NUMERYCZNA REPREZENTACJA LICZB

- Każdą liczbę rzeczywistą $x \neq 0$ można zapisać w postaci ZNORMALIZOWANEJ:

$$x = \pm m \cdot \beta^c = \pm (0.a_1 a_2 \ldots)_\beta \cdot \beta^c$$

gdzie β jest bazą rozwinięcia (np.: 2, 8, 10, 16, ...), liczba $m \in [1/\beta, 1)$ nazywana mantysą, c jest liczbą całkowitą zwaną cechą, $a_i \in \{0, 1, \ldots, \beta - 1\}$, gdzie $a_1 \neq 0$.

- PRZYKŁAD:

$\beta = 10$

$732.501 = 0.732501 \cdot 10^3$

$-0.005612 = -0.5612 \cdot 10^{-2}$

$\beta = 2$

$(1011.11101)_2 = (0.101111101)_2 \cdot 2^4$

$(0.00101)_2 = (0.101)_2 \cdot 2^{-2}$

W ARYTMETYCE ZMIENNOPOZYCYJNEJ lub w skrócie "fl" zakłada się:

- $rd(x) = \pm m_t \cdot \beta^c = \pm (0.a_1, a_2, \ldots, a_t)_\beta \cdot \beta^c$,

ustalony zakres cechy c, $c \in [c_{min}, c_{max}]$. Reprezentację x w arytmetyce fl oznaczamy przez $rd(x)$ lub $\tilde{x}$. W arytmetyce fl wyróżniamy dwa sposoby zastąpienia:

- zaokrąglenie
- obcięcie.

-

Wyliczenie błędu względnego dla zaokrąglenia oraz obcięcia:

$$|\delta| = \frac{|\tilde{x} - x|}{|x|} \qquad rd(x) = x(1+\delta)$$

Precyzja arytmetyki dla ϵ (epsilon)

ZAOKRĄGLENIA:

$$\epsilon = 0.5\,\beta^{1-t}$$

OBCIĘCIA:

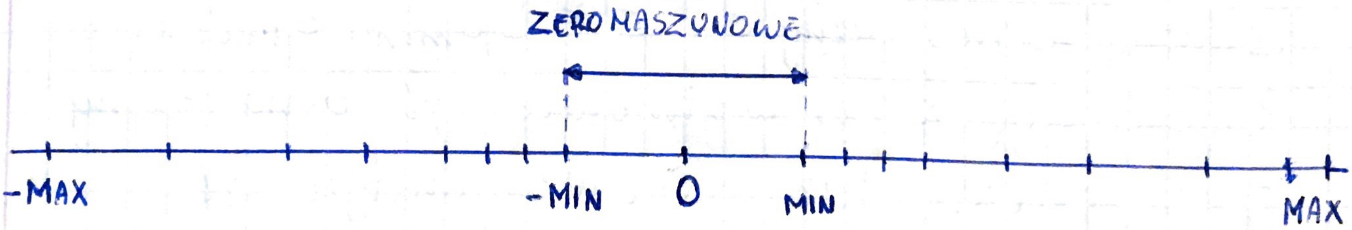
$$\epsilon = \beta^{1-t}$$

Aby sprawdzić dobre uwarunkowanie, czyli nie popełnianie wielkiego błędu względnego, musi zachodzić nierówność:

$$|\delta| \leqslant \epsilon$$

PRECYZJA ARYTMETYKI $\iff$ BŁĄD REPREZENTACJI

LICZBY W ARYTMETYCE fl tworzą siatkę symetryczną niejednorodną (zob. poniżej).



Jeśli $|x| > MAX$, to mówimy o NADMIARZE (przerywane obliczenia).
Jeśli $|x| < MIN$, to $\tilde{x} = 0$, i mówimy o NIEDOMIARZE.

$$\boxed{MIN = \frac{1}{\beta}\beta^{c_{min}}} \leqslant |x| \leqslant \boxed{(1-\beta^{-t})\beta^{c_{max}} = MAX}$$

PRZYKŁAD

Arytmetyka dwójkowa, czyli $\beta=2$, w której 1 bit przeznaczono na zapis znaku s liczby x, 6 bitów na zapis cechy c (wraz z bitem znaku) i 16 bitów na zapis mantysy, $t=16$. Wówczas $c \in [-31, 32]$, bo 2^5, $((\beta^{c-3})) = 32$.

$MIN = \frac{1}{2}\beta^{c_{min}}$ $MAX = (1-\beta^{-t})\beta^{c_{max}}$

$MIN = \frac{1}{2}2^{-31} \approx 4.66 \cdot 10^{-10}$ $MAX = (1-2^{-16})2^{32} \approx 4.29 \cdot 10$

WYLICZENIE c_{min} i c_{max}. ZAKRES REPREZENTOWANYCH LICZB W ARYTMETYCE [C]

$c_{min} \leqslant c \leqslant c_{max}$ gdzie d = ilość bitów cechy.

$-2^d + 1 \leqslant c \leqslant 2^d$

PAMIĘTAJ, że jeżeli cecha jest z nadmiarem, czyli "zapis cechy c wraz z bitem znaku", wtedy

$-2^{d-1}+1 \leqslant c \leqslant 2^d$

ZERO MASZYNOWE - to zakres pomiędzy [-MIN, MIN], ponieważ każda liczba z tego przedziału jest ZEREM.

PSEUDOKOD NA "macheps", epsilon maszynowy.

```
x = 1.0
while (1.0 + x > 1.0)
      x = x / 2.0
printf("Macheps", 2.0 * x)
```

wynik 2^{-t}, gdzie t to liczba cyfr mantysy

UWARUNKOWANIE ZADANIA OBLICZENIOWEGO

PAMIĘTAJ!

Odejmowanie liczb, to pierwsze, na co musisz zwrócić uwagę, w analizowaniu algorytmu, czy poprawnego uwarunkowania zadania.

Odejmowanie ~~od siebie~~ bliskich siebie liczb prowadzi do dużych błędów, a dokładniej, do utraty cyfr znaczących.

Wskaźnik uwarunkowania funkcji:

$$\text{cond}(x) = \frac{|f'(x) \cdot x|}{|f(x)|}$$

Jeżeli cond(x) jest mały (czyli $|f'(x)|$ jest małe), to funkcja jest DOBRZE UWARUNKOWANA.

ZAPAMIĘTAJ PRZY ANALIZIE BŁĘDÓW!

$$fl\left(\frac{a}{b(1+\alpha)}\right) = \frac{a}{b(1+\alpha)}(1+\sigma) = \frac{a}{b}(1+\beta),$$

gdzie $|\sigma| \leqslant \varepsilon = 2^{-t}$, $|\beta| < |\alpha + \sigma|$

- Rozwiązania z listy nr 2, wszystko wyjaśnia.

"Złośliwy wielomian" (Wilkinson)

Wyznaczanie pierwiastków wielomianu jest źle uwarunkowane ze względu na zaburzenia współczynników

SLAJDY 3

ROZWIĄZYWANIE RÓWNAŃ NIELINIOWYCH

Dana jest funkcja: $f: \mathbb{R} \to \mathbb{R}$. Na przykład $f(x) = x - tg\,x$

Szukamy takiego $r \in \mathbb{R}$, dla którego

$$f(r) = 0.$$

Będziemy rozważać metody iteracyjne. Metody te konstruują ciąg przybliżeń $x_0, x_1, x_2, \ldots$, według reguły:

$$x_{n+1} := \Phi(x_n).$$

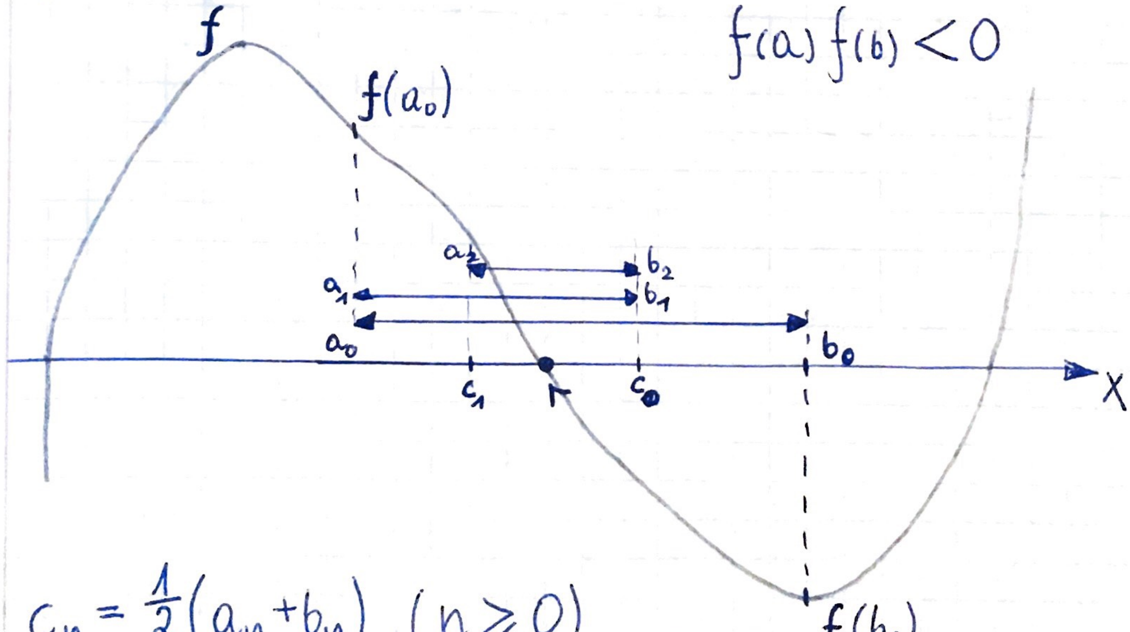
taki, że

$$\lim_{n\to\infty} x_n = r$$

Mówimy o metodzie, że jest zbieżna globalnie jeśli konstruowany ciąg przybliżeń $\{x_n\}$ jest zbieżny do r przy dowolnych przybliżeniach początkowych.

METODA BISEKCJI (POŁOWIENIA)

Założenia: f jest funkcją ciągłą w $[a, b]$ i $f(a)f(b) < 0$ (f zmienia znak).



$$c_n = \frac{1}{2}(a_n + b_n), \quad (n \geq 0)$$

WARUNEK KOŃCA: $|b_n - a_n| \leq \delta$, $|f(c_n)| \leq \varepsilon$

Tw. 1. (O ZBIEŻNOŚCI M. BISEKCJI).

Niech $[a_0, b_0], [a_1, b_1], \dots, [a_n, b_n]$ będzie ciągiem przedziałów konstruowanych przez metodę bisekcji. Wówczas istnieją granice $\lim_{n\to\infty} a_n$ i $\lim_{n\to\infty} b_n$ i są sobie równe, reprezentujące zero r funkcji f. Jeśli $r = \lim_{n\to\infty} c_n$ i $c_n = \frac{1}{2}(a_n + b_n)$, to

$$|r - c_n| \leq 2^{-(n+1)}(b_0 - a_0).$$

① Przykład.

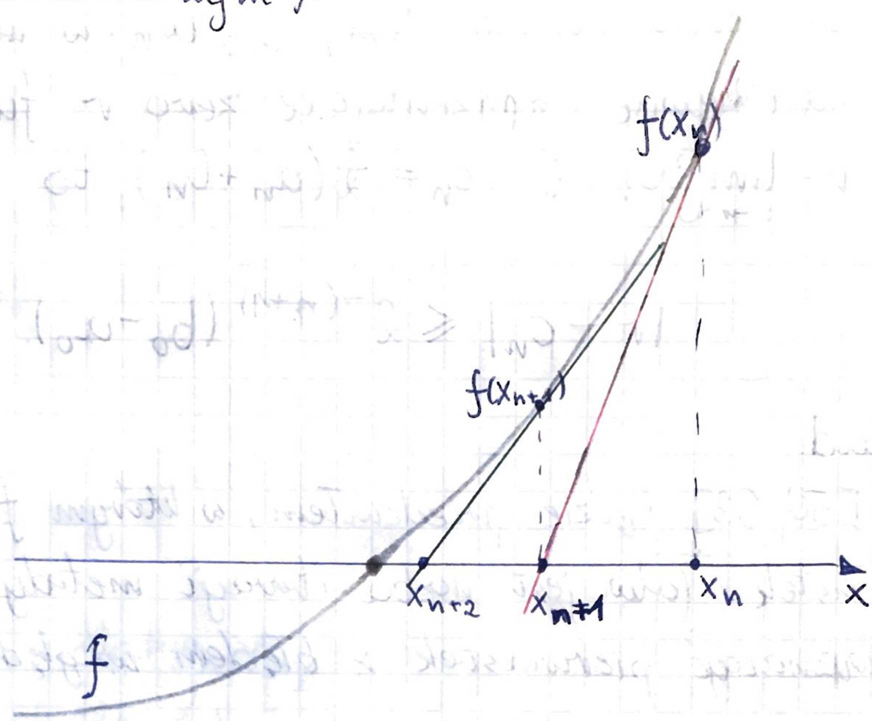
Niech $[50, 63]$ będzie przedziałem, w którym f ma pierwiastek. Jaka jest liczba iteracji metody bisekcji, aby wyznaczyć pierwiastek z błędem względnym 10^{-12}?

$$\frac{|r - c_n|}{|r|} \leq \frac{|r - c_n|}{50} \leq 2^{-(n+1)}\frac{13}{50} \leq 10^{-12}$$

Rozwiązując powyższą nierówność otrzymujemy $n \geq 37$

METODA NEWTONA (METODA STYCZNYCH)

Założenia: $f \in C^2[a,b]$, $f'(r) = 0$ (r jest pierwiastkiem jednokrotnym).



$f(x) = f(x_n) + f'(x_n)(x - x_n) + O((x - x_n)^2)$ sz. Taylora.

$f(x) \approx f(x_n) + f'(x_n)(x - x_n)$, linearyzacja.

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} = x_n + h_n \quad (n \geqslant 0), \quad x_0 \to \text{podane.}$$

WARUNEK KOŃCOWY: $|x_{n+1} - x_n| \leqslant \delta$, $|f(x_{n+1})| \leqslant \varepsilon$

Def. "METODA NEWTONA (M. STYCZNYCH)".
Niech $\{x_n\}$ będzie ciągiem zbieżnym do r. Jeśli istnieją stałe C, α i liczba naturalna $\mathbb{N}$, takie że

$$|x_{n+1} - r| \leqslant C|x_n - r|^\alpha \quad (n \geqslant \mathbb{N}),$$

to mówimy, że wykładnik zbieżności jest rzędu α.

$\alpha = 1 \to$ zbieżność liniowa, $C < 1$.

$\alpha = 2 \to$ zbieżność kwadratowa

$\alpha = 3 \to$ —— " —— sześcienna.

Tw. 2. (O LOKALNEJ ZBIEŻNOŚCI METODY NEWTONA).
Niech $f \in C^2[a,b]$ i r będzie jednokrotnym pierwiastkiem f. Wtedy istnieje otoczenie r i stała C i jeśli przybliżenie początkowe x_0 należy do otoczenia r, to ciąg konstruowanych przez metodę Newtona przybliżeń $\{x_n\}$ spełnia

$$|x_{n+1}-r| \leqslant \cancel{\tfrac{1}{2}} C(x_n-r)^2$$

Ponadto $\lim\limits_{n\to\infty} x_n = r$.

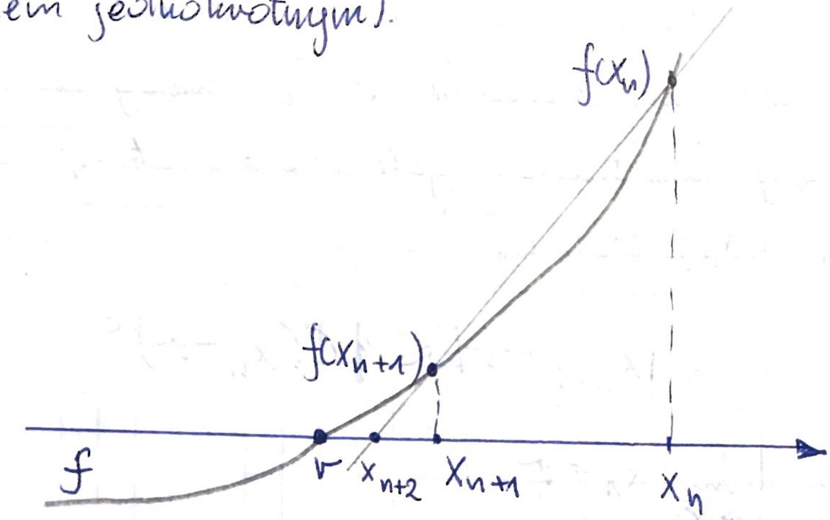
② PRZYKŁAD.
$f(x) = (x/2)^2 + \sin x$. Wyznaczyć metodą stycznych ZERO f w przedziale $[1,5; 2]$ z dokładnością $\delta = 0.5_{10}-5$. $x_0 = 1,5$.

n	x_n	h_n	$\lvert x_n - r\rvert$
0	1.5	0.6439	-0.43374
1	2.14039	-0.18838	0.26644
2	1.95201	-0.018808	0.0111825
3	1.93393	-0.00018	0.00018
4	1.933375	$< 5_{10}-6$	$< 5_{10}-6$.

METODA BISEKCJI WYKONAŁA 17 ITERACJI dla $a=1.5$ i $b=2$.

METODA SIECZNYCH

Założenia: $f \in C^2[a,b]$, $f'(r) = 0$. (r jest pierwiastkiem jednokrotnym).



APROKSYMUJEMY $f'(x_n) \approx \dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$

$$x_{n+1} = x_n - \frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} f(x_n) = x_n + h_n \quad (n \geq 1),\ x_0, x_1 \text{ dane}$$

WARUNEK KOŃCA: $|x_{n+1} - x_n| \leq \delta$, $|f(x_{n+1})| \leq \varepsilon$.

Tw. 4 (O LOKALNEJ ZBIEŻNOŚCI METODY SIECZNYCH).
Niech $f \in C^2[a,b]$ i r będzie jednokrotnym pierwiastkiem f. Wtedy istnieje otoczenie r i stała K i jeśli przybliżenia początkowe x_0, x_1 należą do otoczenia r, to ciąg konstruowanych przez metodę siecznych przybliżeń $\{x_n\}$ spełnia

$$|x_{n+1} - r| \leq K |x_n - r|^{(1+\sqrt{5})/2}.$$

PONADTO $\lim_{n\to\infty} x_n = r$.

PODSUMOWANIE

METODY	ZBIEŻNOŚĆ	WYKŁADNIK α	UWAGI
BISEKCJI	globalna	1	STOSOWAĆ HYBRYDOWO
NEWTONA	lokalna	2	KONIECZNOŚĆ LICZENIA $f'(x)$
SIECZNYCH	lokalna	$(1+\sqrt{5})/2 \approx 1.618$	—

SLAJDY 5

PIERWIASTKI WIELOMIANÓW.

POSTAĆ NATURALNA WIELOMIANU:

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_2 z^2 + a_1 z + a_0, \quad a_n \neq 0$$

gdzie współczynniki a_k i zmienna z mogą być liczbami zespolonymi ($a_k, z \in \mathbb{C}$).

Algorytm Hornera.

Załóżmy, wielomian p jest dany w postaci:

$$p(z) = a_n z^n + a_{n-1} z^{n-1} + \ldots + a_2 z^2 + a_1 z + a_0$$

p możemy zapisać w postaci:

$$p(z) = (\ldots((a_n z + a_{n-1}) z + a_{n-2}) z + \ldots) z + a_0$$

Dla danego z_0, algorytm Hornera wyznacza $p(z_0)$ oraz wielomian $q(z) = b_0 + b_1 z + \ldots + b_{n-1} z^{n-1}$, taki, że

$$p(z) = (z - z_0)\, q(z) + p(z_0).$$

ALGORYTM HORNERA.

Dane: $n, z_0, a_0, \ldots, a_n$

Wyniki: $p(z_0), b_0, \ldots, b_{n-1}$

```
b_{n-1} ← a_n;
for k ← n-1 downto 0 do
    b_{k-1} ← a_k + z_0 b_k;
end for
return b_{-1}, b_0, ..., b_{n-1};
```

RĘCZNE OBLICZENIA $p(z_0)$ i $q(z)$:

	a_n	a_{n-1}	a_{n-2}	$\dots$	a_0
z_0		$z_0 b_{n-1}$	$z_0 b_{n-2}$	$\dots$	$z_0 b_0$
	b_{n-1}	b_{n-2}	b_{n-3}	$\dots$	b_{-1}

① Przykład.

Oblicz za pomocą algorytmu Hornera wartość wielomianu p w punkcie 3 i wielomian $q(z)$.

$$p(z) = z^4 - 4z^3 + 7z^2 - 5z - 2.$$

	1	-4	7	-5	-2
3		3	-3	12	21
	1	-1	4	7	19

$p(3) = 19$, $q(z) = z^3 - z^2 + 4z + 7$

$p(z) = (z-3)(z^3 - z^2 + 4z + 7) + 19$

Za pomocą metody Newtona można obliczać pierwiastki danego wielomianu $p(z)$:

$$z_{k+1} = z_k - \frac{p(z_k)}{p'(z_k)}.$$

Po zróżniczkowaniu równości $p(z) = (z - z_0) q(z) + p(z_0)$ otrzymujemy: $p'(z) = q(z) + (z - z_0) q'(z)$.

Następnie dla $z = z_0$ otrzymujemy $p'(z_0) = q(z_0)$.

DEFLACJA.

Po wyznaczeniu pierwiastka r_1 wielomianu P musimy wyznaczyć pozostałe pierwiastki $r_2, \dots, r_n$. „ODDZIELAMY" więc obliczony pierwiastek r_1, tzn. wyznaczamy wielomian $q(z)$ stopnia $n-1$

$$p(z) = (z - r_1)\, q(z).$$

Proces ten nazywamy deflacją.

SLAJD 6.

- MACIERZ VANDERMONDE'A → ŹLE UWARUNKOWANA.

INTERPOLACJA ZA POMOCĄ WIELOMIANÓW.

Tw. 1.
Dla węzłów x_i i liczb y_i istnieje dokładnie jeden wielomian $p \in \Pi_n$ spełniający warunki interpolacji
$$p(x_i) = y_i \quad (0 \leqslant i \leqslant n).$$

- POSTAĆ LAGRANGE'A WZORU INTERPOLACYJNEGO

$$L_n(x) = \sum_{i=0}^{n} y_i \prod_{\substack{j=0 \\ j \neq i}}^{n} \frac{x - x_j}{x_i - x_j}$$

ILORAZY RÓŻNICOWE

- POSTAĆ NEWTONA WZORU INTERPOLACYJNEGO

$$p(x) = \sum_{k=0}^{n} c_k q_k(x) = \sum_{k=0}^{n} f[x_0, x_1, \ldots, x_k] \prod_{j=0}^{k-1} (x - x_j).$$

Wynika ⇒ $f[x_i] = f(x_i)$, dla $0 \leqslant i \leqslant n$.

Twierdzenie 3 (ILORAZY RÓŻNICOWE WYŻSZYCH RZĘDÓW).
Ilorazy różnicowe spełniają równość:

- $$f[x_0, x_1, x_2, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}.$$

Jeżeli dana jest tablica wartości funkcji $(x_i, f(x_i))$, wówczas ilorazy różnicowe oblicza się konstruując tabelę trójkątną:

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

Twierdzenie 4

Ilorazy różnicowe nie zależą od kolejności węzłów. Jeśli $(z_0, z_1, \ldots, z_n)$ jest permutacją $(x_0, x_1, \ldots, x_n)$, to

$$f[z_0, z_1, \ldots, z_n] = f[x_0, x_1, \ldots, x_n].$$

Twierdzenie 5

Niech będzie wielomianem $p \in \Pi_n$ interpolującym f różnych węzłach $x_0, x_1, \ldots, x_n$. Jeżeli $t \neq x_i$, wtedy

$$f(t) - p(t) = f[x_0, x_1, \ldots, x_n, t] \prod_{j=0}^{n} (t - x_j).$$

Twierdzenie 6

Jeżeli $f \in C^{n+1}[a,b]$ i węzły $x_0, x_1, \ldots, x_n \in (a,b)$. Wtedy istnieje $\delta \in (a,b)$ taka, że

$$f[x_0, x_1, \ldots, x_n] = \frac{1}{n!} f^{(n)}(\delta)$$

SLAJDY 7

POSTAĆ NEWTONA WZORU INTERPOLACYJNEGO:

$$p_n(x) = \sum_{k=0}^{n} c_k q_k(x) = \sum_{k=0}^{n} f[x_0, x_1, \dots, x_k] \prod_{j=0}^{k-1} (x - x_j)$$

Współczynniki $c_0 = f[x_0]$, $c_1 = f[x_0, x_1]$, ..., $c_n = f[x_0, x_1, \dots, x_n]$, wyznaczamy za pomocą tablicą trójkątną.

- Konstrukcja wielomianu interpolacyjnego ~~p~~ $p_{n+1}(x)$:

$$p_n(x_i) = y_i \quad (0 \leq i \leq n)$$

$$p_{n+1}(x_i) = y_i \quad (0 \leq i \leq n+1).$$

Wielomian $p_{n+1}(x)$ można przedstawić w postaci:

$$p_{n+1}(x) = p_n(x) + c_{n+1}(x - x_0) \dots (x - x_n).$$

Zatem wyznaczenie $p_{n+1}(x)$ na podstawie $p_n(x)$ (poprzedniego) wymaga wyznaczenia:

$$c_{n+1} = f[x_0, x_1, \dots, x_{n+1}] \quad \text{(lokalna modyfikacja)}$$

- WIELOMIANY CZEBYSZEWA.

wybór węzłów, aby (~~aby~~ mieć najmniejszą wielkość błędu).

$$\max_{x \in [a,b]} \left| \prod_{i=0}^{n} (x - x_i) \right| \to \min$$

- Proces optymalizacji prowadzi do wielomianów Czebyszewa (I-szego rodzaju), które można zdefiniować rekurencyjnie:

$$\begin{cases} T_0(x) = 1 \quad T_1(x) = x \\ T_{n+1} = 2xT_n(x) - T_{n-1}(x) \quad (n \geq 1). \end{cases}$$

Tw. 1

Dla $x \in [-1, 1]$, wielomiany Czebyszewa można przedstawić w postaci

$$\cancel{T(x)}\ T_n(x) = \cos(n \arccos x) \quad (n \geq 0).$$

Tw. 2

Spośród wszystkich wielomianów stopnia n o współczynniku przy najwyższej potędze równym 1, $w_n(x) = x^k + \dots$, wielomian T_n ma najmniejszą normę:

$$\frac{1}{2^{n-1}} = \max_{x \in [-1,1]} |\tilde{T}_n(x)| \leq \max_{x \in [-1,1]} |w_n(x)|.$$

Tw. 3

Jeżeli w przedziale $[-1,1]$ za węzły interpolacji x_i przyjmiemy zera wielomianu Czebyszewa T_{n+1}, wówczas

$$|f(x) - p(x)| \leq \frac{1}{2^n (n+1)!} \max_{t \in [-1,1]} |f^{(n+1)}(t)|.$$

• FUNKCJE SKLEJANE

Zadanie interpolacji za pomocą funkcji sklejanych 3-go stopnia możemy sformułować następująco. Dla danych $n+1$ punktów (x_i, y_i) $(a = x_0 < x_1 < \ldots < x_n = b)$ znaleźć funkcję S spełniającą warunki:

1. $S \in C^2[x_0, x_n]$.
2. $S|_{[x_{k-1}, x_k]} \equiv p_k \in \Pi_3 \quad (1 \leq k \leq n)$.
3. $S(x_k) = y_k \quad (0 \leq k \leq n)$.

JEŻELI FUNKCJA S SPEŁNIA:

$$S''(x_0) = S''(x_n) = 0,$$

TO S JEST NATURALNĄ FUNKCJĄ SKLEJANĄ 3-GO STOPNIA

Wyznaczanie funkcji sklejanej 3-go stopnia:

KROKI:

1. OBLICZYĆ ILORAZY RÓŻNICOWE $d_k = 6f[x_{k-1}, x_k, x_{k+1}]$. DLA $k = 1, \ldots, n-1$, gdzie
$$f[x_{k-1}, x_k, x_{k+1}] = \frac{f[x_k, x_{k+1}] - f[x_{k-1}, x_k]}{x_{k+1} - x_{k-1}}$$
2. OBLICZYĆ $\lambda_k = \frac{h_k}{(h_k + h_{k+1})}$ DLA $k = 1, \ldots, n-1$
3. WYZNACZYĆ M_k DLA $k = 1, \ldots, n-1$ ROZWIĄZUJĄC UKŁAD (METODĄ PRZEGNANIA, KTÓRĄ OMÓWIMY PÓŹNIEJ).

$$\begin{bmatrix} 2 & 1-\lambda_1 & & & \\ \lambda_2 & 2 & 1-\lambda_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda_{n-2} & 2 & 1-\lambda_{n-2} \\ & & & \lambda_{n-1} & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \vdots \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_{n-2} \\ d_n \end{bmatrix}$$

MACIERZ POWYŻSZEGO UKŁADU JEST DIAGONALNIE SILNIE DOMINUJĄCA $(2 > |\lambda_k| + |1 - \lambda_k|)$.

STĄD JEST NIEOSOBLIWA.