - \gamma_{j+1}^2 p_{j-1}(x) \quad (146)$$

Stąd $(p_{j+1}, p_i) = (xp_j, p_i) = (xp_i, p_j)$, a zatem

$$c_j = -\frac{(p_{j+1}, p_i)}{(p_j, p_j)} = \begin{cases} -\gamma_{j+2}^2 & \text{dla } j = i-1 \\ 0 & \text{dla } j < i-1 \end{cases}, \quad (147)$$

czyli

$$p_{i+1}(x) = (x - \delta_{i+1})p_i(x) - \gamma_{i+1}^2 p_{i-1}(x). \quad (148)$$

Wniosek 15 Jeśli $p \in \Pi_{n-1}$, to $(p, p_n) = 0$.

Twierdzenie 16 Wielomian p_n ma n rzeczywistych i pojedynczych zer - wszystkie w przedziale $[a, b]$.

Twierdzenie 17 Ciąg wielomianów ortogonalnych jest układem Czebyszewa tzn. spełnia następujący warunek Haara:

Dla wzajemnie różnych argumentów $t_i, i = 1, \dots, n$, macierz

$$A = \begin{bmatrix} p_0(t_1) & \cdots & p_{n-1}(t_1) \\ \vdots & \vdots & \vdots \\ p_0(t_n) & \cdots & p_{n-1}(t_n) \end{bmatrix} \quad (149)$$

jest nieosobliwa.

Twierdzenie 18 Niech $x_1, \dots, x_n$ będą zerami wielomianu p_n oraz niech $w_1, \dots, w_n$ będą rozwiązaniem nieosobliwego układu równań

$$\sum_{i=1}^n p_k(x_i)w_i = \begin{cases} (p_0, p_0) & \text{dla } k = 0, \\ 0 & \text{dla } k = 1, \dots, n-1. \end{cases} \quad (150)$$

Wówczas $w_i > 0$ dla $i = 1, \dots, n$, oraz dla każdego wielomianu $p \in \Pi_{2n-1}$

$$\int_a^b \omega(x)p(x)dx = \sum_{i=1}^n w_i p(x_i). \quad (151)$$

Jeśli $w_i, x_i, i = 1, \dots, n$ spełniają (151) dla wszystkich $p \in \Pi_{2n-1}$, to x_i są zerami wielomianu p_n , a w_i spełniają (150).

Nie istnieją takie $x_i, w_i, i = 1, \dots, n$, które spełniają (151) dla wszystkich $p \in \Pi_{2n}$.

Twierdzenie 19 Zera wielomianu ortogonalnego p_n są wartościami własnymi macierzy trójdzielnej

$$J_n = \begin{bmatrix} \delta_1 & \gamma_2 & & 0 \\ \gamma_2 & \delta_2 & \gamma_3 & \\ \vdots & & \ddots & \vdots \\ 0 & & \gamma_n & \delta_n \end{bmatrix}. \quad (152)$$

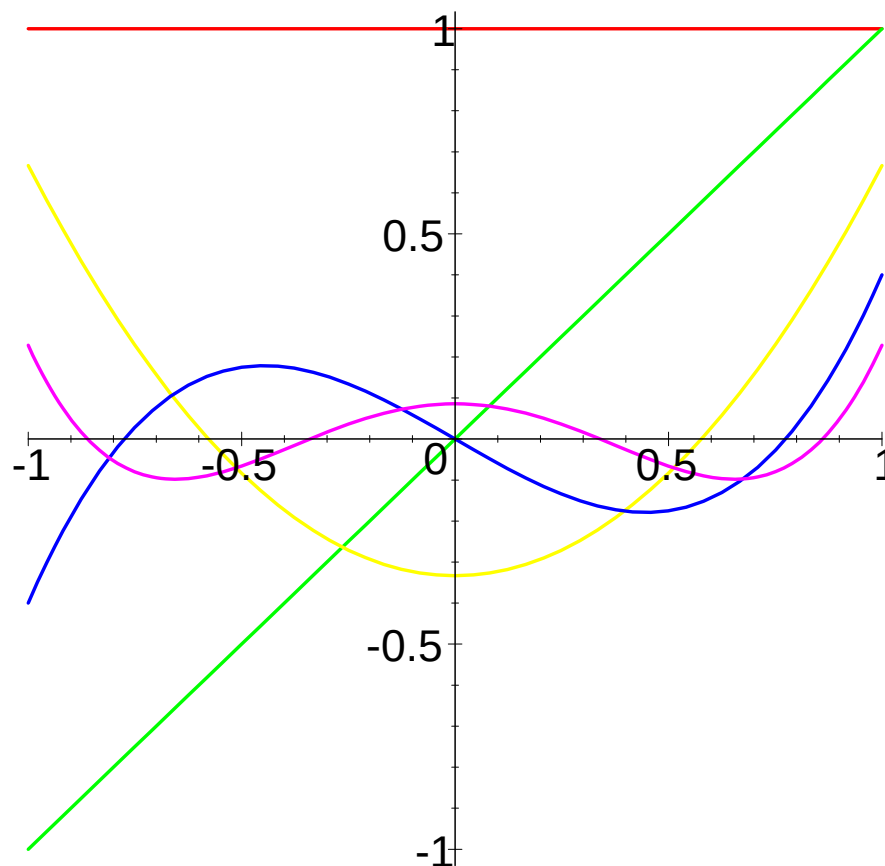
Twierdzenie 20 (Reszta kwadratury Gaussa) Jeżeli $f \in C^{2n}[a, b]$, to dla pewnego $\xi \in (a, b)$

$$\int_a^b \omega(x) f(x) dx - \sum_{i=1}^n w_i f(x_i) = \frac{f^{(2n)}(\xi)}{(2n)!} (p_n, p_n). \quad (153)$$

Najprostsza kwadratura Gaussa (Gaussa-Legendre'a):

Dla funkcji wagowej $\omega \equiv 1$ i przedziału $[a, b] = [-1, 1]$, wielomianami ortogonalnymi są:

$$p_k(x) = \frac{k!}{(2k)!} \frac{d^k}{dx^k} (x^2 - 1)^k, \quad k = 0, 1, \dots \quad (154)$$



Kwadratury Gaussa-Legendre'a:

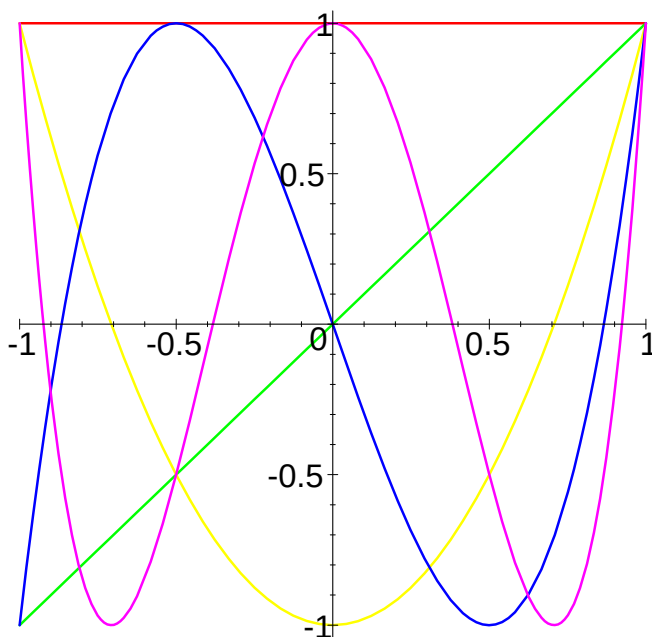
$$\int_{-1}^1 f(x)dx \approx \sum_{i=1}^n w_i f(x_i)$$

n	x_i	w_i
2	$\pm \frac{1}{\sqrt{3}} \approx 0.577350$	1
3	0 $\pm \sqrt{0.6} \approx \pm 0,774597$	$\frac{8}{9} \approx 0.888889$ $\frac{5}{9} \approx 0,555556$
4	$\pm \sqrt{(3 - 4\sqrt{0.3})/7} \approx 0.339981$ $\pm \sqrt{(3 + 4\sqrt{0.3})/7} \approx 0.861136$	$\frac{1}{2} \left(1 + \frac{1}{6} \sqrt{\frac{10}{3}}\right) \approx 0.652145$ $\frac{1}{2} \left(1 - \frac{1}{6} \sqrt{\frac{10}{3}}\right) \approx 0.347855$
5	0 $\pm \frac{1}{3} \sqrt{5 - 2\sqrt{\frac{10}{7}}} \approx 0.538469$ $\pm \frac{1}{3} \sqrt{5 + 2\sqrt{\frac{10}{7}}} \approx 0.906180$	$\frac{122}{255} \approx 0.568889$ $\frac{1}{450} (161 + 65\sqrt{0.7}) \approx 0.478629$ $\frac{1}{450} (161 - 65\sqrt{0.7}) \approx 0.236927$

Inne kwadratury Gaussa:

$[a, b]$	$\omega(x)$	Wielomiany ortogonalne
$[-1, 1]$	$\frac{1}{\sqrt{1-x^2}}$	$T_n(x)$ – Czebyszewa
$[0, \infty]$	e^{-x}	$L_n(x)$ – Laguerre'a
$[-\infty, \infty]$	e^{-x^2}	$H_n(x)$ – Hermite'a

$$T_n(x) = \cos(n \arccos(x)) = \frac{(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n}{2} \quad (155)$$



Metody aproksymacji

Zadanie *aproksymacji funkcji* $f(x)$ polega na wyznaczeniu parametrów $a_0, \dots, a_m$ pewnej funkcji F tak, by zminimalizować $\|f(x) - F(a_0, \dots, a_m; x)\|$.

- przestrzeń parametrów wyznacza przestrzeń funkcji
- aproksymacja *punktowa* i *integralna*
- najczęstszy przypadek: przestrzeń funkcji jest przestrzenią liniową – funkcja F ma postać *wielomianu uogólnionego*

$$F(x) = a_0\varphi_0(x) + a_1\varphi_1(x) + \dots + a_m\varphi_m(x) \quad (156)$$

- aproksymacja wymierna

$$F(x) = \frac{a_0\varphi_0(x) + a_1\varphi_1(x) + \dots + a_k\varphi_k(x)}{b_0\psi_0(x) + b_1\psi_1(x) + \dots + b_l\psi_l(x)} \quad (157)$$

Wybór przestrzeni funkcji i minimalizowanej normy

- Interpolacja - przestrzeń gwarantuje zerową wartość normy
- Aproksymacja średniokwadratowa:

$$\|f\|_{2,\omega} = \left(\int_a^b \omega(x) f(x)^2 dx \right)^{\frac{1}{2}} \quad (158)$$

w przypadku dyskretnym

$$\|f\|_{2,\omega} = \sum_{i=0}^n \omega(x_i) f(x_i)^2 \quad (159)$$

- Aproksymacja jednostajna

$$\|f\| = \sup_{x \in [a,b]} |f(x)| \quad (160)$$

Aproksymacja wielomianowa

Twierdzenie 21 (Weierstrassa) *Jeżeli funkcja $f(x)$ jest ciągła na skończonym przedziale $[a, b]$, to dla każdego $\varepsilon > 0$ można dobrać takie n , że istnieje wielomian P stopnia n , który na całym przedziale $[a, b]$ spełnia nierówność*

$$|f(x) - P(x)| < \varepsilon. \quad (161)$$

Aproksymacja średniokwadratowa wielomianami uogólnionymi

Rozpartujemy funkcje postaci

$$F(x) = \sum_{i=0}^m a_i \varphi_i(x). \quad (162)$$

$\varphi_0, \dots, \varphi_m$ to funkcje bazowe $m + 1$ wymiarowej przestrzeni liniowej.

Różne układy bazowe:

- $1, x, x^2, \dots, x^m$
- rodziny wielomianów z interpolacji Lagrange'a czy Newtona
- wielomiany ortogonalne
- $\sin x, \cos x, \sin 2x, \cos 2x, \dots, \sin kx, \cos kx$

Zadanie: minimalizacja błędu:

$$E(a_0, \dots, a_m) = \sum_{i=0}^n \omega(x_i) \left[f(x_i) - \sum_{j=0}^m a_j \varphi_j(x_i) \right]^2 \quad (163)$$

Minimum E w punkcie, gdzie pochodne cząstkowe są zerowe

$$\frac{\delta E}{\delta a_k} = - \sum_{i=0}^n 2\omega(x_i) \left[f(x_i) - \sum_{j=0}^m a_j \varphi_j(x_i) \right] \varphi_k(x_i) \quad k = 0, \dots, m \quad (164)$$

Układ normalny w postaci macierzowej:

$$\mathbf{D}^T \mathbf{D} \mathbf{A} = \mathbf{D}^T \mathbf{f} \quad (165)$$

gdzie

$$\mathbf{D} = \begin{bmatrix} \varphi_0(x_0) & \cdots & \varphi_m(x_0) \\ \vdots & & \vdots \\ \varphi_0(x_n) & \cdots & \varphi_m(x_n) \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} a_0 \\ \vdots \\ a_m \end{bmatrix}, \quad \mathbf{f} = \begin{bmatrix} f(x_0) \\ \vdots \\ f(x_n) \end{bmatrix} \quad (166)$$

Najprostsze podejście do średniokwadratowej aproksymacji wielomianowej:

Funkcje bazowe:

$$\varphi_i(x) = x^i. \quad (167)$$

Warunki minimum:

$$\frac{\delta E}{\delta a_k} = 0 \quad \equiv \quad \sum_{i=0}^n \left[f(x_i) - \sum_{j=0}^m a_j x_i^j \right] x_i^k = 0 \quad (168)$$

czyli

$$\sum_{i=0}^n f(x_i) x_i^k = \sum_{j=0}^m a_j \sum_{i=0}^n x_i^{j+k}, \quad k = 0, \dots, m \quad (169)$$

W prostszej formie:

$$\sum_{i=0}^m a_i g_{ik} = \rho_k, \quad k = 0, \dots, m \quad (170)$$

$$g_{ij} = \sum_{j=0}^n x_j^{i+k}, \quad \rho_k = \sum_{j=0}^n f(x_j) x_j^k \quad (171)$$

Uwagi:

- Dla $m \leq n$ rozwiązanie jest jednoznaczne
- Dla $m = n$ rozwiązaniem jest wielomian interpolacyjny
- Dla $m = 1$ mamy aproksymację funkcją liniową $y = a_0 + a_1 x$:

$$a_1 = \frac{EXY - EXEY}{EX^2 - (EX)^2} = \frac{cov(X, Y)}{var(X)}, \quad a_0 = EY - a_1 EX \quad (172)$$

- Zmiana stopnia wielomianu aproksymującego wymaga ponownego rozwiązania układu normalnego.

Problem złego uwarunkowania

Założmy, że punkty x_i są równo rozmieszczone wewnątrz przedziału $[0, 1]$. Dla dużych wartości n :

$$g_{ik} = \sum_{j=1}^n x_j^{i+k} \approx n \int_0^1 x_j^{i+k} = \frac{n}{i+k+1}, \quad i, k = 0, \dots, m \quad (173)$$

Macierz współczynników układu normalnego:

$$A = \begin{bmatrix} 1 & \frac{1}{2} & \cdots & \frac{1}{m+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{m+2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{m+1} & \frac{1}{m+2} & \cdots & \frac{1}{2m+1} \end{bmatrix} \quad (174)$$

Macierz odwrotna ma z kolei bardzo duże wartości co powoduje bardzo duże błędy zaokrągleń.

Aproksymacja za pomocą wielomianów ortogonalnych

Definicja 10 Funkcje $f(x)$ i $g(x)$ nazywamy ortogonalnymi na dyskretnym zbiorze punktów $x_0, \dots, x_n$ gdy

$$\sum_{i=0}^n f(x_i)g(x_i) = 0 \quad (175)$$

przy

$$\sum_{i=0}^n f^2(x_i) > 0, \quad \sum_{i=0}^n g^2(x_i) > 0. \quad (176)$$

Uwaga: Zastosowanie wielomianów ortogonalnych jako bazy przestrzeni eliminuje problem złego uwarunkowania macierzy układu normalnego. Macierz ta staje się diagonalną, z elementami przekątnej

$$a_{jj} = \sum_{i=0}^n \varphi_j^2(x_i). \quad (177)$$

Metoda wielomianów ortogonalnych

1. Każdy wielomian jest kombinacją liniową wielomianów ortogonalnych.
2. Wielomian aproksymacji to ten o najmniejszym błędzie kwadratowym.

Niech $P_0, \dots, P_m$ będą wielomianami ortogonalnymi o stopniach równych indeksom oraz $Q_m \in \Pi_m$. Wówczas istnieje jednoznaczny rozkład

$$Q_m(x) = b_0 P_0(x) + \dots + b_m P_m(x). \quad (178)$$

Współczynnik b_k łatwo wyznaczyć mnożąc tożsamość (178) przez $P_k(x)$ i sumując równości uzyskane dla podstawień $x_0, \dots, x_n$. Otrzymujemy:

$$\begin{aligned} \sum_{i=0}^n Q_m(x_i) P_k(x_i) &= b_0 \sum_{i=0}^n P_0(x_i) P_k(x_i) + \dots + b_k \sum_{i=0}^n P_k^2(x_i) + \dots \\ &\quad \dots + b_m \sum_{i=0}^n P_m(x_i) P_k(x_i) \end{aligned} \quad (179)$$

Wobec ortogonalności rodziny $\{P_i\}$ mamy:

$$\sum_{i=0}^n Q_m(x_i)P_k(x_i) = b_k \sum_{i=0}^n P_k^2(x_i). \quad (180)$$

Stąd

$$b_k = \frac{\sum_{i=0}^n Q_m(x_i)P_k(x_i)}{\sum_{i=0}^n P_k^2(x_i)}, \quad k = 0, \dots, m. \quad (181)$$

Szukamy współczynników $b_0, \dots, b_m$ minimalizujących

$$\begin{aligned} S &= \sum_{i=0}^n [b_0 P_0(x_i) + \dots + b_m P_m(x_i) - f(x_i)]^2 \\ &= \sum_{i=0}^n \left[\left(\sum_{j=0}^m b_j P_j(x_i) \right)^2 - 2 \left(\sum_{j=0}^m b_j P_j(x_i) \right) f(x_i) + f^2(x_i) \right] \end{aligned} \quad (182)$$

W wyniku ortogonalności $\{P_i\}$ oraz podstawień:

$$s_j = \sum_{i=0}^n P_j^2(x_i), \quad c_j = \sum_{i=0}^n P_j(x_i)f(x_i), \quad j = 0, \dots, m \quad (183)$$

otrzymujemy

$$\begin{aligned} S &= \sum_{i=0}^n \sum_{j=0}^m b_j^2 P_j^2(x_i) - 2 \sum_{i=0}^n \sum_{j=0}^m b_j P_j(x_i) f(x_i) + \sum_{i=0}^n f^2(x_i) = \\ &= \sum_{j=0}^m b_j^2 s_j - 2 \sum_{j=0}^m b_j c_j + \sum_{i=0}^n f^2(x_i) = \sum_{j=0}^m (b_j^2 s_j - 2b_j c_j) + \sum_{i=0}^n f^2(x_i) \end{aligned} \quad (184)$$

Ponieważ

$$\begin{aligned} b_j^2 s_j - 2b_j c_j &= s_j \left(b_j^2 - 2b_j \frac{c_j}{s_j} \right) = s_j \left(b_j^2 - 2b_j \frac{c_j}{s_j} + \frac{c_j^2}{s_j^2} \right) - \frac{c_j^2}{s_j} = \\ &= s_j \left(b_j - \frac{c_j}{s_j} \right)^2 - \frac{c_j^2}{s_j} \end{aligned} \quad (185)$$

błąd kwadratowy

$$S = \sum_{j=0}^m s_j \left(b_j - \frac{c_j}{s_j} \right)^2 - \sum_{j=0}^m \frac{c_j^2}{s_j} + \sum_{i=0}^n f^2(x_i) \quad (186)$$

osiąga minimum dla

$$b_j = \frac{c_j}{s_j}. \quad (187)$$

Zatem wielomian aproksymacyjny ma postać

$$Q_m(x) = \sum_{j=0}^m \frac{c_j}{s_j} P_j(x). \quad (188)$$

Przypadek równoodległych punktów

Dla zbioru punktów $x_i = x_0 + ih, i = 0, \dots, n$ wprowadzamy $t = \frac{x-x_0}{h}$ (liczby całkowite) i wyznaczamy odpowiednie unormowane wielomiany ortogonalne:

$$\begin{aligned} P_{k,n}(t) &= 1 + b_1 t + b_2 t(t-1) + \dots + b_k t(t-1) \cdots (t-k+1) \\ P_{k,n}(0) &= 0 \quad k = 0, 1, \dots, m \end{aligned} \quad (189)$$

Dowolny wielomian $P_{j,n}(t)$ można zapisać w postaci:

$$P_{j,n}(t) = \sum_{s=0}^j d_s (t+s)^{[s]}, \quad (190)$$

gdzie d_s są stałymi, a *wielomiany czynnikowe* $r^{[j]} \stackrel{\text{ozn}}{=} r(r-1) \cdots (r-j+1)$.

Warunki ortogonalności:

$$\sum_{i=0}^n P_{j,n}(i) P_{k,n}(i) = \sum_{i=0}^n \sum_{s=0}^j d_s (i+s)^{[s]} P_{k,n}(i) = 0, \quad j = 0, \dots, k-1, \quad (191)$$

więc by je spełnić wystarczy by

$$\sum_{i=0}^n (i+j)^{[j]} P_{k,n}(i) = 0, \quad j = 0, \dots, k-1. \quad (192)$$

Sumując

$$(i+j)^{[j]} P_{k,n}(i) = (i+j)^{[j]} + b_1 (i+j)^{[j+1]} + \dots + b_k (i+j)^{[j+k]} \quad (193)$$

po wszystkich $i = 0, \dots, n$, otrzymujemy

$$\sum_{i=0}^n (i+j)^{[j]} P_{k,n}(i) = \sum_{i=0}^n (i+j)^{[j]} + b_1 \sum_{i=0}^n (i+j)^{[j+1]} + \dots + b_k \sum_{i=0}^n (i+j)^{[j+k]} \quad (194)$$

Wykorzystując, że

$$\sum_{i=0}^n (i+j)^{[s]} = \frac{(n+j+1)^{[s+1]}}{s+1}, \quad s = j, \dots, j+k, \quad (195)$$

i dzieląc obie strony przez $(n + j + 1)^{[j+1]}$ dostajemy dla $j = 0, \dots, k - 1$

$$\frac{1}{1+j} + \frac{n}{2+j}b_1 + \dots + \frac{n^{[k]}}{k+1+j}b_k = \frac{Q(j)}{(k+1+j)^{[k]}} = 0, \quad (196)$$

gdzie $Q(j)$ jest wielomianem stopnia co najwyżej k o zerach $j = 0, \dots, k - 1$.

Przyjmujemy

$$a_k = b_k n^{[k]}, \quad Q(j) = j(j-1) \cdots (j-k+1)C_k = C_k j^{[k]}, \quad (197)$$

oraz mnożymy (196) przez $(1+j)$. Dostajemy

$$1 + \frac{1+j}{2+j}a_1 + \dots + \frac{1+j}{k+1+j}a_k = \frac{C_k j^{[k]}}{(2+j)(3+j) \cdots (k+1+j)}, \quad (198)$$

skąd dla $j = -1$ wynika

$$C_k = \frac{1 \cdot 2 \cdots k}{(-1)(-2) \cdots (-k)} = (-1)^k. \quad (199)$$

Aby wyznaczyć a_s dla $s = 1, \dots, k$ mnożymy (196) przez $(s+1+j)$ i przyjmujemy $j = -(s+1)$. Otrzymujemy:

$$\begin{aligned}
 a_s &= \frac{(-1)^k j^{[k]} (s+1+j)}{(k+1+j)^{[k+1]}} = \frac{(-1)^k (-s-1)(-s-2) \cdots (-s-k)}{(k-s)(k-s-1) \cdots 1 \cdot (-1) \cdots (-s)} \\
 &= \frac{(-1)^{2k} (k+s)^{[k]}}{(k-s)! (-1)^s s!} = (-1)^s \frac{(k+s)^{[k]} s!}{s! s! (k-s)!} = (-1)^s \frac{(k+s)!}{s! k!} \frac{k!}{s! (k-s)!} \quad (200) \\
 &= (-1)^s \binom{k+s}{s} \binom{k}{s}
 \end{aligned}$$

Zatem ostateczna postać wielomianów ortogonalnych (Czebyszewa/Grama) to:

$$P_{k,n}(t) = \sum_{s=0}^k (-1)^s \binom{k}{s} \binom{k+s}{s} \frac{t^{[s]}}{n^{[s]}}, \quad k = 0, \dots, m, \quad (201)$$

a wielomian aproksymacyjny ma postać

$$Q_m(x) = \sum_{j=0}^m \frac{c_j}{s_j} P_{j,n} \left(\frac{x - x_0}{h} \right). \quad (202)$$

Aproksymacja funkcji ciągłych

Znamy funkcję, chcemy ją aproksymować inną funkcją.

Dla funkcji aproksymowanej f ciągłej na $[a, b]$, funkcji wagowej $\omega(x)$ oraz rodziny funkcji aproksymujących

$$F(x) = a_0\varphi_0(x) + \cdots + a_n\varphi_n(x) \quad (203)$$

minimalizujemy wartość funkcji

$$E_n = \int_a^b \omega(x) [F(x) - f(x)]^2 dx = \int_a^b \left[\sum_{i=0}^n a_i \varphi_i(x) - f(x) \right]^2 dx. \quad (204)$$

Rozwiązujemy układ równań:

$$\begin{aligned} \frac{\delta E_n}{\delta a_j} &= \int_a^b \left[\sum_{i=0}^n a_i \varphi_i(x) - f(x) \right] \varphi_j(x) dx \\ &= \sum_{i=0}^n a_i \int_a^b \varphi_i(x) \varphi_j(x) dx - \int_a^b f(x) \varphi_j(x) dx = 0, \quad j = 0, \dots, n. \end{aligned} \quad (205)$$

Przykład: Znaleźć najlepszą aproksymację kwadratową funkcji $f(x) = \sqrt{x}$ w przedziale $[0, 1]$ wielomianem stopnia pierwszego $Q(x) = a_0 + a_1x$.

Otrzymujemy układ równań:

$$\begin{aligned} a_0 \int_0^1 1^2 dx + a_1 \int_0^1 x dx &= \int_0^1 \sqrt{x} dx \\ a_0 \int_0^1 x dx + a_1 \int_0^1 x^2 dx &= \int_0^1 \sqrt{x} x dx \end{aligned} \quad (206)$$

Czyli:

$$\begin{aligned} a_0 + \frac{1}{2}a_1 &= \frac{2}{3} \\ \frac{1}{2}a_0 + \frac{1}{3}a_1 &= \frac{2}{5} \end{aligned} \quad (207)$$

Rozwiązanie:

$$a_0 = \frac{4}{15}, \quad a_1 = \frac{4}{5}, \quad Q(x) = \frac{4}{15} + \frac{4}{5}x \quad (208)$$

Obliczenia mogą być mniej uciążliwe przy użyciu **wielomianów ortogonalnych**.

Przykład: Znaleźć najlepszą aproksymację kwadratową funkcji $f(x) = e^x$ w przedziale $[0, 1]$ wielomianem stopnia drugiego przy użyciu wielomianów ortogonalnych Gaussa-Legendre'a (154).

Podstawiamy $x = \frac{t+1}{2}$ i przybliżamy funkcję $e^{\frac{t+1}{2}}$ wielomianem $Q(t) = a_0P_0(t) + a_1P_1(t) + a_2P_2(t)$:

$$\begin{aligned} a_0 \int_{-1}^1 1^2 dt &= \int_{-1}^1 e^{\frac{t+1}{2}} dt \\ a_1 \int_{-1}^1 t^2 dt &= \int_{-1}^1 t e^{\frac{t+1}{2}} dt \\ a_2 \int_{-1}^1 \left(t^2 - \frac{1}{3}\right)^2 dt &= \int_{-1}^1 \left(t^2 - \frac{1}{3}\right) e^{\frac{t+1}{2}} dt \end{aligned} \tag{209}$$

W rozwiązaniu $Q(t)$ podstawiamy $t = 2x - 1$.

Metoda wyrównywania

- W przypadku pewnych nieliniowych zależności:
 - trudno skonstruować odpowiedni wielomian uniwersalny
 - trudno rozwiązać wprost układ równań nieliniowych $\frac{\delta E}{\delta a_i}$
- Zastosowanie znajduje metoda wyrównywania (linearyzacji).
- Jeśli zależność między x a y jest nieliniowa, a istnieją odwzorowania wzajemnie jednoznaczne φ, ψ takie, że zależność między $X = \varphi(x, y)$ a $Y = \psi(x, y)$ daje się opisać wielomianem uniwersalnym, to aproksymuje się zależność między X i Y i minimum przenosi do wyjściowych zmiennych.
- Nieznanej zależności między x a y często szuka się eksperymentalnie (najczęściej wizualnie) pośród następujących:

$$\begin{array}{cccc}
 y = ax + b & y = ax^b & y = ab^x & y = a + \frac{b}{x} \\
 y = \frac{1}{ax + b} & y = \frac{x}{ax + b} & y = a \ln x + b &
 \end{array} \tag{210}$$

Aproksymujemy (przy pomocy równań 172) liniową zależność

$$Y = \alpha X + \beta, \quad (211)$$

gdzie stosujemy następujące podstawienia:

zależność x i y	podstawienia
$y = bx^a$	$X = \log x, Y = \log y, \beta = \log b$
$y = ba^x$	$Y = \log y, \alpha = \log a, \beta = \log b$
$y = a + \frac{b}{x}$	$Y = xy$
$y = \frac{1}{ax+b}$	$Y = \frac{1}{y}$
$y = \frac{x}{ax+b}$	$Y = \frac{x}{y}$
$y = a \ln x + b$	$X = \log x$

Brak specyfikacji podstawienia dla danej zmiennej oznacza odpowiednio:

$$X = x, \quad Y = y, \quad \alpha = a, \quad \beta = b \quad (212)$$

Uwaga: Minimalizacja błędu średniokwadratowego dla ww podstawień gwarantuje optymalną wartość funkcji błędu dla oryginalnych zmiennych tylko wówczas gdy błąd ten jest zerowy.

Aproksymacja funkcjami sklejanymi

Przypadek równoodległych węzłów

Dany jest podział $\Delta = \{x_i : i = 0, \dots, n\}$ przedziału $[a, b]$ taki, że $x_i = x_0 + ih$, gdzie $h = \frac{b-a}{n}$.

Twierdzenie 22 (Baza przestrzeni funkcji sklepanych) Niech funkcje $\Phi_i^3(x), i = -1, 0, \dots, n, n+1$ będą określone wzorem

$$\Phi_i^3(x) = \frac{1}{h^3} \begin{cases} (x - x_{i-2})^3 & \text{dla } x \in [x_{i-2}, x_{i-1}] \\ h^3 + 3h^2(x - x_{i-1}) + 3h(x - x_{i-1})^2 - 3(x - x_{i-1})^3 & \text{dla } x \in [x_{i-1}, x_i] \\ h^3 + 3h^2(x_{i+1} - x) + 3h(x_{i+1} - x)^2 - 3(x_{i+1} - x)^3 & \text{dla } x \in [x_i, x_{i+1}] \\ (x_{i+2} - x)^3 & \text{dla } x \in [x_{i+1}, x_{i+2}] \\ 0 & \text{w pozostałych przyp.} \end{cases} \quad (213)$$

Funkcje $\bar{\Phi}_i(x) = \Phi_i^3(x)$ określone na $[a, b]$ stanowią bazę przestrzeni funkcji sklepanych trzeciego stopnia na Δ .

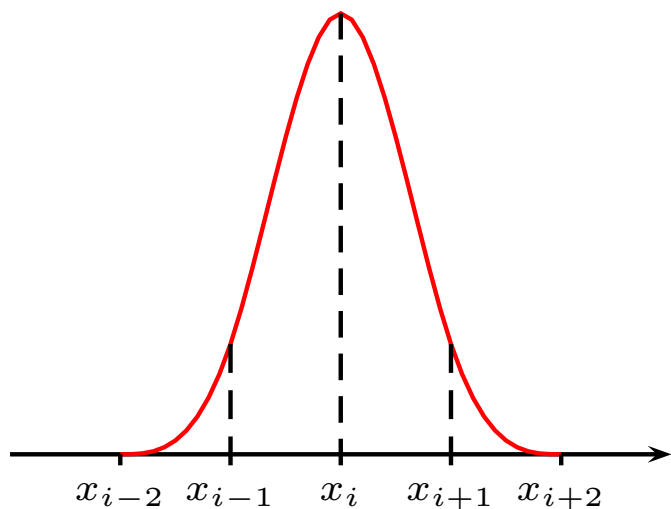
Wniosek 23 Każdą funkcję sklejaną trzeciego stopnia $S_\Delta(x)$ można przedstawić w postaci:

$$S_\Delta(x) = \sum_{i=-1}^{n+1} c_i \bar{\Phi}_i(x), \quad (214)$$

gdzie c_i są liczbami rzeczywistymi.

Własności funkcji bazowych:

Wykres $\Phi_i^3(x)$



Wartości $\Phi_i^3(x)$ i pochodnych

x	x_{i-2}	x_{i-1}	x_i	x_{i+1}	x_{i+2}
$\Phi_i^3(x)$	0	1	4	1	0
$(\Phi_i^3(x))'$	0	$\frac{3}{h}$	0	$-\frac{3}{h}$	0
$(\Phi_i^3(x))''$	0	$\frac{6}{h^2}$	$-\frac{12}{h^2}$	$\frac{6}{h^2}$	0

Są klasy $C^2((-\infty, \infty))$

Aproksymacja funkcjami sklejanymi z równoodległymi węzłami

Minimalizujemy:

$$E = \int_a^b \left[f(x) - \sum_{i=-1}^{n+1} c_i \Phi_i^3(x) \right]^2 dx \quad (215)$$

Minimum wyznacza rozwiązanie układu $n + 3$ równań z $n + 3$ niewiadomymi:

$$\frac{\delta E}{\delta c_j} = 0, \quad j = -1, \dots, n + 1 \quad (216)$$

czyli:

$$\sum_{i=-1}^{n+1} c_i \int_a^b \Phi_i^3(x) \Phi_j^3(x) dx = \int_a^b f(x) \Phi_j^3(x) dx, \quad j = -1, \dots, n + 1 \quad (217)$$

Wprowadzając $a_{ij} = \int_a^b \Phi_i^3(x) \Phi_j^3(x) dx$ otrzymujemy układ równań o symetrycznej siedmiodiagonalnej macierzy głównej.

Aproksymacja funkcjami sklejanymi dla dyskretnego zbioru punktów

Odcięte węzłów: $\{z_i, i = 0, \dots, m\}$, gdzie $m > n + 3$.

$$J = \sum_{k=0}^m \left[f(z_k) - \sum_{i=-1}^{n+1} c_i \Phi_i^3(z_k) \right]^2 \quad (218)$$

Z przyrównania pochodnych cząstkowych do zera mamy układ:

$$\sum_{i=-1}^{n+1} b_{ij} c_i = \sum_{k=0}^m f(z_k) \Phi_j^3(z_k), \quad j = -1, \dots, n+1 \quad (219)$$

gdzie

$$b_{ij} = \sum_{k=0}^m \Phi_i^3(z_k) \Phi_j^3(z_k) \quad (220)$$

jest symetryczną siedmiodiagonalną macierzą.

Jeśli wyznacznik układu jest różny od zera, to jego rozwiązanie określa minimum funkcji J .

Aproksymacja jednostajna

Funkcja błędu:

$$E = \sup_{x \in [a, b]} |f(x) - F(x)| \quad (221)$$

- Z twierdzenia Weierstrassa (tw. **21**) wynika, że dowolną funkcję f ciągłą na $[a, b]$ można aproksymować jednostajnie wielomianem z dowolną dokładnością.
- Twierdzenie Borela mówi, że dla każdej funkcji $f(x)$ na przedziale $[a, b]$ oraz $n \in \mathbb{N}$ istnieje wielomian $W_n(x)$ o minimalnym błędzie aproksymacji jednostajnej.
- Twierdzenie Czebyszewa mówi, że taki wielomian jest tylko jeden.

Brak ogólnej metody wyznaczania wielomianu najlepszego przybliżenia jednostajnego.

Metoda szeregów potęgowych

Jeżeli funkcję $f(x)$ można na $[a, b]$ rozwinąć w szereg Taylora, to przybliżeniem może być odpowiednie obcięcie szeregu:

$$W_n(x) = f(x_0) + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{2!}(x - x_0)^2 + \dots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n \quad (222)$$

Oszacowanie błędu z reszty w postaci Lagrange'a:

$$|f(x) - W_n(x)| = \left| \frac{f^{(n+1)}(c)}{(n+1)!}(x - x_0)^{n+1} \right| \leq \frac{M_{n+1}}{(n+1)!} \delta^{n+1} \quad (223)$$

dla c leżącego między x i x_0 , δ będącego promieniem stosownego otoczenia punktu x_0 oraz $M_{n+1} \geq |f^{(n+1)}(x)|$ dla $x \in [a, b]$.

Odpowiednią dokładność można uzyskać dobierając odpowiednio duże n .

Przykład: Aproksymować metodą szeregów potęgowych funkcję $\sqrt{x}$ wielomianem stopnia 2 w przedziale $[0,5, 1,5]$. Oszacować błąd.

Przybliżenia Padégo

Dokładniejsze od wielomianowych często okazują się być wymierne:

$$R_{n,k}(x) = \frac{L_n(x)}{M_k(x)}, \quad (224)$$

gdzie

$$\begin{aligned} L_n(x) &= a_0 + a_1x + \cdots + a_nx^n \\ M_k(x) &= b_0 + b_1x + \cdots + b_kx^k \end{aligned} \quad (225)$$

- Parametry określa się tak, by w $x_0 = 0$ wartości funkcji f i $R_{n,k}$ oraz jak najwięcej ich pochodnych zrównało się.
- Dla $k = 0$ uzyskujemy przybliżenie szeregiem potęgowym Maclaurina.
- $f(x)$ przybliża się szeregiem Maclaurina i rozwiązuje układ równań wynikających z założenia równości wartości funkcji i pochodnych.
- Stosunkowo małe błędy w pobliżu zera, większe dalej.

Szeregi Czebyszewa

Przybliżanie funkcji $f(x)$ sumą częściową szeregu

$$f(x) \approx \frac{1}{2}c_0 + \sum_{j=1}^n c_j T_j(x), \quad (226)$$

gdzie

$$c_j = \frac{2}{\pi} \int_{-1}^1 \frac{f(x)T_j(x)}{\sqrt{1-x^2}} dx \quad T_j(x) = \cos(j \arccos x). \quad (227)$$

Alternatywnie, funkcję można aproksymować funkcją wymierną

$$T_{n,k}(x) = \frac{\sum_{i=0}^n a_i T_i(x)}{\sum_{i=0}^k b_i T_i(x)}. \quad (228)$$

Wartości a_i, b_i wyznacza się tak, by w różnicy $f(x) - T_{n,k}(x) = \frac{1}{2}c_0 + \sum_{j=1}^{\infty} c_j T_j(x) - T_{n,k}(x)$ zniknęły współczynniki przy T_j dla $j = 0, \dots, N$.

Rozwiązywanie równań nieliniowych

Zadanie: Najczęstsze postaci:

1. Znaleźć pierwiastki rzeczywiste równania $f(x) = 0$, funkcji f zmiennej rzeczywistej.
2. Znaleźć pierwiastki rzeczywiste i zespolone równania $P(x) = 0$ wielomianu P .

Metody iteracyjne

- Ciąg kolejnych przybliżeń $\{x_i\}$.
- n -punktowy wzór iteracyjny

$$x_{i+1} = F_i(x_i, x_{i-1}, \dots, x_{i-n+1}). \quad (229)$$

- Funkcja F_i zwykle zależy od wartości funkcji f i od jej pochodnych.
- Funkcja iteracyjna jest *stacjonarna* gdy jest niezależna od indeksu i . Wówczas pierwiastki są punktami stałymi F :

$$\alpha = F(\alpha, \dots, \alpha). \quad (230)$$

- Niech α będzie szukanym pierwiastkiem. Błąd i -tej iteracji $\varepsilon_i = \alpha - x_i$. Jeśli $\lim_{i \rightarrow \infty} x_i = \alpha$, to mówimy, że metoda jest w punkcie α rzędu $p \geq 1$ wtw gdy

$$\lim_{i \rightarrow \infty} \frac{|x_{i+1} - \alpha|}{|x_i - \alpha|^p} = \lim_{i \rightarrow \infty} \frac{|\varepsilon_{i+1}|}{|\varepsilon_i|^p} = C \neq 0. \quad (231)$$

Wówczas stałą C nazywamy *stałą asymptotyczną błędu*.

- Efektywność metody - odwrotnie proporcjonalna do iloczynu kosztu ϑ jednej iteracji i liczby iteracji I .

Niech rzędy dwóch metod będą odpowiednio równe p_1 i p_2 . Dla dużych i mamy w przybliżeniu:

$$|\varepsilon_{i+1}^{(1)}| = C_1 |\varepsilon_i^{(1)}|^{p_1}, \quad |\varepsilon_{i+1}^{(2)}| = C_2 |\varepsilon_i^{(2)}|^{p_2}. \quad (232)$$

Niech

$$S_i = -\log |\varepsilon_i^{(1)}|, \quad T_i = -\log |\varepsilon_i^{(2)}|. \quad (233)$$

Wtedy

$$S_{i+1} = -\log C_1 + p_1 S_i, \quad T_{i+1} = -\log C_2 + p_2 T_i. \quad (234)$$

Rozwiązania:

$$S_i = S_1 p_1^i - \log C_1^{(p_1^i - 1)/(p_1 - 1)}, \quad T_i = T_1 p_2^i - \log C_2^{(p_2^i - 1)/(p_2 - 1)}. \quad (235)$$

Zakładamy, że $S_1 = T_1$, koszt jednej iteracji to odpowiednio ϑ_1 i ϑ_2 , a liczby kroków potrzebne do uzyskania odpowiedniej dokładności to I_1 oraz I_2 .

Zatem S_{I_1} i T_{I_2} są w przybliżeniu równe. Stąd:

$$S_1(p_1^{I_1} - p_2^{I_2}) + \log \frac{C_2^{(p_2^{I_2} - 1)/(p_2 - 1)}}{C_1^{(p_1^{I_1} - 1)/(p_1 - 1)}} = 0. \quad (236)$$

Najczęściej pierwszy składnik dominuje, więc można drugi zaniedbać. Zatem:

$$I_1 \log p_1 \approx I_2 \log p_2 \approx \text{const}. \quad (237)$$

Ostatecznie wskaźnik efektywności można określić jako:

$$\mathcal{E} = \frac{1}{\vartheta} \log p \quad \text{albo} \quad \mathcal{E} = p^{\frac{1}{\vartheta}}. \quad (238)$$

- Rząd metody jest cechą lokalną.
- Koszt jednej iteracji $\approx$ koszt liczenia wartości funkcji i jej pochodnych.

Metoda interpolacji odwrotnej

Zakładamy, że w okolicy pierwiastka α funkcji f istnieje funkcja odwrotna $g = f^{-1}$ (α jest pierwiastkiem pojedynczym).

Niech $x_{i-n}, \dots, x_i$ będą kolejnymi przybliżeniami α (bądź pewnymi punktami z otoczenia α dla $i = 0$). Przyjmujemy

$$y_j = f(x_{i-j}), \quad j = 0, \dots, n. \quad (239)$$

Możemy interpolować funkcję $g(y)$ wielomianem interpolacyjnym Lagrange'a

$$h(y) = \sum_{j=0}^n L_j(y)g(y_j) = \sum_{j=0}^n L_j(y)x_{i-j}. \quad (240)$$

Wówczas kolejnym przybliżeniem x_{i+1} pierwiastka funkcji f jest:

$$h(0) = \sum_{j=0}^n L_j(0)x_{i-j} = \sum_{j=0}^n \frac{x_{i-j}(-1)^n y_0 \cdots y_{j-1} y_{j+1} \cdots y_n}{(y_j - y_0) \cdots (y_j - y_{j-1})(y_j - y_{j+1}) \cdots (y_j - y_n)}. \quad (241)$$

Ze wzoru Lagrange'a:

$$\alpha - x_{i+1} = \frac{g^{(n)}(\eta)}{n!} (-1)^{n+1} y_0 \cdots y_n, \quad \eta \in I[y_0, \dots, y_n, 0]. \quad (242)$$

- Można użyć dowolnej metody interpolacji.

Metoda bisekcji

Szukamy zera α w przedziale (a, b) , takim, że $f(a)f(b) < 0$. Przyjmujemy $a_0 = a, b_0 = b$, a także

$$x_i = \frac{a_i + b_i}{2}, \quad i = 0, 1, \dots, \quad (243)$$

oraz

$$(a_{i+1}, b_{i+1}) = \begin{cases} (a_i, x_i) & \text{◦ ile } f(a_i)f(x_i) < 0 \\ (x_i, b_i) & \text{◦ ile } f(x_i)f(b_i) < 0 \end{cases}. \quad (244)$$

Zakładamy, że $f(x_i) \neq 0$, bo inaczej mamy rozwiązanie dokładne.

Ponieważ

$$|x_i - x_{i+1}| = \frac{1}{2^{i+1}}(b - a), \quad i = 0, 1, \dots, \quad (245)$$

to

$$\lim_{i \rightarrow \infty} x_i = \alpha. \quad (246)$$

Regula falsi

regula – linia, *falsus* – fałszywy — fałszywość założenia liniowości funkcji

Założenia: funkcja f jest ciągła, $f(a)f(b) < 0$.

Zakładamy $(a_0, b_0) = (a, b)$. W i -tym kroku prowadzimy cięciwę o równaniu

$$y - f(a_i) = \frac{f(b_i) - f(a_i)}{b_i - a_i}(x - a_i) \quad (247)$$

i przybliżamy pierwiastek równania $f(x) = 0$ pierwiastkiem cięciwy:

$$x_i = a_i - \frac{f(a_i)}{f(b_i) - f(a_i)}(b_i - a_i). \quad (248)$$

W kolejnych krokach

$$(a_{i+1}, b_{i+1}) = \begin{cases} (a_i, x_i) & \text{◦ ile } f(a_i)f(x_i) < 0 \\ (x_i, b_i) & \text{◦ ile } f(x_i)f(b_i) < 0 \end{cases}. \quad (249)$$

- na ogół niestacjonarna, stacjonarna np. dla funkcji wypukłych

Oszacowania błędów metody *regula falsi*

Z twierdzenia Lagrange'a o przyrostach:

$$f(x_n) - f(\alpha) = f'(c)(x_n - \alpha), \quad \text{dla pewnego } c \in I[x_n, \alpha] \quad (250)$$

Ponieważ $f(\alpha) = 0$, więc

$$|x_n - \alpha| \leqslant \frac{|f(x_n)|}{m}, \quad m = \inf_{x \in [a, b]} |f'(x)| \quad (251)$$

Przy założeniu, że f' i f'' nie zmieniają znaku w $[a, b]$ jeden z końców przedziału nie zmienia się w kolejnych krokach iteracji. Przyjmijmy $f'(x) > 0, f''(x) > 0$. Stąd $b_i = b, i = 0, 1, \dots$ czyli

$$x_{-1} = a, \quad x_{i+1} = x_i - \frac{f(x_i)}{f(b) - f(x_i)}(b - x_i), \quad i = 0, 1, \dots \quad (252)$$

Mamy więc

$$f(\alpha) - f(x_i) = \frac{f(x_i) - f(b)}{x_i - b}(x_{i+1} - x_i), \quad (253)$$

i z tw. Lagrange'a

$$(\alpha - x_i)f'(\xi_i) = (x_{i+1} - x_i)f'(\bar{x}_i), \quad \xi_i \in (x_i, \alpha), \quad \bar{x}_i \in (x_i, b). \quad (254)$$

Po wymnożeniu i dodaniu obustronnie $-x_{i+1}f'(\xi_i)$ otrzymujemy

$$|\alpha - x_{i+1}| = \frac{|f'(\bar{x}_i) - f'(\xi_i)|}{|f'(\xi_i)|} |x_{i+1} - x_i|. \quad (255)$$

Ponieważ $\xi_i, \bar{x}_i \in [a, b]$ i $f'(x) > 0$, więc

$$|f'(\bar{x}_i) - f'(\xi_i)| \leq M - m, \quad m = \inf_{x \in [a, b]} |f'(x)|, \quad M = \sup_{x \in [a, b]} |f'(x)|. \quad (256)$$

Ostatecznie

$$|\alpha - x_{i+1}| \leq \frac{M - m}{m} |x_{i+1} - x_i|. \quad (257)$$

Oszacowanie to jest na ogół pesymistyczne.

Dla przybliżeń w niewielkim otoczeniu α można szacować:

$$|\alpha - x_{i+1}| \approx \left| \frac{f(x_{i+1})}{f'(x_{i+1})} \right| \approx \left| \frac{x_{i+1} - x_i}{f(x_{i+1}) - f(x_i)} \right| |f(x_{i+1})| \quad (258)$$

Metoda siecznych

- Ta sama zasada, co w *regula falsi*
- Rezygnujemy z założenia $f(a)f(b) < 0$, czasami kosztem zbieżności
- Metoda stacjonarna:

$$x_0 = a, x_1 = b, \quad x_{i+1} = x_i - f(x_i) \frac{x_i - x_{i-1}}{f(x_i) - f(x_{i-1})}, \quad i = 1, 2, \dots \quad (259)$$

- Istotne znaczenie ma maksymalna graniczna dokładność (wynikająca z przyjętej arytmetyki) – gdy $x_i - x_{i-1}$ jest tego samego rzędu co oszacowanie błędu $|\alpha - x_i|$, to kolejne przybliżenia mogą być całkowicie błędne.
Dodatkowe kryterium stopu: zaburzenie zmniejszania się $|f(x_i)|$.
- Wykrywanie rozbieżności – różnica pomiędzy kolejnymi przybliżeniami rośnie.

Metoda Newtona

- Założenia: $f(a)f(b) < 0$, f' oraz f'' nie zmieniają znaku na $[a, b]$.
- Z tego końca przedziału $x_0 \in \{a, b\}$, w którym $f(x)$ ma ten sam znak co $f''(x)$ prowadzimy styczną do wykresu funkcji $y = f(x)$. Punkt x_1 przecięcia stycznej z osią odciętych to pierwsze przybliżenie pierwiastka.
- Prowadzimy styczną w punkcie x_1 i wyznaczamy punkt x_2 przecięcia tej stycznej z osią.
- Uzyskany ciąg $x_1, x_2, \dots$ jest zbieżny monotonicznie do α .

Dowód:

1. Ciąg jest monotoniczny.

2. Zbiega do α .

Przypadek $f'(x) > 0$, $f''(x) > 0$:

Wówczas $x_0 = b$ oraz $f(x_0) > 0$. Pokażemy, że $x_{n+1} < x_n$ dla $n = 0, 1, \dots$

Równanie stycznej:

$$y - f(x_n) = f'(x_n)(x - x_n). \quad (260)$$

Zero stycznej w punkcie

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \quad (261)$$

Ze wzoru Taylora:

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{1}{2}f''(c)(\alpha - x_n)^2, \quad c \in (\alpha, x_n). \quad (262)$$

Stąd

$$\alpha = x_n - \frac{f(x_n)}{f'(x_n)} - \frac{1}{2} \frac{f''(c)}{f'(x_n)} (\alpha - x_n)^2. \quad (263)$$

Z (261):

$$\alpha - x_{n+1} = -\frac{1}{2} \frac{f''(c)}{f'(x_n)} (\alpha - x_n)^2 < 0 \quad (264)$$

Zatem $f(x_{n+1}) > 0$ oraz $x_{n+1} < x_n$, bo

$$x_{n+1} - x_n = -\frac{f(x_n)}{f'(x_n)} < 0. \quad (265)$$

Ciąg $x_n : n = 0, 1, \dots$ jest więc monotoniczny i ograniczony, a zatem zbieżny (do pewnego g). Z (261) przy $n \rightarrow \infty$ mamy

$$g = g - \frac{f(g)}{f'(g)}, \quad (266)$$

czyli $f(g) = 0$, a więc $g = \alpha$. □

Błąd przybliżenia x_n .

Podobnie jak w metodzie *regula falsi* (str. 137)

$$|x_n - \alpha| \leq \frac{|f(x_n)|}{m}, \quad m = \inf_{x \in [a, b]} |f'(x)| \quad (267)$$

Ze wzoru Taylora, dla pewnego $\xi_n \in I[x_n, x_{n+1}]$:

$$f(x_{n+1}) = f(x_n) + f'(x_n)(x_{n+1} - x_n) + \frac{1}{2}f''(\xi_n)(x_{n+1} - x_n)^2 \quad (268)$$

Z (261) $f(x_n) + f'(x_n)(x_{n+1} - x_n) = 0$, więc

$$|f(x_{n+1})| \leq \frac{1}{2} M (x_{n+1} - x_n)^2, \quad M = \sup_{x \in [a, b]} |f''(x)|. \quad (269)$$

Zatem

$$|\alpha - x_{n+1}| \leq \frac{1}{2} \frac{M}{m} (x_{n+1} - x_n)^2 = \frac{M}{2m} \left(\frac{f(x_n)}{f'(x_n)} \right)^2. \quad (270)$$

Alternatywne oszacowanie (w bliskim sąsiedztwie α):

$$|\alpha - x_n| \approx \left| \frac{f(x_n)}{f'(x_n)} \right|. \quad (271)$$

Uwagi:

- Metoda siecznych, to metoda Newtona z przybliżaniem pochodnej ilorazami różnicowymi.
- Znane zastosowanie metody Newtona - szukanie pierwiastków kwadratowych liczb rzeczywistych jako dodatnich pierwiastków równań $x^2 - c = 0$.
- Metoda Newtona zwykle potrzebuje mniej iteracji niż metoda siecznych, ale ma wyższy koszt jednej iteracji.

Rząd stacjonarnej metody jednopunktowej

Dla metod, w których

$$x_{n+1} = \Phi(x_n), \quad (272)$$

gdzie Φ jest wystarczająco wiele razy różniczkowalna w otoczeniu $O(\alpha)$ wyznaczonej wartości α , jeżeli $x_i \in O(\alpha)$ oraz $\Phi^{(k)}(\alpha) = 0$ dla $k = 1, 2, \dots, p-1$ i $\Phi^{(p)}(\alpha) \neq 0$, to zachodzą:

$$x_{i+1} = \Phi(x_i) = \Phi(\alpha) + \frac{(x_i - \alpha)^p}{p!} \Phi^{(p)}(\alpha) + R(\|x_i - \alpha\|^{p+1})$$

$$\lim_{i \rightarrow \infty} \frac{x_{i+1} - \alpha}{(x_i - \alpha)^p} = \frac{\Phi^{(p)}(\alpha)}{p!} \quad (273)$$

Dla $p \geq 2$ metoda jest rzędu p . Metoda jest rzędu pierwszego gdy $p = 1$ oraz $|\phi'(\alpha)| < 1$.

Dla metody Newtona mamy:

$$\Phi(x) = x - \frac{f(x)}{f'(x)}. \quad (274)$$

Metoda jest rzędu 2, bo:

$$\Phi(\alpha) = \alpha, \quad \Phi'(\alpha) = \left. \frac{f(x)f''(x)}{f'(x)^2} \right|_{x=\alpha} = 0, \quad \Phi''(\alpha) = \frac{f''(\alpha)}{f'(\alpha)} \quad (275)$$

Modyfikacje metod

- Dotychczasowe rozważania dotyczyły pierwiastków pojedynczych.

- Pierwiastki wielokrotne

Definicja 11 Liczba α jest pierwiastkiem r -krotnym ($r \geq 2$) równania $f(x) = 0$ wtw gdy jest $r - 1$ -krotnym pierwiastkiem równania $f'(x) = 0$.

- Metoda połowienia, *regula falsi*, metoda siecznych
 - tracą sens dla pierwiastków parzystokrotnych
 - dla krotności nieparzystych przy $f(a)f(b) < 0$ pozostają słuszne, choć dla siecznych obniża się rząd.
- Metoda Newtona
 - pozostaje słuszna dla parzystokrotnych o ile w jednostronnym sąsiedztwie zera spełnione są założenia stałości znaków pochodnych
 - dla pierwiastków wielokrotnych obniżony rząd metody – tylko liniowa zbieżność ze stałą asymptotyczną $C = \frac{r-1}{r}$

Przykłady modyfikacji metod

- Znamy krotność pierwiastka:

$$x_{n+1} = x_n - r \frac{f(x_n)}{f'(x_n)} \quad (276)$$

- Krotność nieznana. Zamiast funkcji $f(x)$ używamy funkcji

$$u(x) = \frac{f(x)}{f'(x)}, \quad (277)$$

która ma te same pierwiastki co $f(x)$, ale wszystkie pojedyncze. Wówczas dla metod *regula falsi* i siecznych wzory (248) i (259) przyjmują postać

$$x_{n+1} = x_n - u(x_n) \frac{x_n - x_{n-1}}{u(x_n) - u(x_{n-1})}, \quad (278)$$

a dla metody Newtona mamy:

$$x_{n+1} = x_n - \frac{u(x)}{u'(x)}, \quad \text{przy } u'(x_n) = 1 - \frac{f''(x_n)}{f'(x_n)} u(x_n). \quad (279)$$

Zera wielomianów

Liczba zer wielomianu

Definicja 12 Ciągiem Sturma dla wielomianu $p(x)$ nazywamy ciąg wielomianów $p_0(x), \dots, p_m(x)$ taki, że:

1. $p_0(x) = p(x)$.
2. $p_1(x) = p'_0(x)$.
3. dla $i = 2, \dots, m+1$ $p_i(x)$ jest resztą z dzielenia p_{i-2} przez p_{i-1} wziętą ze znakiem przeciwnym.
4. $p_{m+1}(x) \equiv 0$

Oznaczenie: $N(z) \stackrel{\text{ozn}}{=} \text{liczba zmian znaku w ciągu } \{p_i(z) : i = 0, \dots, m, p_i(z) \neq 0\}$

Twierdzenie 24 (Sturma) Jeżeli ciąg $\{p_i(x) : i = 0, \dots, m\}$ jest ciągiem Sturma dla wielomianu $p(x)$ na przedziale (a, b) oraz $p_0(a)p_0(b) \neq 0$, to liczba różnych zer rzeczywistych wielomianu $p(x)$ w przedziale (a, b) jest równa $N(a) - N(b)$.

Oznaczenie: $M(z) \stackrel{\text{ozn}}{=} \text{liczba zmian znaku w ciągu } \{p^{(i)}(z) : i = 0, 1, \dots\}.$

Twierdzenie 25 (Fouriera) *Jeżeli $p(x) \in \Pi_n$ i $p(a)p(b) \neq 0$, to liczba zer wielomianu $p(x)$ w przedziale $[a, b]$ jest równa $M(a) - M(b)$ lub jest od tej liczby mniejsza o liczbę parzystą.*

Oznaczenie: Dla wielomianu $p(x) = a_0x^n + \dots + a_{n-1}x + a_n$, gdzie $a_0 \neq 0$, $L(z) \stackrel{\text{ozn}}{=} \text{liczba zmian znaku ciągu } \{p_k(z) = \sum_{i=0}^k a_i z^{k-i} : k = 0, \dots, n\}.$

Twierdzenie 26 (Laguerre'a) *Jeżeli $p(x) \in \Pi_n$ i $p(a)p(b) \neq 0$, to liczba zer wielomianu $p(x)$ w przedziale $[a, b]$ jest równa $L(a) - L(b)$ lub jest od tej liczby mniejsza o liczbę parzystą.*

Przypadek szczególny – *reguła Kartezjusza* – liczba pierwiastków dodatnich wielomianu jest równa liczbie zmian znaku w ciągu współczynników minus liczba parzysta.

Lokalizacja przedziału z zerami

- Zadanie: wyznaczyć przedział zawierający wszystkie zera wielomianu.
- Redukcja problemu: wyznaczyć kres górny R zbioru zer, bo jeśli

$$p_1(x) = x^n p\left(\frac{1}{x}\right), \quad p_2(x) = p(-x), \quad p_3(x) = x^n p\left(\frac{-1}{x}\right) \quad (280)$$

mają kresy górne odpowiednio R_1 , R_2 i R_3 , to wszystkie dodatnie zera wielomianu $p(x)$ leżą w $(\frac{1}{R_1}, R)$, a ujemne w $(-R_2, -\frac{1}{R_3})$

Twierdzenie 27 (Lagrange’a) Niech $a_0 \neq 0$ i a_k ($k \geq 1$) będzie pierwszym ujemnym współczynnikiem wielomianu $p(x) = a_0x^n + \dots + a_{n-1}x + a_n$. Wszystkie zera tego wielomianu są mniejsze niż

$$R = 1 + \sqrt[k]{\frac{A}{|a_0|}}, \quad (281)$$

gdzie A oznacza maksimum modułu ujemnych współczynników wielomianu. Jeżeli wszystkie współczynniki wielomianu są nieujemne, to nie ma on zer dodatnich.

Wyznaczanie zer wielomianów

- Znamy kres górny – można metodą Newtona znaleźć największy pierwiastek.
- Metoda Newtona może wolno zbiegać do największego zera. Metoda podwójnego kroku:

$$x_{k+1} = x_k - 2 \frac{p(x_k)}{p'(x_k)} \quad (282)$$

Twierdzenie 28 Niech $p(x) \in \Pi_n, n \geq 2$ ma wszystkie zera rzeczywiste $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_n$. Niech β_1 będzie największym zerem wielomianu $p'(x)$ ($\alpha_1 \geq \beta_1 \geq \alpha_2$). W przypadku $n = 2$ dodatkowo zakładamy $\alpha_1 > \alpha_2$. Dla każdego $z > \alpha_1$ dobrze określone są liczby

$$z' = z - \frac{p(z)}{p'(z)}, \quad y = z - 2 \frac{p(z)}{p'(z)}, \quad y' = y - \frac{p(y)}{p'(y)} \quad (283)$$

i spełniają nierówności

$$\beta_1 < y, \quad \alpha_1 \leq y' \leq z'. \quad (284)$$

Dla ciągu przybliżeń $\{x_n : i = 1, 2, \dots\}$ (zgodnych z (282)) mamy dwie możliwości:

1. $\forall_n x_n \geq \alpha_1$ oraz $\lim_{n \rightarrow \infty} x_n = \alpha_1$.
2. $\exists_{k_0} p(x_{k_0})p(x_0) < 0$ oraz $\forall_{k < k_0} p(x_k)p(x_0) > 0$. Wówczas ciąg

$$y_0 = x_{k_0}, \quad y_{k+1} = y_k - \frac{p(y_k)}{p'(y_k)}, \quad k = 1, 2, \dots \quad (285)$$

zbiega do α_1 .

Deflacja – „oddzielamy” pierwiastek α_1 od pozostałych – wyznaczamy wielomian stopnia $n - 1$

$$p_1(x) = \frac{p(x)}{x - \alpha_1}. \quad (286)$$

Największym zerem wielomianu $p_1(x)$ jest α_2 , a dobrym punktem startowym jest ewentualnie „przestrzelone” x_k .

Uwaga na błędy zaokrągleń – dla zera o najmniejszej wartości bezwzględnej współczynniki należy wyznaczać od najwyższej potęgi, dla największej wartości bezwzględnej odwrotnie.

Źle uwarunkowane równania algebraiczne

- pierwiastki wielokrotne są na ogół źle uwarunkowane
- bliskie sobie pierwiastki
- nawet „wyraźnie rozdzielone” pierwiastki:

Przykład: Wielomian o pierwiastkach rzeczywistych $1, 2, \dots, 20$:

$$p(z) = (z - 1)(z - 2) \cdots (z - 20) = z^{20} - 210z^{19} + \cdots + 20! \quad (287)$$

Jeśli zmniejszyć a_1 o $2^{-23} \approx 10^{-10}$, to równanie ma pierwiastki:

$$16.730737466 \pm 2.812624894i \quad (288)$$

Dla schematu Hornera błąd obliczonej wartości $p(x)$ może być równy

$$E = \sum_{i=0}^n |(2i + 1)a_{n-i}x^i| 1.06\epsilon \quad (289)$$

Błąd względny pierwiastka może osiągać $\frac{E}{\alpha p'(\alpha)}$

Stąd wskaźnik uwarunkowania:

$$C = \frac{\sum |(2i+1)a_{n-i}\alpha^i|}{|\alpha p'(\alpha)|} = \frac{\sum |(2i+1)a_{n-i}\alpha^i|}{\sum |ia_{n-i}\alpha^i|} \quad (290)$$

Dla badanego wielomianu:

$$C \approx 1.35 \cdot 10^{15}, \quad \text{dla } \alpha = 14. \quad (291)$$

Gdy pierwiastkami są liczby $\alpha_i = 2^{i-21}$, $i = 1, 2, \dots, 20$, to wskaźnik $C = 3.83 \cdot 10^3$.

Miara rozdzielania względnego: Dla k pierwiastków rzeczywistych

$\alpha_i < \dots < \alpha_k$:

$$s \stackrel{\text{def}}{=} \prod_{i=1}^{k-1} \frac{\alpha_{i+1} - \alpha_i}{|\alpha_i|} \quad \alpha_i \neq 0 \quad (292)$$

W przypadku pierwiastków $1, 2, \dots, 20$ mamy $s = 8.2 \cdot 10^{-18}$. Dla pierwiastków $\alpha_i = 2^{i-21}$, $i = 1, 2, \dots, 20$ mamy $s = 1$.

Algorytm Maehly'ego dla wyeliminowania deflacji:

Ponieważ

$$p'_1(x) = \frac{p'(x)}{x - \alpha_1} - \frac{p(x)}{(x - \alpha_1)^2}, \quad (293)$$

więc

$$x_{k+1} = x_k - \frac{p_1(x_k)}{p'_1(x_k)} = x_k - \frac{p(x_k)}{p'(x_k) - \frac{p(x_k)}{x_k - \alpha_1}}. \quad (294)$$

Ogólnie mamy

$$p_j(x) = \frac{p(x)}{(x - \alpha_1) \cdots (x - \alpha_j)} \quad (295)$$

$$p'_j(x) = \frac{p'(x)}{(x - \alpha_1) \cdots (x - \alpha_j)} - \frac{p(x)}{(x - \alpha_1) \cdots (x - \alpha_j)} \sum_{i=1}^j \frac{1}{x - \alpha_i}. \quad (296)$$

Stąd iterujemy

$$x_{k+1} = x_k - \frac{p(x_k)}{p'(x_k) - \sum_{i=1}^j \frac{p(x_k)}{x_k - \alpha_i}}. \quad (297)$$

Metoda Δ^2

Dla metod iteracyjnych rzędu 1 mamy dla dużych wartości k

$$\frac{\alpha - x_{k+2}}{\alpha - x_{k+1}} \approx \frac{\alpha - x_{k+1}}{\alpha - x_k} \approx C. \quad (298)$$

Stąd wyznaczamy

$$\alpha \approx \frac{x_k x_{k+2} - x_{k+1}^2}{x_{k+2} - 2x_{k+1} + x_k} = x_{k+2} - \frac{(\Delta x_{k+1})^2}{\Delta^2 x_k}, \quad (299)$$

gdzie Δ jest operatorem różnicy progresywnej:

$$\Delta x_k = \Delta^1 x_k = x_{k+1} - x_k, \quad \Delta^{i+1} x_k = \Delta^i x_{k+1} - \Delta^i x_k \quad (300)$$

Uwagi:

- proces Δ^2 Aitkena
- znacznie lepsze przybliżenie niż przez x_k
- tylko dla liniowej zbieżności

Minima funkcji jednej zmiennej

Sposoby znajdowania:

- szukanie rozwiązań równania $f'(x) = 0$
- aproksymacja wielomianem
- metody podziału

Metody podziału

Lemat 29 Niech $f : R \rightarrow R$ będzie **unimodalną** na $[a, b]$ tj. ma minimum w punkcie $\alpha \in [a, b]$, oraz f maleje na $[a, \alpha]$ oraz rośnie na $[\alpha, b]$. Aby zlokalizować punkt α w przedziale $[a', b']$ o mniejszej długości niż przedział $[a, b]$, wystarczy obliczyć wartość funkcji w dwóch punktach wewnątrz $[a, b]$.

Dowód: Niech $a < t_1 < t_2 < b$. Jeśli $f(t_1) < f(t_2)$, to $\alpha \in [a, t_2]$. W przec. przyp. $\alpha \in [t_1, b]$. □

Metody podziału polegają na wyznaczaniu ciągów przedziałów zstępujących $[a_i, b_i]$, $i = 0, 1, \dots$, $a_0 = a, b_0 = b$ oraz $b_i - a_i \xrightarrow{n \rightarrow \infty} 0$.

Metoda podziału na trzy części

$$t_1^i = \frac{2}{3}a_i + \frac{1}{3}b_i, \quad t_2^i = \frac{1}{3}a_i + \frac{2}{3}b_i \quad (301)$$

$$b_i - a_i = \left(\frac{2}{3}\right)^i (b - a) \quad (302)$$

W każdej iteracji liczymy 2 wartości funkcji.

Metoda połowienia

$$t_1^i = \frac{3}{4}a_i + \frac{1}{4}b_i, \quad t_2^i = \frac{1}{2}a_i + \frac{1}{2}b_i, \quad t_3^i = \frac{1}{4}a_i + \frac{3}{4}b_i \quad (303)$$

$$b_i - a_i = \left(\frac{1}{2}\right)^i (b - a) \quad (304)$$

W każdej iteracji liczymy 3 lub 2 wartości funkcji.

Metoda optymalnych podziałów – Johnsona

Cel: Minimalizacja liczby wyliczeń wartości funkcji.

Wykorzystujemy liczby Fibonacciego:

$$F_0 = F_1 = 1, \quad F_i = F_{i-1} + F_{i-2}, \quad i = 2, 3, \dots \quad (305)$$

Algorytm: Dla zadanej dokładności ρ

- $c = \frac{b-a}{\rho}$, $a_0 = a$, $b_0 = b$
- Znajdujemy N takie, że $F_{N-1} < c \leq F_N$
- Dla $i = 1, \dots, N-2$:

$$t_1^i = a_{i-1} + \frac{F_{N-i-1}}{F_{N-i+1}}(b_{i-1} - a_{i-1}), \quad t_2^i = a_{i-1} + \frac{F_{N-i}}{F_{N-i+1}}(b_{i-1} - a_{i-1}) \quad (306)$$

Jeżeli $f(t_1^i) \leq f(t_2^i)$, to $a_i \leftarrow a$, $b_i \leftarrow t_2^i$. W przec. przyp. $a_i \leftarrow t_1^i$, $b_i \leftarrow b$

Koszty: Liczymy wartości funkcji w $N-1$ punktach otrzymując ostatecznie

$$b_{N-2} - a_{N-2} = \frac{F_{N-1}}{F_N} \frac{F_{N-2}}{F_{N-1}} \dots \frac{F_2}{F_3} (b - a) = 2 \frac{b - a}{F_N} \leq 2\rho \quad (307)$$

Metoda złotego podziału

Cel:

- Przedział zmniejsza długość w sposób stały.
- Do kolejnych podziałów wykorzystuje się jeden z punktów z poprzedniego kroku.

t_1^i oraz t_2^i wybieramy tak aby:

$$\begin{aligned}t_2^i - a_i &= b_i - t_1^i = \tau(b_i - a_i), & \tau &\in (0, 1) \\ b_i - t_2^i &= \tau(b_i - t_1^i)\end{aligned}\tag{308}$$

Zatem: $\tau^2 + \tau - 1 = 0$ czyli $\tau = \frac{\sqrt{5}-1}{2} \approx 0,62$ – stosunek boków „złotego” prostokąta.

Algorytm:

- $a_0 \leftarrow a, \quad b_0 \leftarrow b, \quad t_1^1 = a + (1 - \tau)(b - a), \quad t_2^1 = b - (1 - \tau)(b - a)$
- Dla kolejnych i :
Jeżeli $f(t_1^i) < f(t_2^i)$, to

$$\begin{aligned} a_{i+1} &\leftarrow a_i, & b_{i+1} &\leftarrow t_2^i \\ t_2^{i+1} &\leftarrow t_1^i, & t_1^{i+1} &\leftarrow a_i + (1 - \tau)(b_i - a_i) \end{aligned} \tag{309}$$

W przeciwnym przypadku:

$$\begin{aligned} a_{i+1} &\leftarrow t_1^i, & b_{i+1} &\leftarrow b \\ t_1^{i+1} &\leftarrow t_2^i, & t_2^{i+1} &\leftarrow b_i - (1 - \tau)(b_i - a_i) \end{aligned} \tag{310}$$

Porównanie metod Johnsona i złotego podziału

Przy założeniu K obliczeń funkcji:

Metoda Johnsona:

$$N = K + 1, \quad |t - \alpha| \leq \frac{b - a}{F_{K+1}} \quad (311)$$

Metoda złotego podziału:

$$|t - \alpha| \leq \frac{b - a}{2G_{K-1}}, \quad \text{gdzie } G_i = \frac{1}{\tau^i}. \quad (312)$$

Porównanie:

$$2G_{K-1} < F_{K+1} < 2G_K \quad (313)$$

Układy równań liniowych

Zadanie: rozwiązać układ

$$\mathbf{Ax} = \mathbf{b}, \quad \mathbf{A} = \begin{bmatrix} a_{11} & \dots & a_{1n} \\ \vdots & & \vdots \\ a_{n1} & \dots & a_{nn} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} \quad (314)$$

Zadanie równoważne znalezieniu macierzy odwrotnej $\mathbf{A}^{-1}$, bo

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \quad (315)$$

Z drugiej strony: i -ta kolumna $\bar{a}_i$ macierzy $\mathbf{A}^{-1}$ jest rozwiązaniem układu

$$\mathbf{Ax} = e_i \quad (316)$$

Eliminacja Gaussa

Układ $\mathbf{Ax} = \mathbf{b}$ przekształcamy do

$$\mathbf{Rx} = \mathbf{c}, \quad \mathbf{R} = \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}, \quad \mathbf{c} = \begin{bmatrix} c_1 \\ \vdots \\ c_n \end{bmatrix}, \quad (317)$$

tak by rozwiązanie wyznaczyć jako:

$$x_i = \frac{c_i - \sum_{k=i+1}^n r_{ik}x_k}{r_{ii}}, \quad i = n, n-1, \dots, 1. \quad (318)$$

Pierwszy krok: Sprowadzić macierz $(\mathbf{A}, \mathbf{b})$ do macierzy $(\mathbf{A}', \mathbf{b}')$:

$$(\mathbf{A}, \mathbf{b}) = \begin{bmatrix} a_{11} & \dots & a_{1n} & b_1 \\ \vdots & & \vdots & \vdots \\ a_{n1} & \dots & a_{nn} & b_n \end{bmatrix}, \quad (\mathbf{A}', \mathbf{b}') = \begin{bmatrix} a'_{11} & a'_{12} & \dots & a'_{1n} & b'_1 \\ 0 & a'_{22} & \dots & a'_{2n} & b'_2 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & a'_{n2} & \dots & a'_{nn} & b'_n \end{bmatrix} \quad (319)$$

tak, aby układy $\mathbf{A}\mathbf{x} = \mathbf{b}$ i $\mathbf{A}'\mathbf{x} = \mathbf{b}'$ miały te same rozwiązania.

Jeśli $a_{11} \neq 0$, to wystarczy wyliczyć

$$a'_{ij} = a_{ij} - a_{1j} \frac{a_{i1}}{a_{11}}, \quad b'_i = b_i - b_1 \frac{a_{i1}}{a_{11}}, \quad i = 2, \dots, n, \quad j = 1, \dots, n. \quad (320)$$

W praktyce tzw. *wybór elementu głównego* o maksymalnym module:

- w kolumnie – konieczna zamiana odpowiednich wierszy
- w całej macierzy – konieczna zamiana odpowiednich wierszy i kolumn

Rozwiązanie pozostaje niezmiennione z dokładnością do permutacji zmiennych.

Kolejne kroki: To samo co w pierwszym dla macierzy $(\mathbf{A}', \mathbf{b}')$ z poprzedniego kroku, bez pierwszego wiersza i pierwszej kolumny.

Eliminacja Gaussa w postaci macierzowej:

$$(\mathbf{A}, \mathbf{b}) = (\mathbf{A}^0, \mathbf{b}^0) \rightarrow (\mathbf{A}^1, \mathbf{b}^1) \rightarrow \dots \rightarrow (\mathbf{A}^{n-1}, \mathbf{b}^{n-1}) = (\mathbf{R}, \mathbf{c}), \quad (321)$$

gdzie

$$\begin{aligned} (\mathbf{A}^j, \mathbf{b}^j) &= \mathbf{G}_j \mathbf{P}_j (\mathbf{A}^{j-1}, \mathbf{b}^{j-1}) \\ (\mathbf{R}, \mathbf{c}) &= \mathbf{G}_{n-1} \mathbf{P}_{n-1} \dots \mathbf{G}_1 \mathbf{P}_1 (\mathbf{A}, \mathbf{b}), \end{aligned} \quad (322)$$

$$G_j = \begin{bmatrix} 1 & & & & 0 \\ & \ddots & & & \\ & & 1 & & \\ & & -l_{j+1,j} & 1 & \\ & & \vdots & & \ddots \\ 0 & & -l_{n,j} & 0 & 1 \end{bmatrix}, \quad P_j = \begin{bmatrix} 1 & & & & \\ & \ddots & & & \\ & & 0 & & 1 \\ & & & \ddots & \\ & & 1 & & 0 \\ & & & \ddots & \\ & & & & \ddots & \\ & & & & & 1 \end{bmatrix} \quad (323)$$

Rozkład na iloczyn macierzy trójkątnych

$$\mathbf{T}_0 = (\mathbf{A}, \mathbf{b}).$$

Dla $j = 1, \dots, n$:

$$\mathbf{T}^j = \left[\begin{array}{cccc|cccc} r_{11} & r_{12} & \dots & r_{1j} & r_{1,j+1} & \dots & r_{1n} & c_1 \\ \lambda_{21} & r_{22} & \dots & r_{2j} & \vdots & & \vdots & \vdots \\ \lambda_{31} & \lambda_{32} & & \vdots & \vdots & & \vdots & \vdots \\ \vdots & \vdots & & r_{jj} & r_{j,j+1} & \dots & r_{jn} & c_j \\ \vdots & \vdots & & \lambda_{j+1,j} & a_{j+1,j+1}^j & \dots & a_{j+1,n}^j & b_{j+1}^j \\ \vdots & \vdots & & \vdots & \vdots & & \vdots & \vdots \\ \lambda_{n,1} & \lambda_{n2} & \dots & \lambda_{nj} & a_{n,j+1}^j & \dots & a_{nn}^j & b_n^j \end{array} \right]. \quad (324)$$

Kolumny z λ_{ij} są permutacją odpowiednich $-l_{ij}$.

Otrzymujemy:

$$\mathbf{LR} = \mathbf{P}_{n-1}\mathbf{P}_{n-2} \dots \mathbf{P}_1\mathbf{A} = \mathbf{PA}, \quad (325)$$

gdzie

$$\mathbf{L} = \begin{bmatrix} 1 & & & 0 \\ t_{21}^n & \ddots & & \\ \vdots & & \ddots & \\ t_{n1}^n & \dots & t_{n,n-1}^n & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} t_{11}^n & \dots & t_{1n}^n \\ & \ddots & \\ 0 & & t_{nn}^n \end{bmatrix}. \quad (326)$$

Uwagi:

- Rozwiązanie układu $\mathbf{Ax} = \mathbf{b}$: ponieważ $\mathbf{PAx} = \mathbf{LRx} = \mathbf{Pb}$, więc x można wyznaczyć w dwóch krokach:

$$\mathbf{Lu} = \mathbf{Pb}, \quad \mathbf{Rx} = \mathbf{u}. \quad (327)$$

- Nie każda nieosobliwa macierz ma rozkład trójkątny $\mathbf{A} = \mathbf{LR}$.
- Złożoność $O(n^3)$.

Schematy zwarte eliminacji Gaussa

- Głównie dla celu obliczeń ręcznych – mniej wyników pośrednich.
- Eliminacja Gaussa w k -tym kroku wyznacza k -tą kolumnę $\mathbf{L}$ i k -ty wiersz $\mathbf{R}$.
- $\mathbf{A} = \mathbf{LR}$ jest równoważne układowi n^2 równań z $n(n+1)$ niewiadomymi:

$$a_{ij} = \sum_{p=1}^{\min(i,j)} l_{ip} r_{pj}. \quad (328)$$

W k -tym kroku:

$$a_{kj} = \sum_{p=1}^k l_{kp} r_{pj}, \quad j \geq k, \quad a_{ik} = \sum_{p=1}^k l_{ip} r_{pk}, \quad i > k. \quad (329)$$

Metoda Doolittle'a: $l_{kk} = 1, \quad k = 1, \dots, n$

$$r_{kj} = a_{kj} - \sum_{p=1}^{k-1} l_{kp} r_{pj}, \quad j = k, \dots, n,$$

$$l_{ik} = \frac{a_{ik} - \sum_{p=1}^{k-1} l_{ip} r_{pk}}{r_{kk}}, \quad i = k+1, \dots, n.$$
(330)

Metoda Crouta: $r_{kk} = 1, \quad k = 1, \dots, n$

$$l_{ik} = a_{ik} - \sum_{p=1}^{k-1} l_{ip} r_{pk}, \quad i = k, \dots, n,$$

$$r_{kj} = \frac{a_{kj} - \sum_{p=1}^{k-1} l_{kp} r_{pj}}{l_{kk}}, \quad j = k+1, \dots, n.$$
(331)

Algorytm Gaussa–Jordana

Cel: Wyznaczenie macierzy odwrotnej $\mathbf{A}^{-1}$. Macierz $\mathbf{A}$ realizuje odwzorowanie:

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= y_1 \\ &\vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= y_n \end{aligned} \tag{332}$$

Wyznaczamy odwzorowanie odwrotne poprzez zamiany zmiennych. Zmienną x_1 zamieniamy na takie y_r , dla którego

$$|a_{r1}| = \max_i |a_{i1}|. \tag{333}$$

Jeśli $a_{r1} = 0$, to macierz $\mathbf{A}$ jest osobliwa.

Zamieniamy miejscami równania 1 i r otrzymując układ $\overline{\mathbf{A}}\mathbf{x} = \overline{\mathbf{y}}$, gdzie $\overline{a}_{11} = a_{r1}$ oraz $\overline{y}_1 = y_r$.

Po zamianie zmiennych x_1 i $\bar{y}_1 = y_r$ otrzymujemy układ:

$$\begin{aligned} a'_{11}\bar{y}_1 + a'_{12}x_2 + \cdots + a'_{1n}x_n &= x_1 \\ a'_{21}\bar{y}_1 + a'_{22}x_2 + \cdots + a'_{2n}x_n &= \bar{y}_2 \\ &\vdots \\ a'_{n1}\bar{y}_1 + a'_{n2}x_2 + \cdots + a'_{nn}x_n &= \bar{y}_n \end{aligned} \quad (334)$$

gdzie dla $i, j = 2, \dots, n$

$$a'_{11} = \frac{1}{\bar{a}_{11}}, \quad a'_{1j} = -\frac{\bar{a}_{1j}}{\bar{a}_{11}}, \quad a'_{i1} = \frac{\bar{a}_{i1}}{\bar{a}_{11}}, \quad a'_{ij} = \bar{a}_{ij} - \frac{\bar{a}_{i1}\bar{a}_{1j}}{\bar{a}_{11}}. \quad (335)$$

W każdym kolejnym kroku $k = 2, \dots, n$ zamieniamy w analogiczny sposób zmienną x_k przez jedną ze zmiennych $\bar{y}_k, \dots, \bar{y}_n$.

Ogólnie: konstruujemy ciąg macierzy:

$$\mathbf{A} = \mathbf{A}^0 \rightarrow \dots \rightarrow \mathbf{A}^n = \mathbf{A}^{-1}, \quad (336)$$

gdzie dla $k = 1, \dots, n$ macierz $\mathbf{A}^k = (a'_{ij})$ powstaje z $\mathbf{A}^{k-1} = (a_{ij})$ w następujący sposób:

1. Wyznaczamy r takie, że

$$|a_{rk}| = \max_{i \geq k} |a_{ik}|. \quad (337)$$

Jeśli $a_{rk} = 0$, to macierz $\mathbf{A}$ jest osobliwa – Koniec!.

2. Zamieniamy wiersze k i r macierzy $\mathbf{A}^{k-1}$ otrzymując macierz $\overline{\mathbf{A}} = (\overline{a}_{ik})$.

3. Dla $i, j \neq k$ liczymy:

$$a'_{kk} = \frac{1}{\overline{a}_{kk}}, \quad a'_{kj} = -\frac{\overline{a}_{kj}}{\overline{a}_{kk}}, \quad a'_{ik} = \frac{\overline{a}_{ik}}{\overline{a}_{kk}}, \quad a'_{ij} = \overline{a}_{ij} - \frac{\overline{a}_{ik}\overline{a}_{kj}}{\overline{a}_{kk}}. \quad (338)$$

Rozkład Choleskiego

Definicja 13 Macierz zespoloną $\mathbf{A}$ wymiaru $n \times n$, nazywamy **dodatnio określona**, gdy:

1. $\mathbf{A} = \mathbf{A}^H$, tzn. $\mathbf{A}$ jest macierzą hermitowską.
2. Dla wszystkich $\mathbf{x} \in C^n, \mathbf{x} \neq 0$ zachodzi $\mathbf{x}^H \mathbf{A} \mathbf{x} > 0$.

Hermitowską macierz nazywamy **dodatnio półokreślona** gdy dla wszystkich $\mathbf{x} \in C^n$ zachodzi $\mathbf{x}^H \mathbf{A} \mathbf{x} \geq 0$.

Twierdzenie 30 Dla każdej dodatnio określonej macierzy $\mathbf{A}$ wymiaru $n \times n$ istnieje jednoznaczna trójkątna dolna macierz $\mathbf{L}$ wymiaru $n \times n$, spełniająca warunki $l_{kk} > 0, k = 1, \dots, n$, oraz

$$\mathbf{A} = \mathbf{L}\mathbf{L}^H. \quad (339)$$

Jeżeli $\mathbf{A}$ jest rzeczywista, to również $\mathbf{L}$ jest rzeczywista.

Równanie (339) oznacza, że dla $k = 1, \dots, n$, $i = k, \dots, n$ zachodzi:

$$a_{ik} = l_{i1}\bar{l}_{k1} + l_{i2}\bar{l}_{k2} + \dots + l_{in}\bar{l}_{kn}. \quad (340)$$

Algorytm:

Dla $k = 1, \dots, n$:

$$l_{kk} = \sqrt{a_{kk} - \sum_{p=1}^{k-1} l_{kp}^2}, \quad (341)$$

Dla $i = k + 1, \dots, n$:

$$l_{ik} = \frac{a_{ik} - \sum_{p=1}^{k-1} l_{ip}l_{kp}}{l_{kk}}. \quad (342)$$

Uwagi:

- Wykorzystujemy tylko górną trójkątną część macierzy $\mathbf{A}$.
- Złożoność $O(n^3)$.
- n razy liczymy pierwiastek
- $|l_{kj}| \leq \sqrt{a_{kk}}, \quad k = 1, \dots, n \quad j = 1, \dots, k.$

Błędy zaokrągleń w eliminacji Gaussa

Częściowy wybór elementu głównego

- zabezpiecza przed zerowym elementem głównym w macierzy nieosobliwej
- ma wpływ na błędy numeryczne

Przykład: Układ

$$\begin{bmatrix} 0.005 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \quad (343)$$

ma rozwiązanie $(\frac{5000}{9950}, \frac{4950}{9950}) \approx (0.503, 0.497)$.

Stosujemy 2-cyfrową precyzję obliczeń.

Jeśli elementem głównym jest a_{11} , to mamy:

$$\begin{bmatrix} 0.005 & 1 \\ 0 & -200 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0.5 \\ -99 \end{bmatrix}, \quad (x, y) = (0, 0.5). \quad (344)$$

Jeśli wybierzemy a_{21} jako element główny, to otrzymamy:

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 0.5 \end{bmatrix}, \quad (x, y) = (0.5, 0.5). \quad (345)$$

Częściowy wybór elementu głównego to nie wszystko:
Z równoważnym do (343) układem

$$\begin{bmatrix} 1 & 200 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 100 \\ 1 \end{bmatrix} \quad (346)$$

są te same problemy, choć $a_{11} = a_{21}$.

Skalowanie macierzy: Przemnożenie pierwszego wiersza przez 200, to zastąpienie macierzy $\mathbf{A}$ przez $\tilde{\mathbf{A}} = \mathbf{D}\mathbf{A}$ (i równocześnie $\mathbf{b}$ przez $\tilde{\mathbf{b}} = \mathbf{D}\mathbf{b}$), gdzie

$$\mathbf{D} = \begin{bmatrix} 200 & 0 \\ 0 & 1 \end{bmatrix}. \quad (347)$$

Skalować można zarówno wiersze jak i kolumny, wówczas otrzymujemy układ:

$$\mathbf{D}_1 \mathbf{A} \mathbf{D}_2 (\mathbf{D}_2^{-1} \mathbf{x}) = \mathbf{D}_1 \mathbf{A} \mathbf{D}_2 \mathbf{y} = \mathbf{D}_1 \mathbf{b}, \quad (348)$$

gdzie macierze $\mathbf{D}_1$ i $\mathbf{D}_2$ są diagonalne.

Nie ma sposobu skalowania zapewniającego numeryczną stabilność dla dowolnej macierzy.

Heurystyka: macierze *zrównoważone*, tzn o zbliżonych sumach wartości bezwzględnych elementów poszczególnych wierszy.

Najczęstsze skalowanie:

$$\mathbf{D}_2 = \mathbf{I}, \quad \mathbf{D}_1 = \text{diag}(s_1, \dots, s_n), \quad s_i = \frac{1}{\sum_{k=1}^n |a_{ik}|}. \quad (349)$$

Alternatywa: modyfikacja zasady wyboru elementu głównego w kolumnie (skalowanie przez s_i).

Błędy zaokrągleń dla rozkładu trójkątnego

Niech $\mathbf{A} = \mathbf{LR}$ będzie rozkładem trójkątnym macierzy $\mathbf{A}$ wymiaru $n \times n$. $\mathbf{L}$ i $\mathbf{R}$ liczymy np. wzorami (330).

Wystarczy ocenić wyrażenia typu:

$$b_n = fl \left(\frac{c - a_1 b_1 - \dots - a_{n-1} b_{n-1}}{a_n} \right). \quad (350)$$

Analiza rozchodzenia się błędów (str. 17) pozwala oszacować:

$$\left| c - \sum_{j=1}^n a_j b_j \right| \leq \frac{eps}{1 - eps} \left[\delta |s_{n-1}| + \sum_{j=1}^{n-1} (|s_j| + |a_j b_j|) \right], \quad (351)$$

gdzie

$$s_0 = c, \quad s_j = fl(s_{j-1} - a_j b_j), \quad \delta = \begin{cases} 0 & \text{jeśli } a_n = 1 \\ 1 & \text{w przec. przyp.} \end{cases} \quad (352)$$

Otrzymujemy dla $\mathbf{F} = (f_{ij}) = \mathbf{A} - \bar{\mathbf{L}}\bar{\mathbf{R}}$ oszacowania:

$$\begin{aligned}
 |f_{ij}| = |a_{ij} - \sum_{k=1}^i \bar{l}_{ik} \bar{r}_{kj}| &\leq \frac{eps}{1-eps} \sum_{k=1}^{i-1} (|\bar{a}_{ij}^k| + |\bar{l}_{ik}| |\bar{r}_{kj}|) && \text{dla } j \geq i \\
 |f_{ij}| = |a_{ij} - \sum_{k=1}^j \bar{l}_{ik} \bar{r}_{kj}| &\leq \frac{eps}{1-eps} \left[|\bar{a}_{ij}^{j-1}| + \sum_{k=1}^{j-1} (|\bar{a}_{ij}^k| + |\bar{l}_{ik}| |\bar{r}_{kj}|) \right] && \text{dla } j < i,
 \end{aligned}
 \tag{353}$$

gdzie $\bar{a}_{ij}^k = fl(a_{ij} - \sum_{s=1}^k \bar{l}_{is} \bar{r}_{sj})$.

Przyjmując $a_k = \max_{i,j} |a_{ij}^k|$, $a = \max_{0 < i < n-1} a_i$ oraz uwzględniając, że $\bar{r}_{ij} = \bar{a}_{ij}^{i-1}$, dla $i \leq j$ i zakładając, że $|\bar{l}_{ij}| \leq 1$ szacujemy:

$$\begin{aligned}
 |f_{ij}| &\leq \frac{eps}{1-eps} (a_0 + 2a_1 + \cdots + 2a_{i-2} + a_{i-1}) \leq 2(i-1)a \frac{eps}{1-eps} && \text{dla } j \geq i \\
 |f_{ij}| &\leq \frac{eps}{1-eps} (a_0 + 2a_1 + \cdots + 2a_{j-2} + 2a_{j-1}) \leq 2ja \frac{eps}{1-eps} && \text{dla } j < i
 \end{aligned}
 \tag{354}$$

$$|f_{ij}| \leq 2a \frac{eps}{1 - eps} \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & 0 \\ 1 & 1 & 1 & \dots & 1 & 1 \\ 1 & 2 & 2 & \dots & 2 & 2 \\ 1 & 2 & 3 & \dots & 3 & 3 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ 1 & 2 & 3 & \dots & n-1 & n-1 \end{bmatrix} \quad (355)$$

Oszacowanie wartości a

Z definicji $a_0 = \max_{i,j} a_{ij}$. Dla częściowego wyboru elementu głównego można wykazać, że $a_{k-1} \leq 2^k a_0$. Mamy więc (zwykle pesymistyczne, ale czasem osiągalne) oszacowanie

$$a \leq 2^{n-1} a_0. \quad (356)$$

Dla górnej macierzy Hessenberga ($a_{ij} = 0$ dla $i > j + 1$) można wykazać, że

$$a \leq (n-1)a_0. \quad (357)$$

Dla macierzy trójdzielnej

$$a \leq 2a_0. \quad (358)$$

Pełny wybór elementu głównego

- Wilkinson wykazał, że

$$a_k \leq f(k+1)a_0, \quad (359)$$

gdzie

$$f(k) = \sqrt{k 2^1 3^{\frac{1}{2}} 4^{\frac{1}{3}} \dots k^{\frac{1}{k-1}}}. \quad (360)$$

- W praktyce nie znaleziono kontrprzykładu dla oszacowania

$$a_k \leq (k+1)a_0. \quad (361)$$

- Bardziej kosztowny niż częściowy.
- Niszczy szczególną strukturę macierzy (np. trójdagonalnych).

Błędy zaokrągleń dla układów o trójkątnych macierzach

Rozwiązanie układu $\mathbf{Ax} = \mathbf{b}$ sprowadziliśmy do dwóch układów trójkątnych: $\bar{\mathbf{L}}\mathbf{y} = \mathbf{b}$ oraz $\bar{\mathbf{R}}\mathbf{x} = \mathbf{y}$.

Korzystamy z oszacowania:

$$\left| c - \sum_{i=1}^n a_i b_i \right| \leq \frac{eps}{1 - n \cdot eps} \left[\sum_{i=1}^{n-1} i |a_i b_i| + (n + 1 + \delta) |a_n b_n| \right]. \quad (362)$$

Dla układu $\bar{\mathbf{L}}\mathbf{y} = \mathbf{b}$ dostajemy:

$$\left| b_r - \sum_{i=1}^r \bar{l}_{ri} \bar{y}_i \right| \leq \frac{eps}{1 - n \cdot eps} \left[\sum_{i=1}^r i |\bar{l}_{ri} \bar{y}_i| + |\bar{y}_r| \right], \quad (363)$$

czyli

$$|\mathbf{b} - \bar{\mathbf{L}}\bar{\mathbf{y}}| \leq \frac{eps}{1 - n \cdot eps} (|\bar{\mathbf{L}}| \mathbf{D} - \mathbf{I}) |\bar{\mathbf{y}}|, \quad \mathbf{D} = \begin{bmatrix} 1 & & & 0 \\ & 2 & & \\ & & \ddots & \\ 0 & & & n \end{bmatrix}. \quad (364)$$

A zatem $\bar{\mathbf{y}}$ jest rozwiązaniem dokładnym nieco zaburzonego układu:

$$(\bar{\mathbf{L}} + \Delta\bar{\mathbf{L}})\bar{\mathbf{y}} = \mathbf{b}, \quad |\Delta\bar{\mathbf{L}}| \leq \frac{eps}{1 - n \cdot eps} (|\bar{\mathbf{L}}| \mathbf{D} - \mathbf{I}). \quad (365)$$

Podobne wnioskowanie można przeprowadzić dla układu $\mathbf{R}\mathbf{x} = \mathbf{y}$, co ostatecznie prowadzi do wniosku, że liczenie rozwiązania układu metodą eliminacji Gaussa jest **algorytmem numerycznie stabilnym**.

Normy macierzy

Definicja 14 *Normą macierzy* A jest liczba $\|A\|$, gdy $\|\cdot\|$ spełnia następujące warunki:

- 1° $\|A\| \geq 0$ oraz $\|A\| = 0$ wtw. gdy $A = 0$,
- 2° $\|cA\| = |c| \|A\|$ dla dowolnej $c \in R$,
- 3° $\|A + B\| \leq \|A\| + \|B\|$ (nierówność trójkąta),
- 4° $\|AB\| \leq \|A\| \|B\|$ (nierówność Schwarza).

Z 4° wynika, że

$$\|A^m\| \leq \|A\|^m. \quad (366)$$

Definicja 15 *Norma euklidesowa:*

$$\|A\|_E = \sqrt{\sum_{i=1}^m \sum_{j=1}^n a_{ij}^2}. \quad (367)$$

Definicja 16 Norma spektralna:

$$\|A\|_S = \sqrt{\max_i |\lambda_i(AA^T)|}. \quad (368)$$

Definicja 17 Norma euklidesowa wektora $\mathbf{w}$:

$$\|\mathbf{w}\| = \sqrt{\mathbf{w}^T \mathbf{w}}. \quad (369)$$

Definicja 18 Promień spektralny macierzy:

$$\varrho(A) = \max_i |\lambda_i(A)|. \quad (370)$$

Ważne własności:

Dla macierzy kwadratowych:

$$\|A\|_E^2 = \text{Tr}(AA^T) = \sum_{i=1}^n \lambda_i(AA^T). \quad (371)$$

Stąd

$$\|A\|_S \leq \|A\|_E. \quad (372)$$

Dla macierzy symetrycznych

$$\|A\|_S = \varrho(A). \quad (373)$$

Twierdzenie 31 *Dla macierzy kwadratowych A stopnia n*

$$\max_{\mathbf{x} \neq 0} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|} = \|A\|_S. \quad (374)$$

Metody iteracyjne rozwiązywania układów równań liniowych

- Zwykle dokładność porównywalna z metodami dokładnymi uzyskiwana jest w dłuższym czasie.
- Dla macierzy rzadkich, metody iteracyjne mogą być znacznie szybsze.
- Mniejsza złożoność pamięciowa.

Jednopunktowy liniowy wzór iteracyjny dla układu równań:

$$\mathbf{x}_{i+1} = B_i \mathbf{x}_i + \mathbf{c}_i. \quad (375)$$

Rozwiązujemy układ równań $A\mathbf{x} = \mathbf{b}$. Inaczej

$$(I + A)\mathbf{x} = \mathbf{x} + \mathbf{b}. \quad (376)$$

Narzuca się wzór iteracyjny

$$\mathbf{x}_{i+1} = (I + A)\mathbf{x}_i - \mathbf{b}. \quad (377)$$

Chcemy, aby rozwiązanie dokładne $\mathbf{x}_d$ było punktem stałym przekształcenia (375):

$$\mathbf{x}_d = B_i \mathbf{x}_d + \mathbf{c}_i. \quad (378)$$

Skoro

$$\mathbf{x}_d = A^{-1} \mathbf{b}, \quad (379)$$

więc dostajemy

$$A^{-1} \mathbf{b} = B_i A^{-1} \mathbf{b} + \mathbf{c}_i. \quad (380)$$

Zatem:

$$\mathbf{c}_i = (I - B_i) A^{-1} \mathbf{b} = C_i \mathbf{b}. \quad (381)$$

Warunek zgodności dla B_i i C_i :

$$(I - B_i) A^{-1} = C_i, \quad \text{czyli} \quad B_i + C_i A = I \quad (382)$$

Częstsza postać wzoru iteracyjnego (375):

$$\mathbf{x}_{i+1} = B_i \mathbf{x}_i + C_i \mathbf{b}. \quad (383)$$

Analiza zbieżności wzoru (383)

Wprowadźmy:

$$\varepsilon_i = \mathbf{x}_i - \mathbf{x}_d. \quad (384)$$

Zatem:

$$\begin{aligned} \varepsilon_{i+1} &= B_i \mathbf{x}_i + C_i \mathbf{b} - \mathbf{x}_d = B_i \mathbf{x}_i + C_i \mathbf{b} - A^{-1} \mathbf{b} = \\ &= B_i \mathbf{x}_i - B_i A^{-1} \mathbf{b} = B_i \varepsilon_i. \end{aligned} \quad (385)$$

A więc:

$$\varepsilon_{i+1} = K_i \varepsilon_1, \quad \text{gdzie } K_i = B_i B_{i-1} \dots B_1. \quad (386)$$

Warunek konieczny i dostateczny zbieżności $\{\mathbf{x}_i\}$ do $\mathbf{x}_d$:

$$\forall_{\mathbf{y}} \lim_{i \rightarrow \infty} K_i \mathbf{y} = 0. \quad (387)$$

Warunek dostateczny:

$$\lim_{i \rightarrow \infty} \|\varepsilon_i\| = 0. \quad (388)$$

Z twierdzenia 31 mamy

$$\max_{\varepsilon_1} \frac{\|\varepsilon_{i+1}\|}{\|\varepsilon_1\|} = \max_{\varepsilon_1} \frac{\|K_i \varepsilon_1\|}{\|\varepsilon_1\|} = \|K_i\|_S. \quad (389)$$

Warunkiem dostatecznym jest więc

$$\lim_{i \rightarrow \infty} \|K_i\|_S = 0. \quad (390)$$

Dla metod stacjonarnych mamy również warunek dostateczny

$$\lim_{i \rightarrow \infty} \varrho(K_i) = 0. \quad (391)$$

Stacjonarne metody iteracyjne

W (383) $B_i = B$, $C_i = C$, oraz

$$K_i = B_i B_{i-1} \dots B_1 = B^i. \quad (392)$$

Wartości własne B^i są i -tymi potęgami wartości własnych macierzy B . Stąd gwarancją zbieżności jest by moduły wartości własnych B były mniejsze od 1 (jest to również warunek konieczny).

Z relacji (372) między normami macierzy dostajemy ostrzejszy warunek dostateczny:

$$\|B\|_E < 1. \quad (393)$$

Szybkość zbieżności:

$$\|\varepsilon_{i+1}\| \leq \|K_i\| \|\varepsilon_1\| \leq \|B\|^i \|\varepsilon_1\|. \quad (394)$$

Metoda prostej iteracji

Niech

$$A = D + L + U, \quad (395)$$

gdzie D jest diagonalna, L oraz U – trójkątnymi dolną i górną z zerami na przekątnej.

Układ $A\mathbf{x} = \mathbf{b}$ można przedstawić w postaci:

$$D\mathbf{x} = -(L + U)\mathbf{x} + \mathbf{b}. \quad (396)$$

Metoda prostej iteracji:

$$\mathbf{x}_{i+1} = -D^{-1}(L + U)\mathbf{x}_i + D^{-1}\mathbf{b}. \quad (397)$$

Uwagi:

1. Na przekątnej macierzy A nie mogą występować zera. Dla macierzy nieosobliwych można poprzestawiać kolumny i wiersze.
2. Sprawdzenie czy $B = -D^{-1}(L + U)$ gwarantuje zbieżność może być trudne.

Metoda Seidla

Modyfikacja metody prostej iteracji: po obliczeniu dowolnej składowej wektora $\mathbf{x}_{i+1}$, korzysta się z niej przy obliczaniu następnych składowych.

Oznaczając j -tą składową wektora $\mathbf{x}_i$ przez $x_i^{(j)}$ otrzymujemy:

$$x_{i+1}^{(1)} = -\frac{1}{a_{11}} \left(\sum_{j=2}^n a_{1j} x_i^{(j)} - b_1 \right), \quad (398)$$

$$x_{i+1}^{(2)} = -\frac{1}{a_{22}} \left(a_{21} x_{i+1}^{(1)} + \sum_{j=3}^n a_{2j} x_i^{(j)} - b_2 \right). \quad (399)$$

Ogólnie:

$$x_{i+1}^{(r)} = -\frac{1}{a_{rr}} \left(\sum_{j=1}^{r-1} a_{rj} x_{i+1}^{(j)} + \sum_{j=r+1}^n a_{rj} x_i^{(j)} - b_r \right). \quad (400)$$

W postaci macierzowej:

$$\mathbf{x}_{i+1} = -D^{-1}(L\mathbf{x}_{i+1} + U\mathbf{x}_i) + D^{-1}\mathbf{b}. \quad (401)$$

Stąd:

$$\mathbf{x}_{i+1} = -(D + L)^{-1}U\mathbf{x}_i + (D + L)^{-1}\mathbf{b}. \quad (402)$$

Warunek dostateczny zbieżności:

$$\|B\| = \|(D + L)^{-1}U\| < 1. \quad (403)$$

Twierdzenie 32 *Metoda Seidla jest zbieżna dla dodatnio określonych macierzy A dla dowolnego wektora początkowego.*

Uwaga: Metody iteracyjne rozwiązywania układów równań liniowych są zbieżne liniowo (p. równanie (385)). Można więc zastosować metodę Δ^2 .

Relaksacja

Idea: Rozważamy wektor reszt

$$\mathbf{r} = \mathbf{b} - A\mathbf{x}. \quad (404)$$

Modyfikujemy składowe przybliżonego rozwiązania $\mathbf{x}$ tak, aby wyzerować jedną lub więcej składowych wektora $\mathbf{r}$.

Uwaga: Metoda Seidla jest metoda relaksacyjną.

Algorytm wygodny w obliczeniach ręcznych:

1. Wybierz wektor początkowy $\mathbf{x}_1$.
2. Dla i od 1 w górę tak długo jak trzeba
 - (a) Oblicz $\mathbf{r}_i = \mathbf{b} - A\mathbf{x}_i$.
 - (b) Odszukaj składową wektora $\mathbf{r}_i$ o największym module ($r_i^{(L)}$) i element o największym module z L -tym wierszu macierzy A (a_{LJ})
 - (c) Wyznacz wektor $\mathbf{x}_{i+1}$ mający wszystkie składowe identyczne ze składowymi wektora $\mathbf{x}_i$, z wyjątkiem

$$x_{i+1}^{(J)} = x_i^{(J)} + \frac{r_i^{(L)}}{a_{LJ}} \quad (405)$$

Uwagi: $\mathbf{r}_{i+1}$ lepiej wyznaczać jako

$$\mathbf{r}_i + (x_i^{(J)} - x_{i+1}^{(J)})A_L = \mathbf{r}_i - \frac{r_i^{(L)}}{a_{LJ}}A_L, \quad \text{gdzie } A_L \text{ to } L\text{-ta kolumna } A. \quad (406)$$

Jeśli elementy maksymalne w wierszu są na głównej przekątnej, to $J = L$.

Metoda nadrelaksacji

Szybkość zbieżności zależy od promienia spektralnego macierzy B – zmniejszenie promienia spektralnego oznacza przyspieszenie zbieżności.

Przekształcamy układ $A\mathbf{x} = \mathbf{b}$ tak, by na przekątnej macierzy A dostać same jedynki. Metoda Seidla (400) przybiera postać:

$$\begin{aligned} x_{i+1}^{(r)} &= - \sum_{j=1}^{r-1} a_{rj} x_{i+1}^{(j)} - \sum_{j=r+1}^n a_{rj} x_i^{(j)} + b_r \\ &= x_i^{(r)} - \sum_{j=1}^{r-1} a_{rj} x_{i+1}^{(j)} - \sum_{j=r}^n a_{rj} x_i^{(j)} + b_r. \end{aligned} \quad (407)$$

Rozważamy

$$x_{i+1}^{(r)} = x_i^{(r)} - \omega \left(\sum_{j=1}^{r-1} a_{rj} x_{i+1}^{(j)} + \sum_{j=r}^n a_{rj} x_i^{(j)} - b_r \right), \quad (408)$$

gdzie ω jest czynnikiem nadrelaksacji.

W postaci macierzowej:

$$\mathbf{x}_{i+1} = B_\omega \mathbf{x}_i + C_\omega \mathbf{b}, \quad (409)$$

gdzie

$$B_\omega = (I + \omega L)^{-1}((1 - \omega)I - \omega U), \quad C_\omega = \omega(I + \omega L)^{-1}. \quad (410)$$

Problem: Jak dobrać ω , by promień $\varrho(B_\omega)$ był minimalny?

Ogólnie: trudne zadanie.

Czasami: np. prosta zależność między optymalnym ω a promieniem $\varrho(L + U)$.

Metody iteracyjne wynikające z minimalizacji formy kwadratowej

Cel: Rozwiązanie układu $Ax = b$ poprzez minimalizację formy kwadratowej.

Jeśli A nie jest symetryczna i dodatnio określona, można minimalizować

$$R = \mathbf{r}^T \mathbf{r} = (\mathbf{b} - A\mathbf{x})^T (\mathbf{b} - A\mathbf{x}). \quad (411)$$

Zwykle ogromna ilość obliczeń.

Jeśli A jest symetryczna i dodatnio określona, można minimalizować

$$Q = \frac{1}{2} \mathbf{x}^T A \mathbf{x} - \mathbf{x}^T \mathbf{b}. \quad (412)$$

Minimum w rozwiązaniu $Ax = b$ bo:

$$\begin{aligned} Q(\mathbf{x} + \Delta \mathbf{x}) - Q(\mathbf{x}) &= \frac{1}{2} (\mathbf{x} + \Delta \mathbf{x})^T A (\mathbf{x} + \Delta \mathbf{x}) - (\mathbf{x} + \Delta \mathbf{x})^T \mathbf{b} - \frac{1}{2} \mathbf{x}^T A \mathbf{x} + \mathbf{x}^T \mathbf{b} \\ &= \Delta \mathbf{x}^T A \mathbf{x} - \Delta \mathbf{x}^T \mathbf{b} + \frac{1}{2} \Delta \mathbf{x}^T A \Delta \mathbf{x} = \frac{1}{2} \Delta \mathbf{x}^T A \Delta \mathbf{x}. \end{aligned}$$

Ogólna metoda minimalizacji

W kolejnym kroku mając wektor $\mathbf{x}_i$, wybieramy kierunek $\mathbf{v}_i$ i odległość α_i :

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{v}_i. \quad (413)$$

Metoda największego spadku

Kierunek $\mathbf{v}_i$ wybieramy jako kierunek najszybszego spadku formy Q czyli kierunek gradientu w punkcie $\mathbf{x}_i$:

$$\mathbf{v}_i \leftarrow \nabla Q = A\mathbf{x}_i - \mathbf{b} = -\mathbf{r}_i. \quad (414)$$

Aby wyznaczyć α_i liczymy:

$$Q(\mathbf{x}_i - \alpha_i \mathbf{r}_i) = -\frac{1}{2} \mathbf{x}_i^T \mathbf{r}_i + \alpha_i \mathbf{r}_i^T \mathbf{r}_i + \frac{1}{2} \alpha_i^2 \mathbf{r}_i^T A \mathbf{r}_i - \frac{1}{2} \mathbf{x}_i^T \mathbf{b}. \quad (415)$$

Zatem

$$\frac{\delta Q}{\delta \alpha_i} = \mathbf{r}_i^T \mathbf{r}_i + \alpha_i \mathbf{r}_i^T A \mathbf{r}_i, \quad (416)$$

czyli minimum w punkcie

$$\alpha_i = -\frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{r}_i^T A \mathbf{r}_i}. \quad (417)$$

Ostatecznie

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \frac{\mathbf{r}_i^T \mathbf{r}_i}{\mathbf{r}_i^T A \mathbf{r}_i} \mathbf{r}_i. \quad (418)$$

Uwagi:

- Zbieżność zapewniona, bo $\mathbf{Q}(\mathbf{x}_{i+1}) < \mathbf{Q}(\mathbf{x}_i)$.
- Zbieżność może być bardzo wolna dla układów źle uwarunkowanych.

Metoda sprzężonych gradientów

Idea: Odpowiedniejszy dobór kolejnych kierunków.

Niech $\mathbf{v}_1, \dots, \mathbf{v}_n$ będą bazą n wymiarowej przestrzeni euklidesowej i niech $\mathbf{x}$ będzie przybliżeniem rozwiązania dokładnego $\mathbf{x}_d$. Wtedy:

$$\mathbf{x}_d - \mathbf{x} = \sum_{j=1}^n \alpha_j \mathbf{v}_j. \quad (419)$$

Jeśli wektory $\mathbf{v}_j$ są ortogonalne, to

$$\alpha_j = \frac{\mathbf{v}_j^T (\mathbf{x}_d - \mathbf{x})}{\mathbf{v}_j^T \mathbf{v}_j}, \quad j = 1, \dots, n. \quad (420)$$

Jeśli wektory $\mathbf{v}_j$ są **A -ortogonalne** (lub **A -sprzężone**), tzn.

$$\mathbf{v}_j^T A \mathbf{v}_i = 0 \quad \text{dla } i \neq j, \quad (421)$$

to

$$\alpha_j = \frac{\mathbf{v}_j^T A (\mathbf{x}_d - \mathbf{x})}{\mathbf{v}_j^T A \mathbf{v}_j} = \frac{\mathbf{v}_j^T \mathbf{r}}{\mathbf{v}_j^T A \mathbf{v}_j}, \quad j = 1, \dots, n. \quad (422)$$

Metoda sprzężonych gradientów polega na wyznaczaniu wektorów $\mathbf{v}_j$ i współczynników α_j , by ostatecznie wyznaczyć $\mathbf{x}_d$.

Konstrukcja wektorów A -ortogonalnych: założmy, że mamy bazę wektorów $\mathbf{u}_1, \dots, \mathbf{u}_n$. Stosujemy metodę ortogonalizacji Grama-Schmidta:

$$\mathbf{v}_1 = \mathbf{u}_1, \quad \mathbf{v}_{i+1} = \mathbf{u}_{i+1} + \sum_{k=1}^i \beta_{i+1,k} \mathbf{v}_k, \quad i = 1, \dots, n, \quad (423)$$

gdzie

$$\beta_{i+1,k} = -\frac{\mathbf{v}_k^T A \mathbf{u}_{i+1}}{\mathbf{v}_k^T A \mathbf{v}_k}. \quad (424)$$

Pozostaje wyznaczyć ciąg wektorów $\{\mathbf{u}_i\}$

- Wektory jednostkowe.
- Bardziej wygodny sposób: założmy, że

$$\mathbf{x}_i = \mathbf{x}_1 + \sum_{j=1}^{i-1} \alpha_j \mathbf{v}_j, \quad (425)$$

gdzie α_j spełniają (422). Zatem

$$\mathbf{x}_{i+1} = \mathbf{x}_i + \alpha_i \mathbf{v}_i, \quad (426)$$

$$\mathbf{r}_i = \mathbf{b} - A\mathbf{x}_i. \quad (427)$$

Przyjmujemy

$$\mathbf{u}_i = \mathbf{r}_i, \quad i = 1, \dots, n. \quad (428)$$

Rekurencyjne wyznaczanie $\mathbf{r}_i$, α_i oraz $\mathbf{v}_i$ definiuje metodę sprzężonych gradientów:

$$\begin{aligned}\mathbf{v}_1 &= \mathbf{r}_1 = \mathbf{b} - A\mathbf{x}_1 \\ \alpha_i &= \frac{\mathbf{v}_i^T \mathbf{r}_i}{\mathbf{v}_i^T A \mathbf{v}_i} \\ \mathbf{x}_{i+1} &= \mathbf{x}_i + \alpha_i \mathbf{v}_i \\ \mathbf{r}_{i+1} &= \mathbf{r}_i - \alpha_i A \mathbf{v}_i \\ \beta_i &= -\frac{\mathbf{v}_i^T A \mathbf{r}_{i+1}}{\mathbf{v}_i^T A \mathbf{v}_i} \\ \mathbf{v}_{i+1} &= \mathbf{r}_{i+1} + \beta_i \mathbf{v}_i\end{aligned}\quad \text{dla } i = 1, \dots, n. \quad (429)$$

Uwagi:

- Na użytek $\mathbf{v}_i$ wyznaczamy tylko jedno β_i .
- $\mathbf{x}_{n+1} = \mathbf{x}_d$ – metoda skończona, choć błędy zaokrągleń psują tę równość.
- Złożoność $O(n^3)$. Trochę więcej operacji niż eliminacja Gaussa. Korzystniej dla macierzy rzadkich, ale wtedy skuteczniejsze prostsze metody (np. Seidla).

Metoda ortogonalizacji Householdera

Idea: Kolejne transformacje macierzy dobierane tak, by wskaźnik uwarunkowania zadania nadmiernie nie wzrósł.

Analiza prowadzi do wniosku: przekształcenia o macierzach unitarnych zachowują wskaźnik uwarunkowania.

Propozycja Householdera: unitarna macierz transformacji $\mathbf{P}$ spełniająca

$$\mathbf{P} = \mathbf{I} - 2\mathbf{w}\mathbf{w}^H, \quad \text{gdzie } \mathbf{w}^H\mathbf{w} = 1, \quad \mathbf{w} \in C^n. \quad (430)$$

Macierz $\mathbf{P}$ jest hermitowska:

$$\mathbf{P}^H = \mathbf{I} - (2\mathbf{w}\mathbf{w}^H)^H = \mathbf{I} - 2\mathbf{w}\mathbf{w}^H = \mathbf{P}, \quad (431)$$

i unitarna:

$$\begin{aligned} \mathbf{P}^H\mathbf{P} &= \mathbf{P}\mathbf{P} = \mathbf{P}^2 = (\mathbf{I} - 2\mathbf{w}\mathbf{w}^H)(\mathbf{I} - 2\mathbf{w}\mathbf{w}^H) \\ &= \mathbf{I} - 2\mathbf{w}\mathbf{w}^H - 2\mathbf{w}\mathbf{w}^H + 4\mathbf{w}\mathbf{w}^H\mathbf{w}\mathbf{w}^H = \mathbf{I}. \end{aligned} \quad (432)$$

Wektor $\mathbf{w}$ dobierany jest tak, aby pewien wektor $\mathbf{x} = [x_1, \dots, x_n]^T$ był przekształcony na kierunek wektora jednostkowego $\mathbf{e}_1$:

$$\mathbf{P}\mathbf{x} = k\mathbf{e}_1. \quad (433)$$

Dokładniejsza analiza prowadzi do:

$$\mathbf{P} = \mathbf{I} - \frac{\mathbf{u}\mathbf{u}^H}{\gamma}, \quad (434)$$

gdzie

$$\sigma = \sqrt{\sum_{i=1}^n |x_i|^2}, \quad x_1 = e^{i\alpha}|x_1|, \quad k = -\sigma e^{i\alpha},$$

$$\mathbf{u} = \mathbf{x} - k\mathbf{e}_1 = \begin{bmatrix} e^{i\alpha}(|x_1| + \sigma) \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \quad \gamma = \sigma(\sigma + |x_1|). \quad (435)$$

Algorytm:

Przekształcamy macierz A krok po kroku zerując kolejne kolumny pod przekątną:

$$\mathbf{A} = \mathbf{A}^{(0)} \rightarrow \mathbf{A}^{(1)} \rightarrow \dots \rightarrow \mathbf{A}^{(n-1)} = \mathbf{R} = \begin{bmatrix} r_{11} & \dots & r_{1n} \\ & \ddots & \vdots \\ 0 & & r_{nn} \end{bmatrix}. \quad (436)$$

W j -tym kroku przekształcamy macierz

$$\mathbf{A}^{(j-1)} = \left[\begin{array}{ccc|ccc} r_{11} & \dots & x_{1,j-1} & a_{1j}^{(j-1)} & \dots & a_{1n}^{(j-1)} \\ & \ddots & \vdots & \vdots & & \vdots \\ 0 & & r_{j-1,j-1} & a_{j-1,j}^{(j-1)} & \dots & a_{j-1,n}^{(j-1)} \\ \hline & & & a_{jj}^{(j-1)} & \dots & a_{jn}^{(j-1)} \\ & & & \vdots & & \vdots \\ & 0 & & a_{nj}^{(j-1)} & \dots & a_{nn}^{(j-1)} \end{array} \right] = \begin{bmatrix} \mathbf{D} & \mathbf{B} \\ 0 & \tilde{\mathbf{A}}^{(j-1)} \end{bmatrix}. \quad (437)$$

Szukamy takiej unitarnej macierzy przekształcenia $\mathbf{P}_j$, aby

$$\mathbf{P}_j = \begin{bmatrix} \mathbf{I}_{j-1} & 0 \\ 0 & \tilde{\mathbf{P}}_j \end{bmatrix}, \quad \tilde{\mathbf{P}}_j \begin{bmatrix} a_{jj}^{(j-1)} \\ \vdots \\ a_{nj}^{(j-1)} \end{bmatrix} = k\mathbf{e}_1 \in C^{n-j+1}. \quad (438)$$

Uwagi:

- Mnożenie przez macierz

$$\tilde{\mathbf{P}}_j = \mathbf{I} - \frac{\mathbf{u}_j \mathbf{u}_j^H}{\gamma_j} \quad (439)$$

realizuje się w następujący sposób:

$$\tilde{\mathbf{P}}_j \tilde{\mathbf{A}}^{(j-1)} = \tilde{\mathbf{A}}^{(j-1)} - \mathbf{u}_j \mathbf{y}_j^H, \quad \text{gdzie } \mathbf{y}_j^H = \frac{\mathbf{u}_j^H \tilde{\mathbf{A}}^{(j-1)}}{\gamma_j}. \quad (440)$$

- Algorytm wymaga około $\frac{2n^3}{3}$ działań.
- Współrzędne wektora $\mathbf{u}$ wpisuje się w macierz $\mathbf{A}$ pod przekątną – brakuje tylko miejsca na jeden wektor n wartości.
- Kolumny macierzy $\mathbf{Q} = \mathbf{P}^{-1} = \mathbf{P}_1 \cdots \mathbf{P}_{n-1}$ są ortonormalne

Metoda Householdera – przykład

Rozwiązać układ równań

$$\begin{bmatrix} 2 & 4 & -1 \\ 2 & -1 & 2 \\ 1 & -3 & 1 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 0 \\ 6 \\ 3 \end{bmatrix} \quad (441)$$

sprowadzając macierz główną do postaci trójkątnej metodą Householdera.

Metoda ortogonalizacji Grama-Schmidta

Jeśli macierz A ma rozkład:

$$A = QR, \quad (442)$$

gdzie macierz Q ma ortonormalne kolumny, to można je wyznaczyć metodą ortogonalizacji Grama-Schmidta.

Dla k -tej kolumny macierzy $A = (a_1, \dots, a_n)$ mamy:

$$a_k = \sum_{i=0}^k r_{ik} q_i, \quad k = 1, \dots, n. \quad (443)$$

Kolumna a_k jest kombinacją liniową wektorów $q_1, \dots, q_k$ i odwrotnie: wektor q_k jest kombinacją liniową $a_1, \dots, a_k$.

Kolejne q_k wyznaczamy rekurencyjnie tak, by zachowywać ortonormalność.

Algorytm:

- Wyznaczamy q_1 i r_{11} :

$$r_{11} \leftarrow \|a_1\|, \quad q_1 \leftarrow \frac{a_1}{r_{11}}. \quad (444)$$

- Znamy $q_1, \dots, q_{k-1}$ oraz r_{ij} dla $j \leq k-1$.
 - Wyznaczamy wektor

$$b_k \leftarrow a_k - r_{1k}q_1 - \dots - r_{k-1,k}q_{k-1} \quad (445)$$

tak, aby był ortogonalny do $q_1, \dots, q_{k-1}$:

$$q_i^H b_k = 0 \quad \Rightarrow \quad r_{ik} \leftarrow q_i^H a_k, \quad i = 1, \dots, k-1 \quad (446)$$

- Wyznaczamy q_k i r_{kk} :

$$r_{kk} \leftarrow \|b_k\|, \quad q_k \leftarrow \frac{b_k}{r_{kk}}. \quad (447)$$

Uwagi:

- Dla wszystkich $1 \leq i, j \leq k$ mamy:

$$q_i^H q_j = \begin{cases} 1 & \text{dla } i = j, \\ 0 & \text{w przec. przyp.} \end{cases} \quad (448)$$

- Algorytm jest niestabilny numerycznie gdy wektory a_i są bliskie liniowej zależności – błędy numeryczne sprawiają, że $\bar{b}_k$ nie jest ortogonalny do q_i .

Reortogonalizacja wektora $\bar{b}_k$

Obliczamy poprawki w postaci skalarów

$$\Delta r_{ik} \leftarrow q_i^H \bar{b}_k, \quad i = 1, \dots, k-1, \quad (449)$$

oraz wektor

$$\tilde{b}_k = \bar{b}_k - \Delta r_{1k} q_1 - \dots - \Delta r_{k-1,k} q_{k-1}. \quad (450)$$

Poprawiamy:

$$r_{kk} \leftarrow \|\tilde{b}_k\|, \quad q_k \leftarrow \frac{\tilde{b}_k}{r_{kk}}, \quad r_{ik} \leftarrow \bar{r}_{ik} + \Delta r_{ik}. \quad (451)$$

Wyznaczanie wartości i wektorów własnych

Definicja 19 Wektor x jest **wektorem własnym** macierzy A jeśli istnieje taka liczba λ , że $Ax = \lambda x$. λ to **wartość własna** macierzy A .

Twierdzenie 33 Liczba λ jest wartością własną macierzy A wtedy i tylko wtedy, gdy jest pierwiastkiem **wielomianu charakterystycznego** $\det(A - \lambda I)$ macierzy A .

Definicja 20 Macierze A i B są **podobne** jeśli istnieje nieosobliwa **macierz podobieństwa** P tj. taka, że $P^{-1}AP = B$.

Twierdzenie 34 Macierze podobne mają takie same wartości własne.

Twierdzenie 35 Wartości i wektory własne macierzy symetrycznej są rzeczywiste.

Twierdzenie 36 Wartości własne macierzy $A + cI$ to wartości własne macierzy A przesunięte o c .

Metoda Kryłowa

- Szukanie wartości własnych jako pierwiastków wielomianu charakterystycznego (1931 r.).
- Zagadnienie wyznaczenia współczynników wielomianu charakterystycznego jest znacznie gorzej uwarunkowane niż zadanie wyznaczenia wartości własnych.
- W zasadzie metoda nie do użytku.

Metoda potęgowa

Dla wyznaczania pojedynczych wartości i wektorów własnych.

Algorytm:

- $i \leftarrow 0$, wybieramy wektor $\mathbf{x}_0$ taki, że $\|\mathbf{x}_0\|_\infty = 1$
- Tak długo jak i jest mniejsze niż założona liczba iteracji:
 - $\mathbf{v}_{i+1} \leftarrow \mathbf{A}\mathbf{x}_i$, $m_{i+1} \leftarrow \|\mathbf{v}_{i+1}\|_\infty$
 - Jeżeli $m_{i+1} = 0$ to KONIEC!
 - $\mathbf{x}_{i+1} \leftarrow \frac{\mathbf{v}_{i+1}}{m_{i+1}}$, $i \leftarrow i + 1$

Uwagi:

- Jeśli $\mathbf{x}_{2i} \xrightarrow{i \rightarrow \infty} \mathbf{x}$, to $m_i \xrightarrow{i \rightarrow \infty} m$.
- Jeśli $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_\infty \xrightarrow{i \rightarrow \infty} 0$, to $\mathbf{A}\mathbf{x} = m\mathbf{x}$.
- Jeśli $\|\mathbf{x}_{i+1} - \mathbf{x}_i\|_\infty \xrightarrow{i \rightarrow \infty} 2$, to $\mathbf{A}\mathbf{x} = -m\mathbf{x}$.
- W ogólnym przypadku trudno ocenić kiedy ciąg $(\mathbf{x}_{2i})$ jest zbieżny.
- Można wykazać zbieżność np. gdy macierz jest podobna do diagonalnej.

Metoda QR

Dla macierzy Hessenberga $\mathbf{H}$ ($h_{ij} = 0$ dla $j < i - 1$).

Schemat algorytmu:

- $i \leftarrow 1, \mathbf{H}^{(1)} \leftarrow \mathbf{H}$
- Do uzyskania stosownej zbieżności:
 - Wyznaczamy rozkład $\mathbf{H}^{(i)} = \mathbf{Q}^{(i)} \mathbf{R}^{(i)}$
 - $\mathbf{H}^{(i+1)} \leftarrow \mathbf{R}^{(i)} \mathbf{Q}^{(i)}$
 - $i \leftarrow i + 1$

Uwagi:

- Wszystkie $\mathbf{Q}^{(i)}$ oraz $\mathbf{H}^{(i)}$ są macierzami Hessenberga.
- Wszystkie $\mathbf{H}^{(i)}$ są podobne, bo $\mathbf{H}^{(i+1)} = \mathbf{Q}^{(i)-1} \mathbf{H}^{(i)} \mathbf{Q}^{(i)}$.
- Jeśli wartości własne $\mathbf{H}$ są rzeczywiste i mają różne moduły, to $\mathbf{H}^{(i)}$ zbiegają do macierzy trójkątnej, a elementy diagonalni do wartości własnych $\mathbf{H}$.
- Jeśli wśród wartości własnych są pary sprzężone (poza tym nie ma wartości o równych modułach), to nie wszystkie elementy pod diagonalą dążą do zera.

Problemy zbieżności

- Analiza szybkości zbieżności pokazuje, że zależy ona od stosunku modułów odpowiednich wartości własnych.
- Dotyczy zarówno metody potęgowej jak i QR.
- Sposób na przyspieszenie: przesunięcie $\mathbf{B} = \mathbf{A} - p\mathbf{I}$.
 - Metoda potęgowa, gdy znamy wartości własne, a szukamy wektorów:
 $p = \frac{1}{2}(\lambda_i + \lambda_n)$ dla λ_1 , gdzie $i = \min\{k > 1 : \lambda_k < \lambda_1\}$ oraz $p = \frac{1}{2}(\lambda_1 + \lambda_j)$ dla λ_n , gdzie $j = \max\{k < n : \lambda_k > \lambda_n\}$.
 - Metoda QR:
 $\mathbf{H}^{(i)} - p_i\mathbf{I} = \mathbf{Q}^{(i)}\mathbf{R}^{(i)} \quad \mathbf{H}^{(i+1)} = \mathbf{R}^{(i)}\mathbf{Q}^{(i)} + p_i\mathbf{I},$
 p_i równe otrzymanym w poprzednich iteracjach wartościom własnym.

Sprowadzanie macierzy do postaci Hessenberga

- eliminacja Gaussa
- metoda Householdera
- metoda Givensa

Szukanie wektorów własnych

- Jeżeli $\mathbf{P}^{-1}\mathbf{A}\mathbf{P} = \mathbf{B}$, oraz $\mathbf{x}$ jest wektorem własnym macierzy $\mathbf{B}$, to $\mathbf{y} = \mathbf{P}\mathbf{x}$ jest wektorem własnym macierzy $\mathbf{A}$.
- Jeśli macierz $\mathbf{B}$ jest trójkątna górna, to wartości własnej $\lambda_i = b_{ii}$ odpowiada wektor własny $\mathbf{x}^{(i)}$ taki, że:

$$\begin{aligned} x_j^{(i)} &= 0, & j &= n, n-1, \dots, i+1 \\ x_i^{(i)} &= 1 \\ x_j^{(i)} &= -\frac{\sum_{k=j+1}^i b_{jk} x_k^{(i)}}{b_{jj} - b_{ii}}, & j &= i-1, \dots, 1 \end{aligned} \tag{452}$$

- Jeśli $\mathbf{B}$ ma wielokrotną wartość własną λ , to wektor własny można wyliczyć tylko dla najmniejszego i takiego, że $b_{ii} = \lambda$.

Metoda QR

- Macierze $\mathbf{Q}^{(i)}$ można wykorzystać do obliczenia wektorów własnych.

- **Algorytm:** Dane: macierz $\mathbf{A}$.

- Sprowadzamy macierz $\mathbf{A}$ do postaci Hessenberga

$$\mathbf{P}^{-1} \mathbf{A} \mathbf{P} = \mathbf{H}. \quad (453)$$

- Obliczamy ciąg macierzy $\mathbf{H}^{(i)}$ oraz $\mathbf{Q}^{(i)}$ algorytmem QR dla macierzy Hessenberga $\mathbf{H}$. Jednocześnie wyznaczamy

$$\mathbf{Z}^{(1)} = \mathbf{P}, \quad \mathbf{Z}^{(i+1)} = \mathbf{Z}^{(i)} \mathbf{Q}^{(i)}. \quad (454)$$

- Po uznaniu w $\mathbf{H}^{(N)}$ zbieżności do macierzy trójkątnej górnej wyznaczamy jej wektory własne zgodnie z (452).
- Wyznaczamy wektory własne macierzy $\mathbf{A}$:

$$\mathbf{y}^{(k)} = \mathbf{Z}^{(N)} \mathbf{x}^{(k)}. \quad (455)$$

Metoda LR

- **Algorytm:** Dane: macierz $\mathbf{A}$.
 - Obliczamy ciąg macierzy $\mathbf{A}^{(i)}$ dla $i = 1, 2, \dots$:

$$\begin{aligned} \mathbf{A}^{(1)} &\leftarrow \mathbf{A} \\ \mathbf{A}^{(i)} &= \mathbf{L}^{(i)} \mathbf{U}^{(i)}, \quad \mathbf{A}^{(i+1)} \leftarrow \mathbf{U}^{(i)} \mathbf{L}^{(i)} \end{aligned} \tag{456}$$

do uzyskania zbieżności do macierzy trójkątnej.

Jednocześnie wyznaczamy:

$$\mathbf{Z}^{(1)} = \mathbf{I}, \quad \mathbf{Z}^{(i+1)} = \mathbf{Z}^{(i)} \mathbf{L}^{(i)}. \tag{457}$$

- Wyznaczamy wektory własne macierzy $\mathbf{A}^{(N)}$ zgodnie z (452).
- Wyznaczamy wektory własne macierzy $\mathbf{A}$:

$$\mathbf{y}^{(k)} = \mathbf{Z}^{(N)} \mathbf{x}^{(k)}. \tag{458}$$

- 3 razy mniej działań niż w QR ale gorsze uwarunkowanie.
- Stabilna numerycznie dla $\mathbf{A}$ symetrycznych i dodatnio określonych.
- Jak w QR: przesunięcia i do postaci Hessenberga ($O(n^3) \rightarrow O(n^2)$).

Metoda iteracji odwrotnej

- Przyspieszenie zbieżności metody potęgowej
- **Algorytm:**
 - $i \leftarrow 0$, wybieramy wektor $\mathbf{x}_0$ taki, że $\|\mathbf{x}_0\|_\infty = 1$
 - Tak długo jak i jest mniejsze niż założona liczba iteracji:
 - * $\mathbf{v}_{i+1} \leftarrow (\mathbf{A} - k_i \mathbf{I})^{-1} \mathbf{x}_i$, $m_{i+1} \leftarrow \|\mathbf{v}_{i+1}\|_\infty$
 - * $\mathbf{x}_{i+1} \leftarrow \frac{\mathbf{v}_{i+1}}{m_{i+1}}$, $i \leftarrow i + 1$
- Jeśli $\mathbf{A}$ ma wartości własne $\lambda_1, \dots, \lambda_n$ takie, że $|\lambda_1| \geq \dots \geq |\lambda_n|$, to wartościami własnymi $(\mathbf{A} - k_i \mathbf{I})^{-1}$ są:

$$\frac{1}{\lambda_1 - k_i}, \frac{1}{\lambda_2 - k_i}, \dots, \frac{1}{\lambda_n - k_i}. \quad (459)$$

- Ciąg $m_1, m_2, \dots$ pozwala wyznaczyć wartość własną λ_n jak w metodzie potęgowej.
- Jeśli k_i jest bliskie λ_n , to macierz $\mathbf{A} - k_i \mathbf{I}$ jest źle uwarunkowana, co można zlekceważyć gdy $\mathbf{x}$ jest dobrze uwarunkowany i norma $\|\mathbf{v}_{i+1}\|_\infty$ jest duża.