

Prowadzący: dr hab. L.K. Bieniasz

MATERIAŁY POMOCNICZE

DO WYKŁADÓW

Z PRZEDMIOTU:

„METODY OBliczeniowe”

(Informatyka, I stopień)

Metody obliczeniowe

dotyczą głównie rozwiązywania, za pomocą komputerów cyfrowych, takich zadań matematycznych, w których niezbędne są obliczenia na liczbach rzeczywistych.

Tego typu obliczenia wykonywane są ZAWSZE w sposób przybliżony, tzn. że wyniki obliczeń obarczone są błędami.

Dlatego też, wykonując obliczenia nie możemy poprzestać na uzyskaniu wyniku obliczeń, ale również musimy określić (choćby w przybliżeniu) jak dużym błędem obarczony jest ten wynik.

Do szacowania błędów obliczeń służy m.innymi tzw. ESTYMATORY BŁĘDÓW

Błędы zwykle definiuje się następująco:

Jeśli W = dokładny wynik zadania

$W_{\text{przybl.}}$ = przybliżony wynik zadania

(uzyskany np. za pomocą obliczeń na komputerze)

To:

$$\text{Błąd bezwzględny} \stackrel{\text{df}}{=} \begin{cases} W_{\text{przybl.}} - W & \text{albo} \\ W - W_{\text{przybl.}} & \text{albo} \\ |W_{\text{przybl.}} - W| \end{cases}$$

UWAGA: Studując literaturę należy sprawdzać jakiej definicji używają autorzy!

$$\text{Błąd względny} \stackrel{\text{df}}{=} \begin{cases} \frac{W_{\text{przybl.}} - W}{W} & \text{albo} \\ \frac{W - W_{\text{przybl.}}}{W} & \text{albo} \\ \left| \frac{W_{\text{przybl.}} - W}{W} \right| \end{cases}$$

UWAGA: Tutaj zawsze dzielimy przez W a nie przez $W_{\text{przybl.}}$!

Zauważmy, że do wyznaczenia błędów niezbędna jest znajomość W , a tej na ogół nie mamy. Dlatego też niezbędne są ESTYMATORY BŁĘDÓW

Źródła błędów w obliczeniach komputerowych

- 1) Uproszczenia modeli matematycznych
 - 2) Przybliżona znajomość danych wejściowych
 - 3) Przybliżenia zastosowane w algorytmach,
np. → błędy obcięcia
→ błędy dyskretyzacji
 - 4) Błędy maszynowe → przybliżona reprezentacja liczb rzeczywistych w komputerach
→ błędy wytworzone w działaniach arytmetycznych
- W Informatyce zajmujemy się źródłami 3 i 4.
Pozostałe źródła błędów są przedmiotem zainteresowania innych dziedzin wiedzy.

Przykład blądu obcięcia:

wyznaczenie przybliżenia funkcji $f(x) = e^x = \exp(x)$ poprzez rozwinięcie w szereg.

Wzory Taylora: $f(x) = \begin{cases} \sum_{k=0}^{\infty} \frac{1}{k!} f^{(k)}(0) x^k & \text{lub} \\ \sum_{k=0}^n \frac{1}{k!} f^{(k)}(0) x^k + \frac{1}{(n+1)!} f^{(n+1)}(\xi) x^{n+1} \end{cases}$

gdzie $0 < \xi < x$ dla $x > 0$ lub
 $x < \xi < 0$ dla $x < 0$

Ograniczymy się do $x > 0$

$$f^{(k)}(x) = e^x \quad \text{dla } k=0, 1, \dots$$

$$f^{(k)}(0) = 1$$

$$f^{(n+1)}(\xi) = e^{\xi}$$

Jeśli za przybliżenie $f(x)$ przyjmiemy $\sum_{k=0}^n \frac{1}{k!} f^{(k)}(0) x^k$

to popełniamy **BLĄD OBCIĘCIA** (bezwarunkowy)

który wyniesie

$$\left| e^x - \sum_{k=0}^n \frac{x^k}{k!} \right| = \left| \frac{1}{(n+1)!} e^{\xi} x^{n+1} \right| = \frac{1}{(n+1)!} e^{\xi} |x|^{n+1}$$

$\left| \frac{1}{(n+1)!} e^{\xi} x^{n+1} \right| < \frac{1}{(n+1)!} e^x x^{n+1}$ ponieważ $x > 0$.

A zatem względny BŁĄD OBCIĘCIA

ma oszacowanie:

$$\left| \frac{e^x - \sum_{k=0}^n \frac{x^k}{k!}}{e^x} \right| < \frac{1}{(n+1)!} x^{n+1}$$

np. dla $x=1$ uzyskamy:

$n =$	1	2	3	4	5	6
błęd względny	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{24}$	$\frac{1}{128}$	$\frac{1}{768}$	$\frac{1}{5376}$

To można uznać za ESTYMATORE BŁĘDU OBCIĘCIA, bo daje się wyznaczyć bez znajomości dokładnej wartości e^x

Arytmetyka zmiennoprzecinkowa (i jej konsekwencje - błędy maszynowe)

Podstawowa idea:

Każda liczbę rzeczywistą (niezerową) można przedstawić w postaci

$$X = s \cdot m \cdot \beta^c$$

gdzie $s \in \{-1, 1\}$ ZNAK LICZBY

$$m \in [1, 2) \text{ MANTYSA} = 1 + \sum_{j=1}^{\infty} f_j \beta^{-j}$$

gdzie $f_j \in \{0, 1\}$

$\beta = 2$ PODSTAWA (BAZA)

$c \in$ liczb całkowitych CECHA

Jest to reprezentacja DOKŁADNA, która sugeruje następującej PRZYBLIŻONA, REPREZENTACJĘ LICZB RZECZYWISTYCH

Reprezentacja zmiennopozycinkowa

$x \in \mathbb{R}$

1) dla liczb niezerowych:

$$rd_y(x) = (-1)^e \cdot \underbrace{(1+f)}_{\text{mantysa}} \cdot \underbrace{2^{z-b}}_{\text{cecha}} \quad \text{gdy } z > 0$$

Jest to tzw. LICZBA ZNORMALIZOWANA

lub też

$$rd_y(x) = (-1)^e \cdot \underbrace{(0+f)}_{\text{mantysa}} \cdot \underbrace{2^{0-b+1}}_{\text{cecha}} \quad \text{gdy } z = 0$$

Jest to tzw. LICZBA ZDENORMALIZOWANA

w pow. wzorach

$e \in \{0,1\}$ BIT ZNAKU

$$f = (0.f_1 \cdots f_t)_2 = \sum_{j=1}^t f_j 2^{-j}, \quad f_j \in \{0,1\}$$

(liczba dwójkowa ułamkowa)

BITY MANTYSY

$$z = (z_1 \cdots z_p)_2$$

(liczba dwójkowa naturalna lub zero)

BITY CECHY

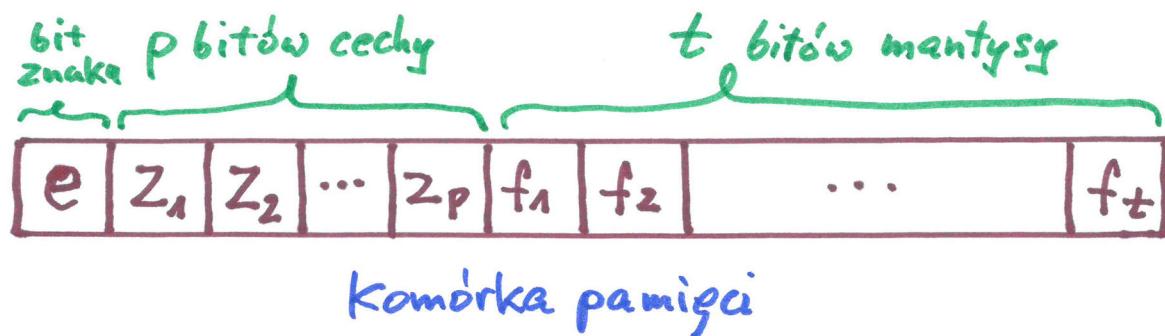
$$z_j \in \{0,1\}$$

b = przesunięcie (bias) stała liczba całkowita dobrana tak by uzyskać porównywanie liczb dodatnich i ujemnych wartości cechy.

t = liczba bitów mantysy

p = liczba bitów cechy

Sposób przechowywania e, f, z



UWAGA : Nie przechowujemy ani cechy ani mantysy, a jedynie

liczby "z" cechy oraz
liczby "f" mantysy.

Nie przechowujemy też przesunięcia "b"

2) Reprezentacja zera rzeczywistego

+0	[0 0	...	[0]
-0	[1 0	...	[0]

lub

Paradoks: mamy zero dodatnie i zero ujemne

3) Reprezentacja nieskończoności

+INF	[0 1	...	[1 0	...	[0]
	P bitów cechy równych 1			t bitów mantysy równych 0	
-INF	[1 1	...	[1 0	...	[0]

4) Reprezentacja symboli NaN (Not a Number), które oznaczają wyniki obliczeń, którym nie da się przypisać wartości liczbowych ani też nieskończoności. Np. wyniki typu $0/0$, INF/INF , $INF-INF$, $0\cdot INF$ etc.

NaN	[0 1	...	[1 0	...	[1 ... 0]
	P bitów cechy równych 1			przynajmniej jeden bit ustawiony na "1"	

Precyza arytmetyki 22

to maksymalny względem błąd reprezentacji
liczby (znormalizowanej)

Niech

$$x = \underbrace{(1.f_1 f_2 \dots)}_m \cdot 2^c > 0 \quad (\text{wartość dokładna } x)$$

Obcięcie dalszych bitów powyżej t (tzn. $t+1, t+2, \dots$) daje

$$x_- = (1.f_1 \dots f_t)_2 \cdot 2^c$$

zaokrąglenie w góry dając

$$x_+ = [(1.f_1 \dots f_t)_2 + 2^{-t}] \cdot 2^c$$

mamy $x_- \leq x < x_+$, $x_+ - x_- = 2^{c-t}$

przyjmujemy zasadę: $\text{rd}_y(x) = \begin{cases} x_- \text{ gdy } |x-x_-| \leq |x_+-x| \\ x_+ \text{ gdy } |x-x_-| > |x_+-x| \end{cases}$

A zatem

$$|\text{rd}_y(x) - x| \leq \frac{1}{2} (x_+ - x_-) = 2^{c-t-1}$$

$$\left| \frac{\text{rd}_y(x) - x}{x} \right| \leq \frac{2^{c-t-1}}{m \cdot 2^c} = \frac{1}{m} 2^{-(t+1)} \leq \boxed{2^{-(t+1)}}$$

Ogólnie zatem

$$\text{rd}_y(x) = x(1+e)$$

↑ błąd względny reprezentacji
|e| ≤ γ

precyza arytmetyki
(zależy tylko od t .)

Standard IEEE 754

(Institute of Electrical and Electronic Engineers)

Literatura:

- 1) E. Wantuch, M. Drabowski, Wstęp do Informatyki, PK, 2006, Kraków.
- 2) D. Goldberg : "What every Computer Scientist should know about Floating Point Arithmetic". www.physics.ohio-state.edu/~dws/grouplinks/floating-point-math.pdf

zmienna	pojedynczej precyzji	podwójnej precyzji
typ "C"	float	double
P	8	11
t	23	52
b	127	1023
zakres (l. znorm.)	$\approx 10^{-38} \dots 10^{+38}$	$\approx 10^{-308} \dots 10^{+308}$
precyzja arytmetyki $\gamma \approx$	$6 \cdot 10^{-8}$	10^{-16}
liczba bajtów	4	8

Są też zmienne rozszerzonej precyzji
(typ "C" long double) ale nie objęte do końca standardem.

Niektóre charakterystyki zmiennych standardu IEEE 754, na przykładzie zmiennych pojedynczej precyzji

1) Najmniejsza i największa liczba znormalizowana (dodatnia)

$$Z \in \{0, \dots, \underbrace{2^8 - 1}_{255}\}$$

ale 255 zarezerwowane na $\pm\infty$ oraz NaN, a 0 zarezerwowane na ± 0 oraz liczby zdenormalizowane.

$$\text{Zatem } Z \in \{1, \dots, 254\}$$

$$\text{cecha } C = \underbrace{Z - b}_{127} \in \{-126, \dots, 0, \dots, 127\}$$

mantysa

$$m = 1 + \sum_{j=1}^{23} f_j 2^{-j} \in \left[1.0, \underbrace{1.0 + \left(\frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{23}}\right)}_{1 - \frac{1}{2^{24}}} \right]$$

$$1 \cdot \frac{1 - \left(\frac{1}{2}\right)^{24}}{1 - \frac{1}{2}} = 2 - 2^{-23}$$

$$\text{czyli } m \in [1, 2 - 2^{-23}]$$

$$\text{największa liczba znormalizowana} = (2 - 2^{-23}) \cdot 2^{127} \approx 3.4 \cdot 10^{+38}$$

$x_{\text{max, znorm}}$

$$\text{najmniejsza liczba znormalizowana} = 1 \cdot 2^{-126} \approx 1.2 \cdot 10^{-38}$$

$x_{\text{min, znorm}}$

2) Najmniejsza i największa liczba zdenormalizowana (dodatnia)

$z=0$ (w pamięci) ale zastępujemy przez $z=1$

Cecha $c = z - \frac{b}{2^e} = -126$

mantysa

$$m = 0 + \sum_{j=1}^{23} f_j 2^{-j} \in \left[2^{-23}, \underbrace{\left(\frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{23}} \right)}_{\substack{\\ \text{}}}, \underbrace{\left(1 + \frac{1}{2} + \frac{1}{4} + \dots + \frac{1}{2^{23}} \right) - 1}_{\substack{\\ \text{}}} \right]$$

czyli

$$m \in \left[2^{-23}, 1 - 2^{-23} \right]$$

$$(2 - 2^{-23}) - 1 = 1 - 2^{-23}$$

największa liczba zdenormalizowana = $(1 - 2^{-23}) \cdot 2^{-126}$

$x_{\max, \text{zdenorm}}$

najmniejsza liczba zdenormalizowana = $2^{-23} \cdot 2^{-126} = 2^{-149}$

$x_{\min, \text{zdenorm}}$

3) Zachodzą relacje:

$$+\infty < x_{\min, \text{zdenorm}} < x_{\max, \text{zdenorm}} < x_{\min, \text{znorm}} < x_{\max, \text{znorm}} < +\infty$$

$$x < x_{\min, \text{zdenorm}} \Rightarrow \text{rnd}_y(x) = +\infty \quad \begin{matrix} \text{niedomiar} \\ \text{underflow} \end{matrix}$$

$$x > x_{\max, \text{znorm}} \Rightarrow \text{rnd}_y(x) = +\infty \quad \begin{matrix} \text{nadmiar} \\ \text{overflow} \end{matrix}$$

Działania arytmetyczne

Oznaczmy $(+ - * /) \equiv \square$

W dobrze zaprojektowanym komputerze, jeśli

$$x = \text{rd}_y(x) \text{ oraz } y = \text{rd}_y(y)$$

to

$$\underbrace{f_{L_y}(x \square y)}_{\substack{\text{wynik działania} \\ \text{zmiennoprzecinkowego}}} = \underbrace{(x \square y)}_{\substack{\text{wynik} \\ \text{dokładny}}} (1 + \delta), \text{ gdzie } |\delta| \leq 2$$

czyli $f_{L_y}(x \square y) \approx \text{rd}_y(x \square y)$

(błąd działania powinien być na poziomie błędu reprezentacji dokładnego wyniku)

Uwaga: Tak jest tylko przy założeniu, że nie wystąpi nadmiar lub niedomiar.

NADMIAR \Rightarrow wyniku nie da się reprezentować

NIEDOMIAR \Rightarrow błąd względny wynosi 100%

Należy też zachować ostrożność przy odejmowaniu (otym później)

Należy pamiętać, że w obliczeniach zmiennoprzecinkowych:

Działania są przemienne ($+, *$): $f_{L_y}(x \square y) = f_{L_y}(y \square x)$

ale

nie są łączne ($+, *$): $f_{L_y}((x \square y) \square z) \neq f_{L_y}(x \square (y \square z))$

i nie są rozdzielne: $f_{L_y}((x+y)*z) \neq f_{L_y}(x*z + y*z)$

Jeśli natomiast $x \neq \tau d_y(x)$ i $y \neq \tau d_x(y)$

to

$$f_{L_y}(\tau d_y(x) \square \tau d_y(y)) = (x(1+e_x) \square y(1+e_y))(1+\delta)$$

Ocena błędów maszynowych powstających
w bardziej złożonych obliczeniach to trudny
problem

Stosuje się :

- 1) Analizę teoretyczną (bardzo pracochłoną)
— przykłady poznamy później
- 2) Arytmetykę przedziałową. Polega to na zastąpieniu działań na liczbach działaniami na przedziałach liczbowych, przy użyciu stasowych bibliotek. Jest to metoda kosztowna obliczeniowo.

UWAGA :

Żaden komputer nie daje gwarancji, że wyniki naszych obliczeń są nieobarzone (dużymi błędami maszynowymi). Nie jest też prawda, że duże błędy maszynowe powstają tylko przy odpowiednio złożonych obliczeniach. Można napisać program złożony z jednej instrukcji, który daje całkowicie niepoprawny wynik (chodzi o utratę cyfr znaczących przy odejmowaniu, o której powiemy później).

Własności Zadań

Typowe zadanie jakie chcemy rozwiązać wygląda następująco:

Dla danych obliczyć wynik

$$\vec{d} = [d_1, d_2, \dots, d_n]^T$$

$$\vec{W} = [w_1, w_2, \dots, w_m]^T$$

taki, że

$$\vec{W} = \vec{F}(\vec{d})$$

Podstawową właściwością zadań, istotną w obliczeniach jest ich UWARUNKOWANIE.

Uwarunkowanie zadania dotyczy wrażliwości wyniku na zaburzenie danych.

Zadanie nazywamy źle uwarunkowanym, jeśli niewielkie względne zmiany danych powodują duże względne zmiany wyniku.

Zadanie nazywamy dobrze uwarunkowanym, jeśli niewielkie względne zmiany danych powodują niewielkie względne zmiany wyniku.

W zasadzie, na komputerze możemy rozwiązywać jedynie zadania dobrze uwarunkowane, gdyż spodziewamy się danych obarczonych błędami.

Ilościowo, uwarunkowanie zadania określają WSKAŻNIKI UWARUNKOWANIA.

Zazwyczaj względne zaburzenia danych i wyniku są ze sobą powiązane różnymi wzorami. Na przykład:

$$\frac{|\vec{\Delta \tilde{W}}|}{|\vec{W}|} = \underbrace{\text{cond}(\vec{W})}_{\substack{\text{względne zaburzenie} \\ \text{wyniku}}} \cdot \frac{|\vec{\Delta d}|}{|\vec{d}|}$$

wskaźnik uwarunkowania

względne zaburzenie danych

Jeśli $\text{cond}(\vec{W}) \approx 1$ lub jest < 1 to zadanie będzie dobre uwarunkowane (well conditioned)

Jeśli $\text{cond}(\vec{W}) > 1$ to zadanie będzie źle uwarunkowane (ill conditioned)

Jeśli (dla pewnych dopuszczalnych danych)

$\text{cond}(\vec{W}) \rightarrow \infty$ to zadanie jest źle postawione (ill posed)

UWAGA: Wartość "1" można tu zastąpić przez inną niewielką liczbę - to może być 2, 10 czy 100

Właściwości algorytmów obliczeniowych

1) Numeryczna poprawność

Algorytm numerycznie poprawny, to taki algorytm, który daje rozwiązanie (zadania) będące nieco zaburzonym dokładnym rozwiązaniem zadania o nieco zaburzonych danych

(nieco zaburzone \Leftrightarrow zaburzone na poziomie błędu reprezentacji liczb rzeczywistych)

Jest to najbardziej pożądana ale rzadko spotykana właściwość algorytmów obliczeniowych

2) Numeryczna stabilność

Jest to niezbędna właściwość algorytmów wieloetapowych



Etap k dostaje dane $W_{k-1} + E_{k-1}$, które są wynikiem obliczeń etapu $k-1$, obarczonym błędem E_{k-1} , i produkuje wynik W_k obarczony błędem E_k . Jeżeli błędy są niewielkie, to mamy związek liniowy pomiędzy błędami:

$$E_k = g_k \cdot E_{k-1} + E_k^*$$

gdzie

- E_k^* - błąd wytworzony w etapie k
- E_{k-1} - błąd przeniesiony z etapu $k-1$

$g_k \rightarrow$ to tzw. współczynnik wzmacniania błędu

Algorytm jest numerycznie stabilny, jeżeli
będzie przeniesiony z poprzedniego etapu jest
zmieniony, tzn. jeśli $|g_k| \leq 1$

Algorytmy numerycznie niestabilne

NIE NADAJĄ SIĘ DO UŻYTKU !

Zachodzi implikacja:

Numeryczna poprawność \Rightarrow Numeryczna stabilność

Ale nie na odwrót !

3) Złożoność obliczeniowa

Zazwyczaj preferujemy algorytmy o możliwie najmniejszej
złożoności, tak aby algorytm dawał możliwie jak
najlepsze wyniki możliwie jak najmniejszym kosztem
obliczeniowym.

Przykłady analizy teoretycznej zadań obliczeniowych, i błędów przy rozwiązywaniu tych zadań

1

Obliczanie sumy 2 liczb : $S = X + Y$

Zbadamy uwarunkowanie tego zadania, a jednocześnie oszacujemy błąd maszynowy wyniku, w przypadku sumowania na komputerze.

$$f_{L_2}(s) = f_{L_2}(rd_2(x) + rd_2(y)) = [x(1+e_x) + y(1+e_y)](1+\delta)$$

$$|f_{L_2}(s) - s| = |x(1+e_x)(1+\delta) + y(1+e_y)(1+\delta) - x - y|$$

$$\text{ale } (1+e_x)(1+\delta) \approx 1+e_x+\delta$$

$$(1+e_y)(1+\delta) \approx 1+e_y+\delta \quad \text{zatem}$$

$$\begin{aligned} |f_{L_2}(s) - s| &\approx |x(e_x+\delta) + y(e_y+\delta)| \leq |x| \cdot |e_x+\delta| + |y| \cdot |e_y+\delta| \\ &\leq |x| \cdot (|e_x| + |\delta|) + |y| \cdot (|e_y| + |\delta|) \\ &\leq (|x| + |y|) \cdot \max \{ |e_x| + |\delta|, |e_y| + |\delta| \} \end{aligned}$$

$$\left| \frac{f_{L_2}(s) - s}{s} \right| \leq \underbrace{\frac{|x| + |y|}{|x + y|}}_{\substack{\text{względne} \\ \text{zaburzenie} \\ \text{wyniku}}} \cdot \underbrace{\max \{ |e_x| + |\delta|, |e_y| + |\delta| \}}_{\substack{\text{względne} \\ \text{zaburzenie} \\ \text{danych}}} \cdot \underbrace{\text{cond}(s)}_{\substack{\text{wskaźnik} \\ \text{uwarunkowania}}}$$

Jeśli x, y są tego samego znaku, to $\text{cond}(s) = 1$, czyli zadanie jest dobrze uwarunkowane.

$$\begin{aligned} \text{Z kolei } |e_x| &\leq 2 \\ |e_y| &\leq 2 \\ |\delta| &\leq 2 \end{aligned}$$

A zatem

$$\left| \frac{f_{L_2}(s) - s}{s} \right| \lesssim 22$$

co daje ocenę błędu maszynowego przy rozwiązyaniu zadania na komputerze.

② Obliczanie różnicy 2 liczb: $\tau = x - y$

$$f_{L_2}(\tau) = f_{L_2}(\tau d_{L_2}(x) - \tau d_{L_2}(y)) = [x(1+e_x) - y(1+e_y)](1+\delta)$$

$$|f_{L_2}(\tau) - \tau| = |x(1+e_x)(1+\delta) - y(1+e_y)(1+\delta) - (x-y)|$$

$$\approx |x(e_x + \delta) - y(e_y + \delta)| =$$

$$= |x(e_x + \delta) + (-y)(e_y + \delta)|$$

$$\leq |x| \cdot (|e_x| + |\delta|) + |y| \cdot (|e_y| + |\delta|)$$

$$\leq (|x| + |y|) \cdot \max \{ |e_x| + |\delta|, |e_y| + |\delta| \}$$

$$\left| \frac{f_{L_2}(\tau) - \tau}{\tau} \right| \leq \frac{|x| + |y|}{|x - y|} \cdot \max \{ |e_x| + |\delta|, |e_y| + |\delta| \}$$

względne
zaburzenie
wyniku

wskaznik
uwarunkowania
 $\text{cond}(\tau)$

względne
zaburzenie
danych

$$\text{Mamy } \text{cond}(r) = \frac{|x| + |y|}{|x - y|}$$

a więc dla $x \approx y$ $\text{cond}(r) \gg 1$ czyli
zadanie jest **ZLE UWARUNKOWANE!**

Ponadto błąd maszynowy wyniku może być
dowolnie duży.

Ze wszystkich działań arytmetycznych jedynie
ODEJMOWANIE wykazuje że uwarunkowanie,
przy odejmowaniu **PORÓWNYWALNYCH** liczb
($x \approx y$).

W przypadku odejmowania reprezentacji zmienoprzecinkowych
prowadzi to do tzw. **UTRATY CYFR ZNACZĄCYCH**
PRZY ODEJMOWANIU

Dlatego też, w programach numerycznych należy unikać
odejmowania, jeśli to tylko możliwe.

③ Uwarunkowanie zadania obliczania wartości funkcji $f(x)$.

Zakładały, że funkcja jest różniczkowalna.

Jeśli $y = f(x)$ to z własności różniczki:

$$\Delta y = f'(x) \Delta x$$

A zatem

x pełni rolę danych
 y pełni rolę wyniku
 $\Delta x, \Delta y$ - zaburzenia danych i wyniku

$$\frac{\Delta y}{y} = \frac{f'(x) \Delta x}{f(x)} = \frac{f'(x)}{f(x)} \times \frac{\Delta x}{x}$$

$$\left| \frac{\Delta y}{y} \right| = \left| \frac{f'(x) x}{f(x)} \right| \cdot \left| \frac{\Delta x}{x} \right|$$

względne
zaburzenie
wyniku

wskaznik
uwarunkowania
 $\text{cond}(f(x))$

względne
zaburzenie
danych

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

Przykład: $f(x) = x \Rightarrow f'(x) = 1 \Rightarrow \text{cond}(f(x)) = \left| \frac{1 \cdot x}{x} \right| = 1$
 a więc zadanie jest dobrze uwarunkowane.

Rozwiązywanie
algebraicznych
równań nieliniowych

Definicje

Postać ogólna równania algebraicznego nieliniowego z jedną niewiadomą:

$$f(x) = 0$$

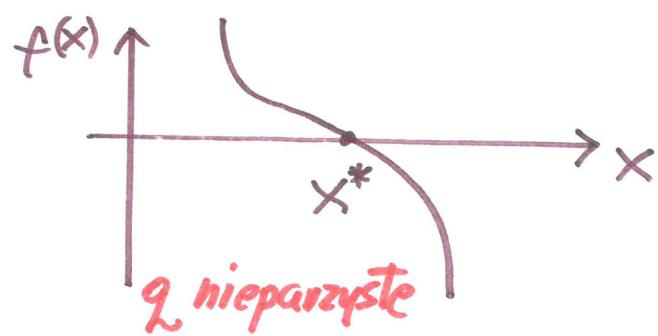
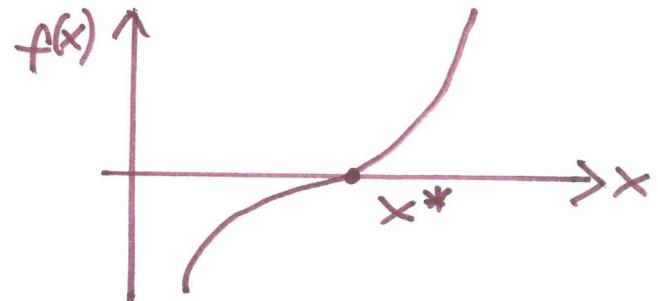
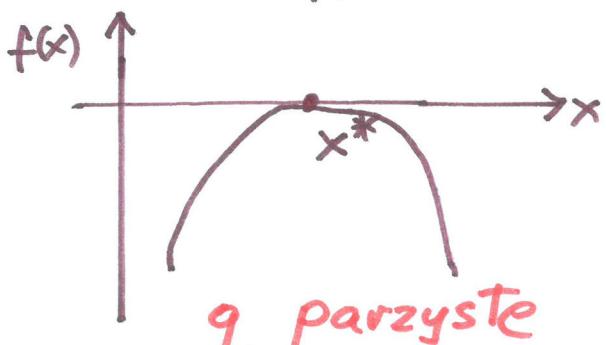
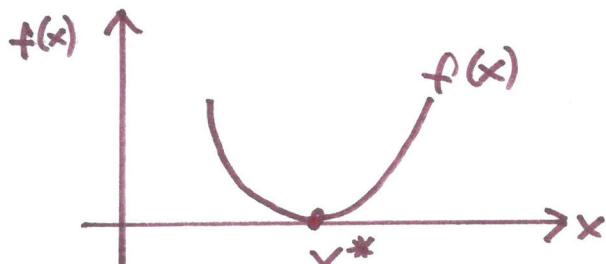
$$x \in \mathbb{R}, f(x) \in \mathbb{R}$$

↗ dowolna funkcja

Jeżeli dla pewnego $x^* \in \mathbb{R}$ mamy $f(x^*) = 0$
to x^* nazywamy **PIERWIASTKIEM RÓWNANIA**
lub też **MIEJSCEM ZEROVYM FUNKCJI** $f(x)$

Pierwiastek x^* nazywamy **q -krotnym**
jeżeli w pewnym otoczeniu x^* zachodzi

$$f(x) \approx (x-x^*)^q \cdot g(x) \quad \text{gdzie } 0 < |g(x)| < \infty$$



Metody iteracyjne

Ogólna Koncepcja:

Tworzymy ciąg kolejnych przybliżeń x_0, x_1, \dots pierwiastka x^* które wyliczamy z pewnego wzoru, startując z jednego lub kilku przybliżeń początkowych.

Wzór do obliczania kolejnych przybliżeń:

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

ITERACJE
JEDNOARGUMENTOWE

lub

$$x_{n+1} = \bar{\Phi}(x_n, x_{n-1}, \dots, x_{n-k})$$

ITERACJE
WIELOARGUMENTOWE

Funkcja definiująca
daną metodę iteracyjną.

Na ogólną zależy ona też od funkcji $f(x)$!

UWAGA:

Należy rozróżnić $f(x)$ i $\Phi(x)$



Problem zbieżności metod iteracyjnych

(Kiedy $x_n \xrightarrow{n \rightarrow \infty} x^*$?)

Obserwacja:

Dla iteracji jednoargumentowych

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

Jeśli iteracje są zbieżne to

$$\begin{array}{c} \downarrow \quad \quad \quad \downarrow \\ x^* = \underline{\Phi}(x^*) \end{array}$$

Czyli x^* jest punktem STĄSYM odwzorowania

$$\underline{\Phi}(x)$$

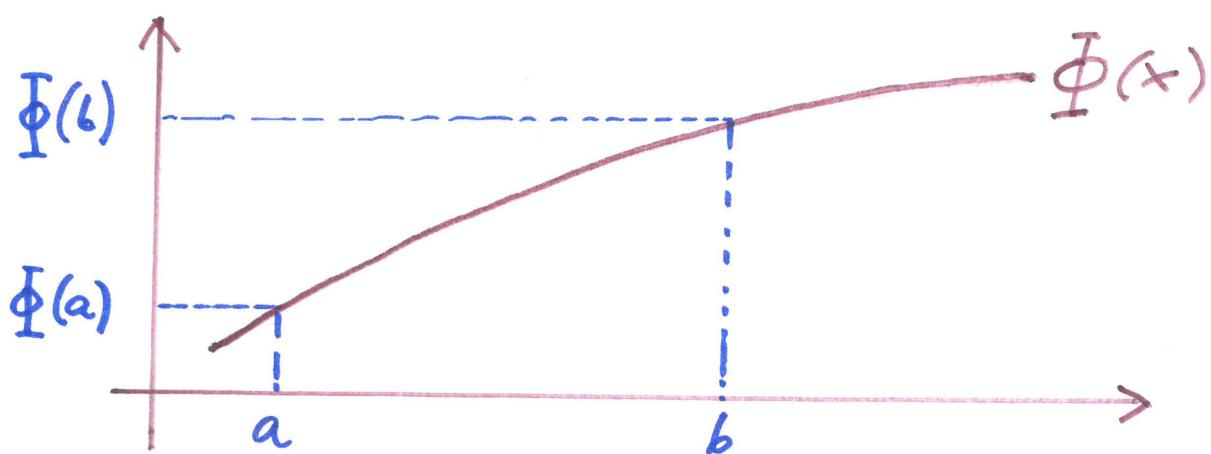
Odwzorowanie zwężające (definicja)

$\Phi(x)$ jest zwężające \Leftrightarrow

$$\forall \gamma \in [0,1] \quad \text{takie, że} \quad \bigwedge_{x,y \in D} \quad |\Phi(x) - \Phi(y)| \leq \gamma \cdot |x-y|$$

↗ Dziedzina $\Phi(x)$

Ilustracja graficzna:



Przedział $[\Phi(a), \Phi(b)]$ jest "węższy" od przedziału $[a, b]$

Twierdzenie Banacha o kontrakcji

Niech C będzie podzbiorem domkniętym R .
Jeśli $\Phi(x)$ jest odwzorowaniem zwężającym
zbioru C w siebie, to $\Phi(x)$ ma
JEDYNY PUNKT STĄTY. Ponadto ten
punkt stały jest granicą każdego ciągu
 $x_{n+1} = \Phi(x_n)$, dla $n=0, 1, \dots$ z punktu
początkowego $x_0 \in C$

WYNIÓSEK:

Aby metoda iteracyjna $x_{n+1} = \Phi(x_n)$
była zbieżna, funkcja $\Phi(x)$ musi być
odwzorowaniem zwężającym

Sprawdzanie zwężalności $\Phi(x)$
w szczególnym przypadku gdy $\Phi(x)$ jest
różniczkowalna

Σ rozwinięcia w szereg:

$$\Phi(y) = \Phi(x) + \Phi'(x)(y-x)$$

gdzie ξ leży między x i y

$$|\Phi(y) - \Phi(x)| = |\Phi'(\xi)| \cdot |y-x|$$

Jeśli $|\Phi'(\xi)| < 1$ dla dowolnego ξ
to $\Phi(x)$ jest zwężającej i mamy zbieżność
iteracji

① Metoda iteracji prostych (Picarda)

Równanie $f(x) = 0$ przekształcamy do postaci $x = \Phi(x)$, o ile to możliwe, i do iteracji stosujemy funkcję $\Phi(x)$ tak uzyskaną.

Przykład 1

$$e^{-x} - \tan x = 0$$

$$x \in \left(-\frac{\pi}{2}, \frac{\pi}{2}\right)$$

$$e^{-x} = \tan x$$

$$\ln e^{-x} = \ln(\tan x)$$

$$x = -\ln(\tan x)$$

$\underbrace{\quad}_{\Phi(x)}$

Przykład 2

$$e^{-x} - \tan x = 0$$

$$\tan x = e^{-x}$$

$$\arctan(\tan x) = \arctan(e^{-x})$$

$$x = \arctan(e^{-x})$$

$\underbrace{\quad}_{\Phi(x)}$

Przykład 3 (ogólny)

$$f(x) = 0$$

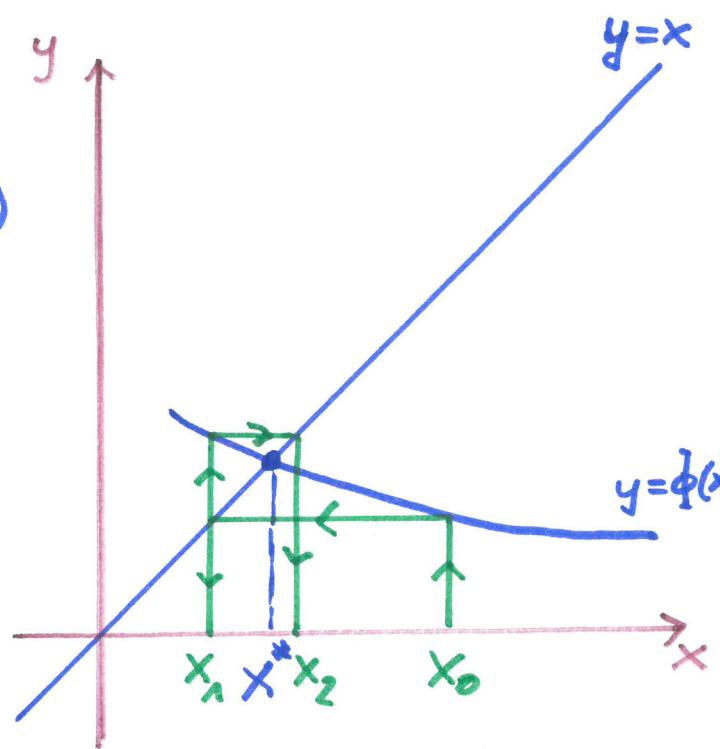
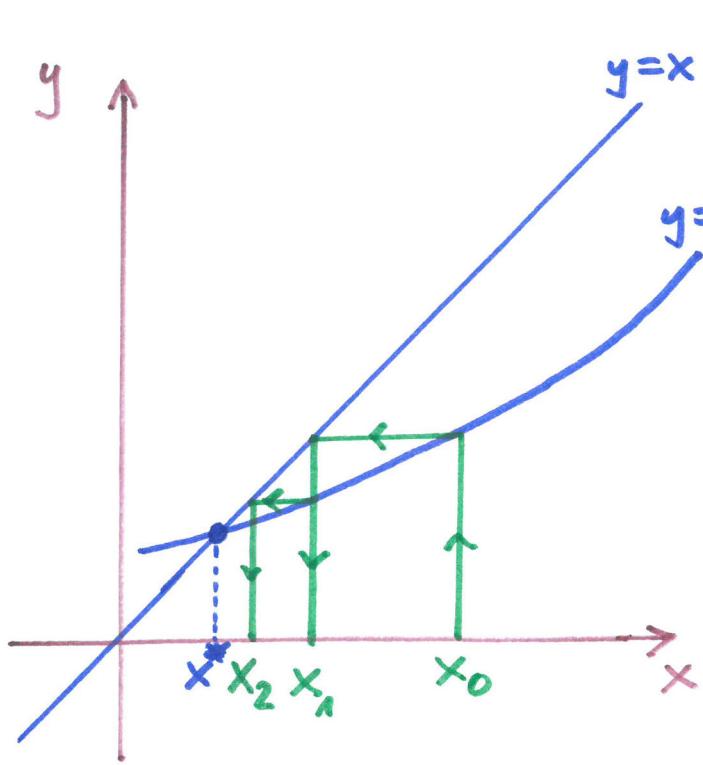
$$f(x) + x - x = 0$$

$$x = f(x) + x$$

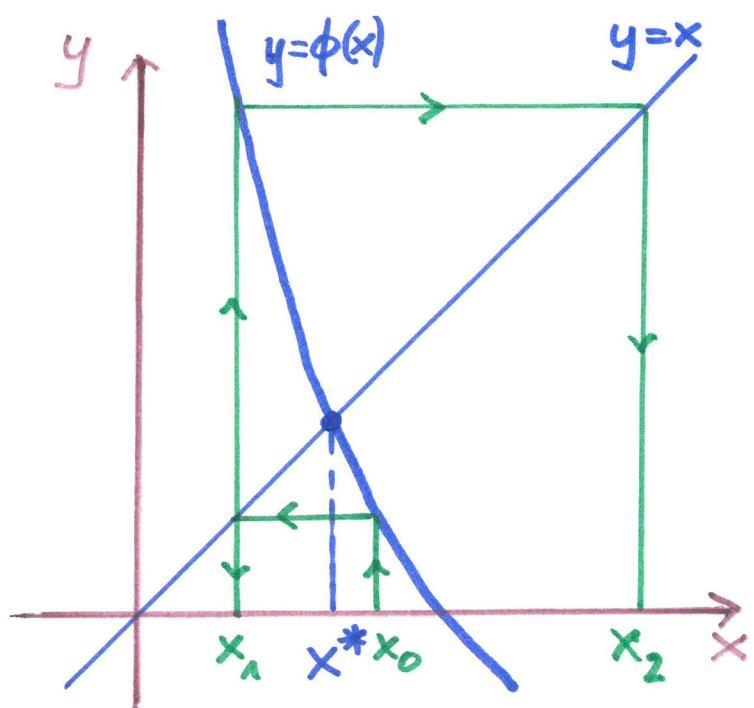
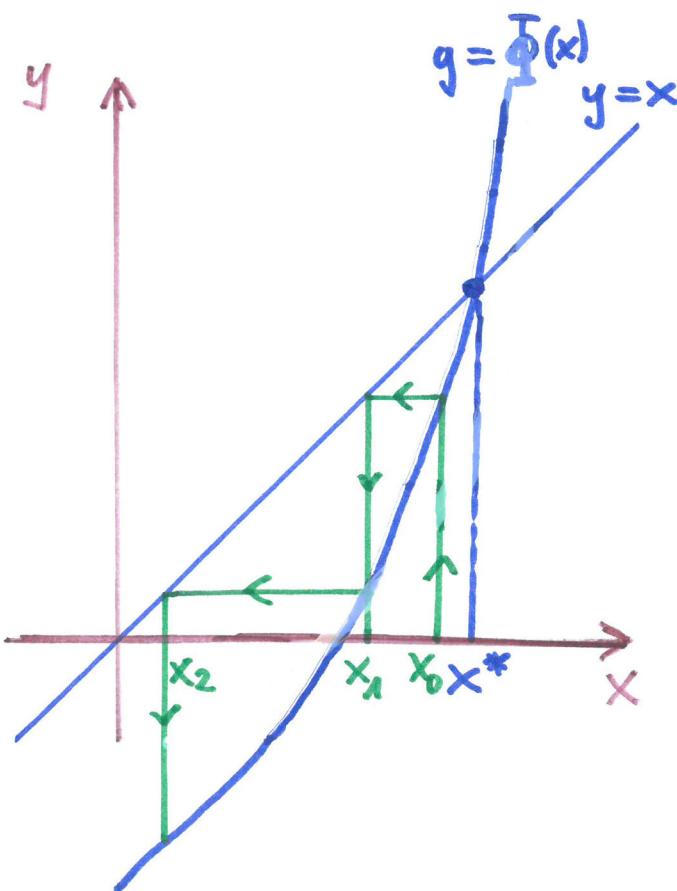
$\underbrace{\quad}_{\Phi(x)}$

Możliwe sytuacje jeśli $\Phi(x)$ różniczkowalna

$|\Phi'(x)| < 1 \Rightarrow$ ZBIEŻNOŚĆ



$|\Phi'(x)| > 1 \Rightarrow$ ROZBIEŻNOŚĆ



2 Metoda bisekcji

Założenia: $f(x)$ ciągła w przedziale $[a, b]$

$f(a)$ i $f(b)$ różnego znaku

(musi istnieć conajmniej jeden pierwiastek w przedziale $[a, b]$)

Procedura:

Zaczynając od przedziału $[a, b]$, dzielimy go na dwie połówki, i wybieramy ten podprzedział dla którego $f(x)$ jest różnego znaku na końcach. Dzielimy go ponownie, etc. Powstaje ciąg podprzedziałów $[a_n, b_n]$.

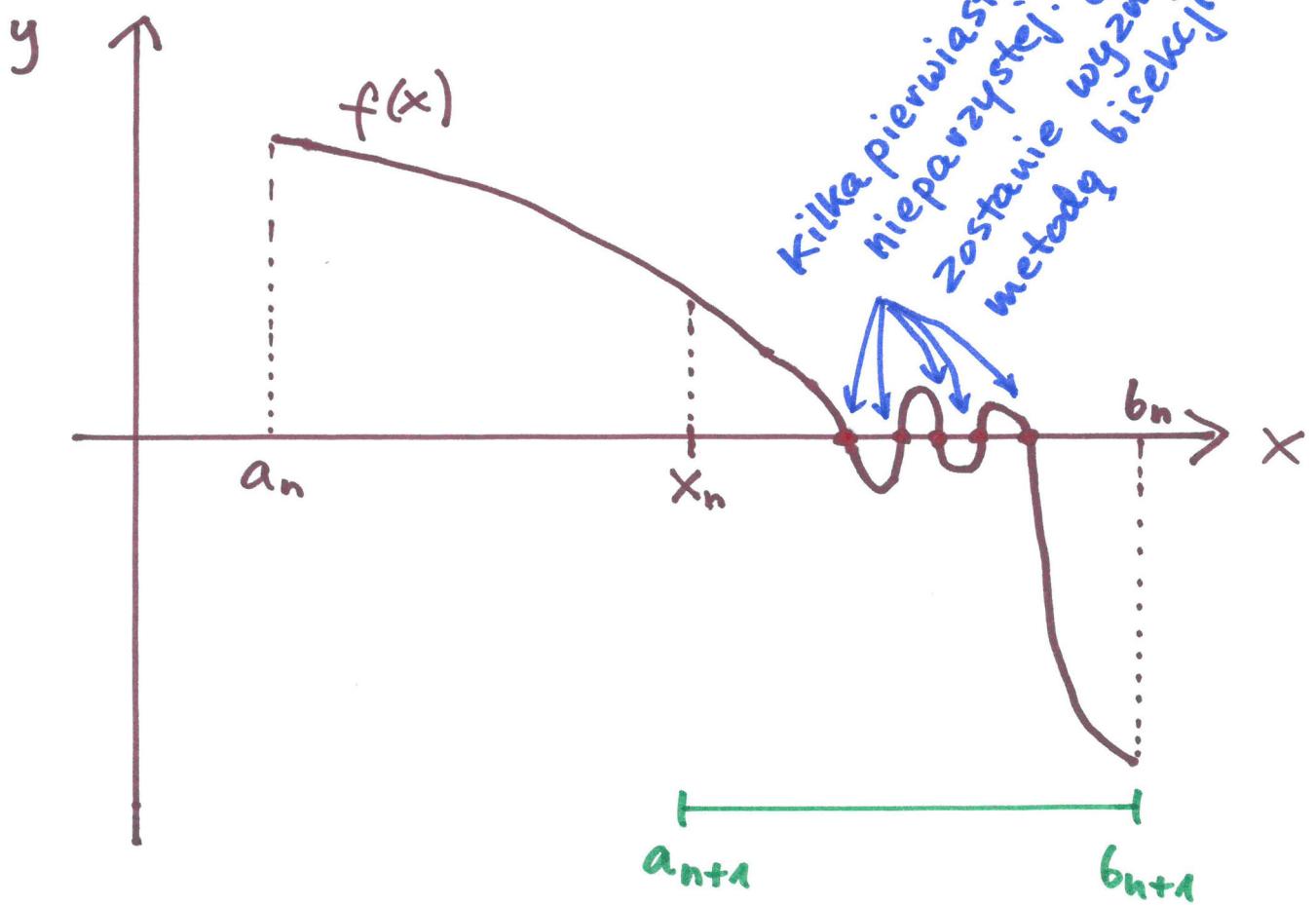
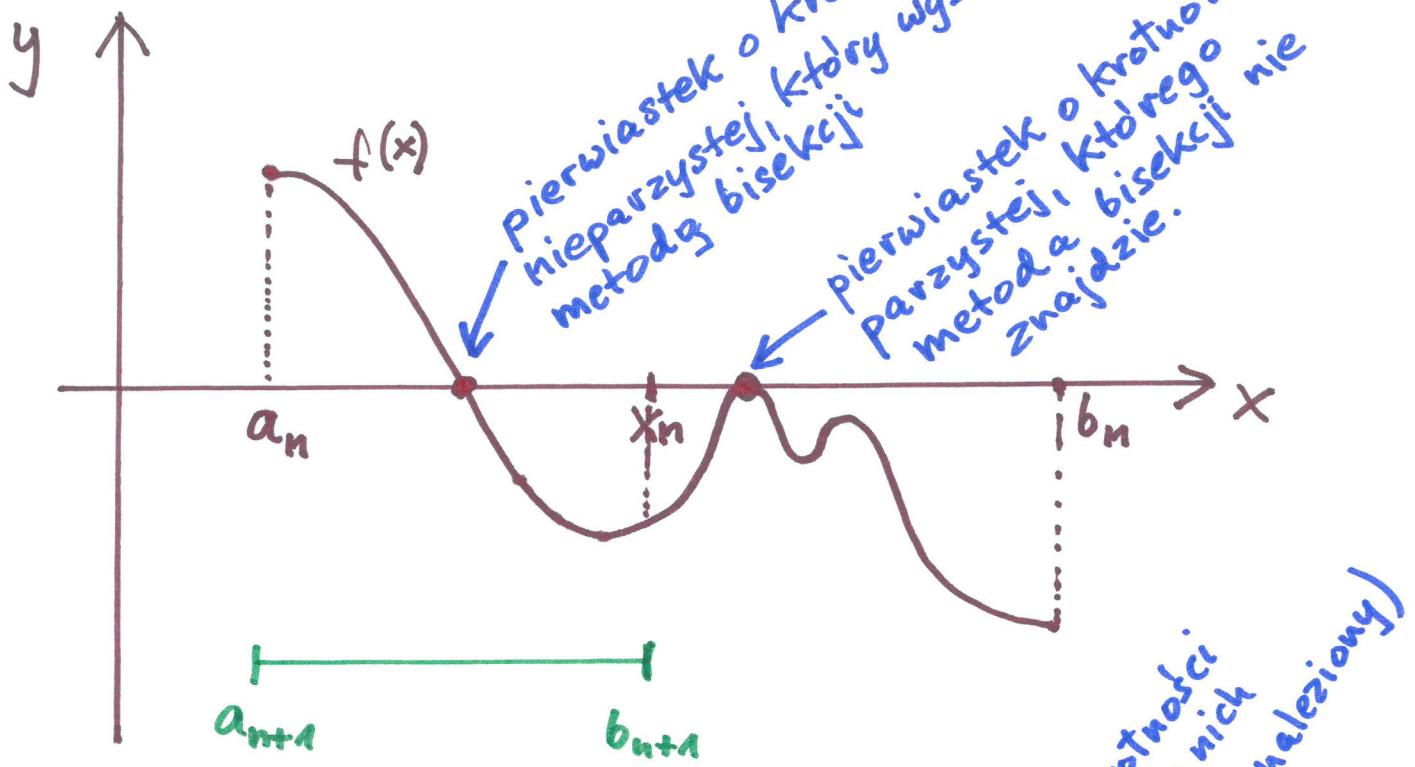
Aktualne przybliżenie pierwiastka to

$$x_n = \frac{a_n + b_n}{2} \quad \left(\begin{array}{l} \text{współrzędna} \\ \text{środka podprzedziału} \end{array} \right)$$

Estymator błędu pierwiastka to

$$e_n = \frac{b_n - a_n}{2} \quad \left(\begin{array}{l} \text{połowa długości} \\ \text{podprzedziału} \end{array} \right)$$

Przykładowe sytuacje:

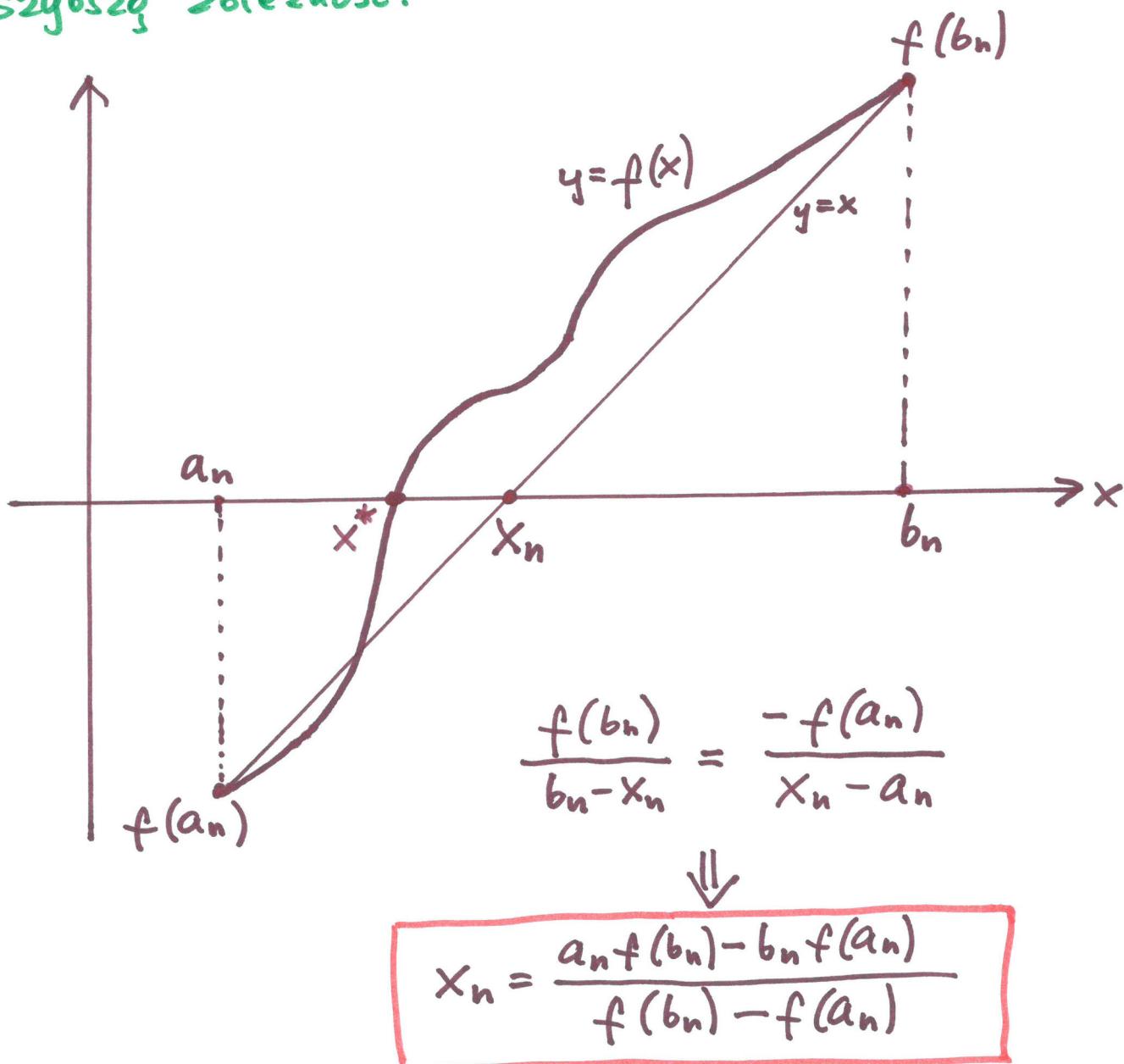


③ Regula Falsi

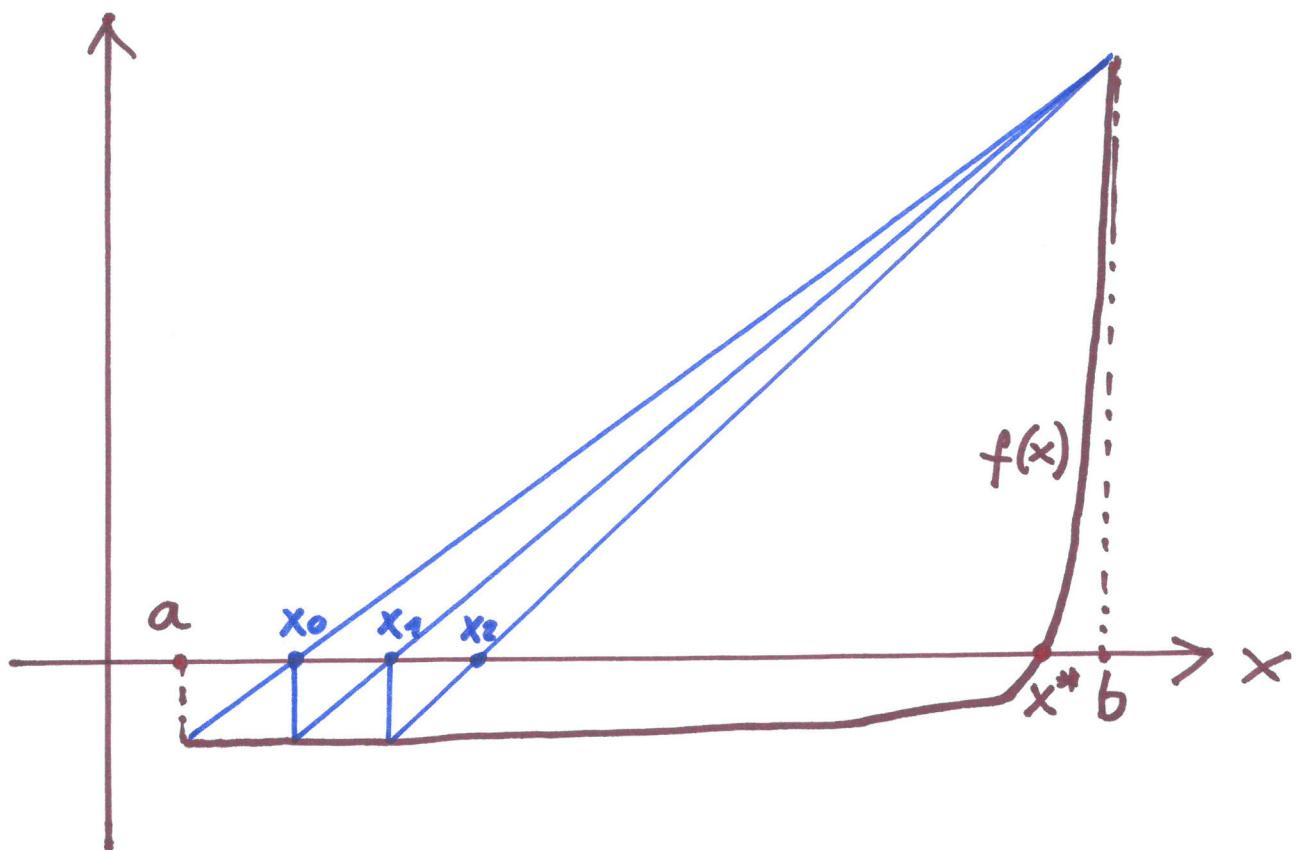
(regula \equiv linia, falsus \equiv fałszywy)

Metoda fałszywego założenia liniowości funkcji

Podobna do bisekcji, ale punkt podziału wyznaczany jest w inny sposób, w nadziei na szybszą zbieżność.



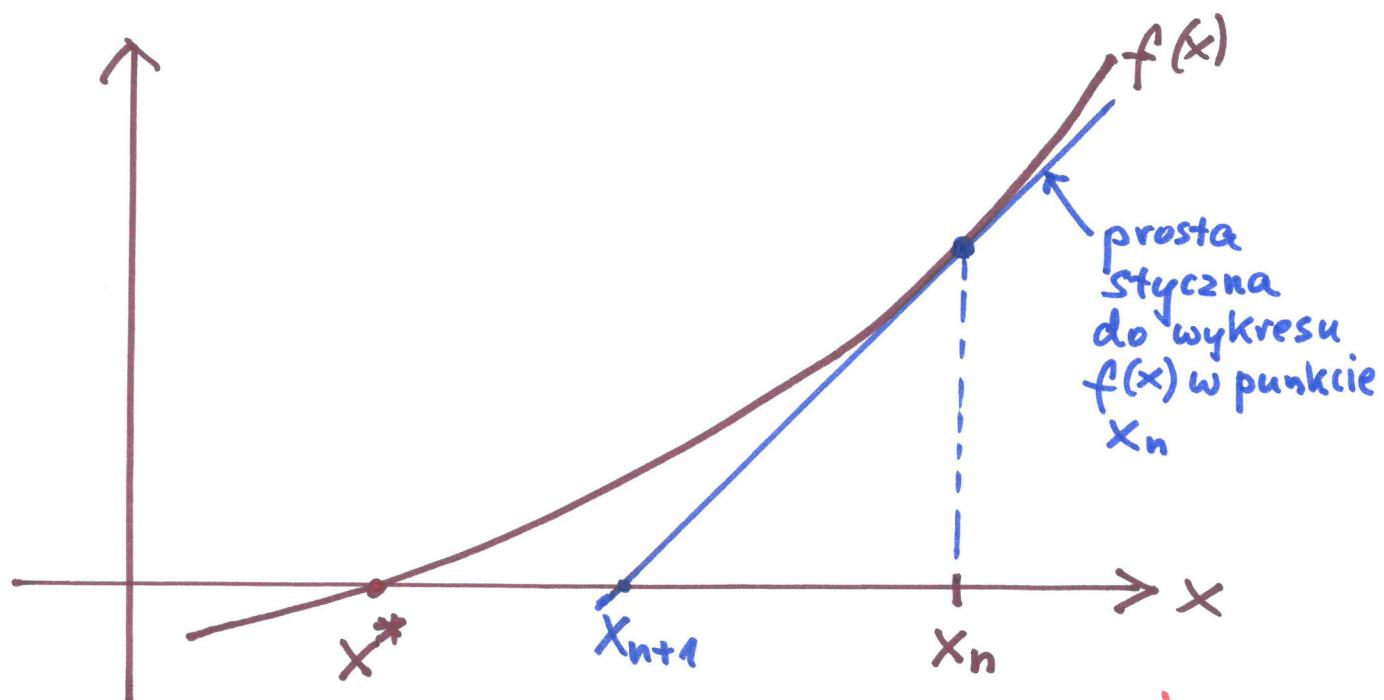
Przykład 6. wolnej zbieżności
metody regula falsi



④ Metoda Newtona (metoda stycznych)

Założenia: $f(x)$ conajmniej różniczkowalna

Koncepcja: Funkcję $f(x)$ liniaryzujemy wokoło x_n i znajdujemy miejsce zerowe uzyskanej funkcji liniowej, jako następne przybliżenie.



Z rozwinięcia w szereg (dalej wyrazy pomijamy):

$$y \approx f(x_n) + f'(x_n)(x - x_n)$$

$$y=0 \Rightarrow x = x_{n+1}, \text{ a zatem}$$

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n)$$



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

czyli że

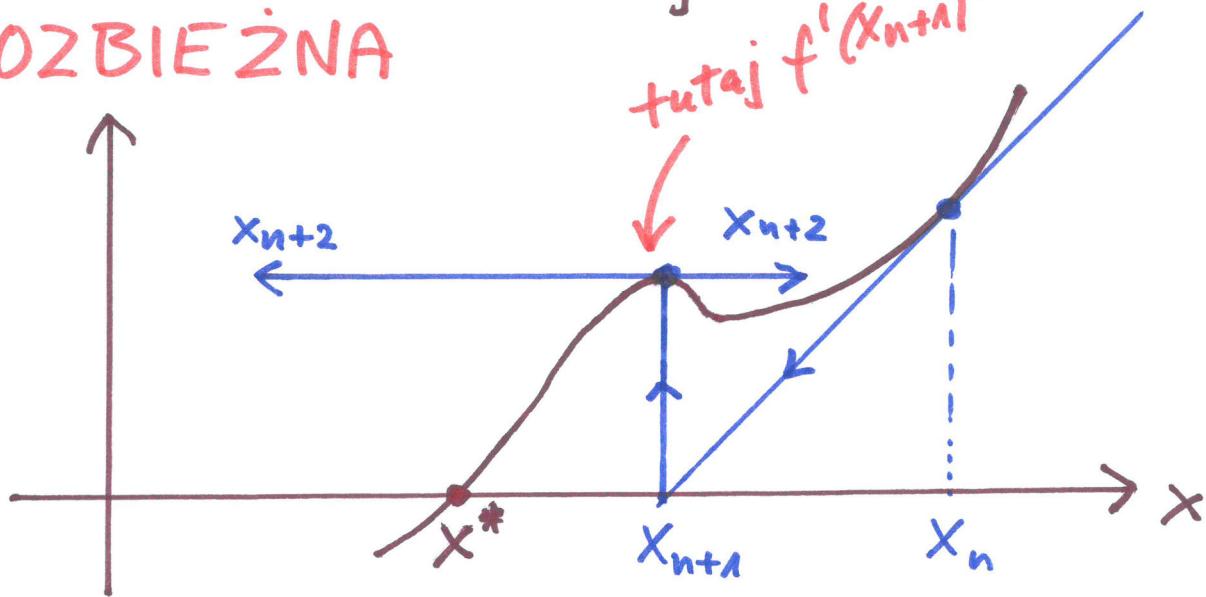
$$\phi(x) = x - \frac{f(x)}{f'(x)}$$

↑ UWAGA! nie może być
 $f'(x_n) = 0$

Potencjalne trudności

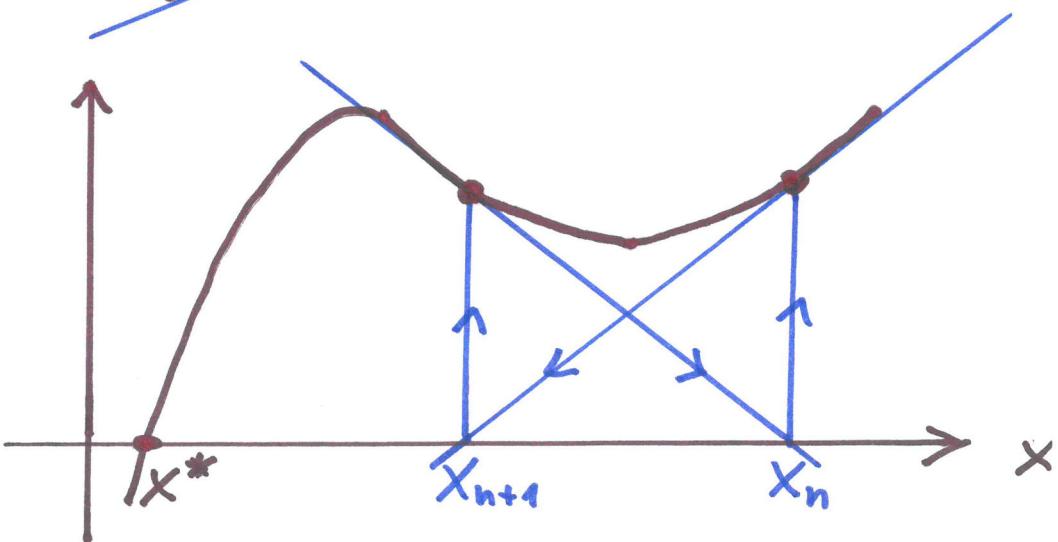
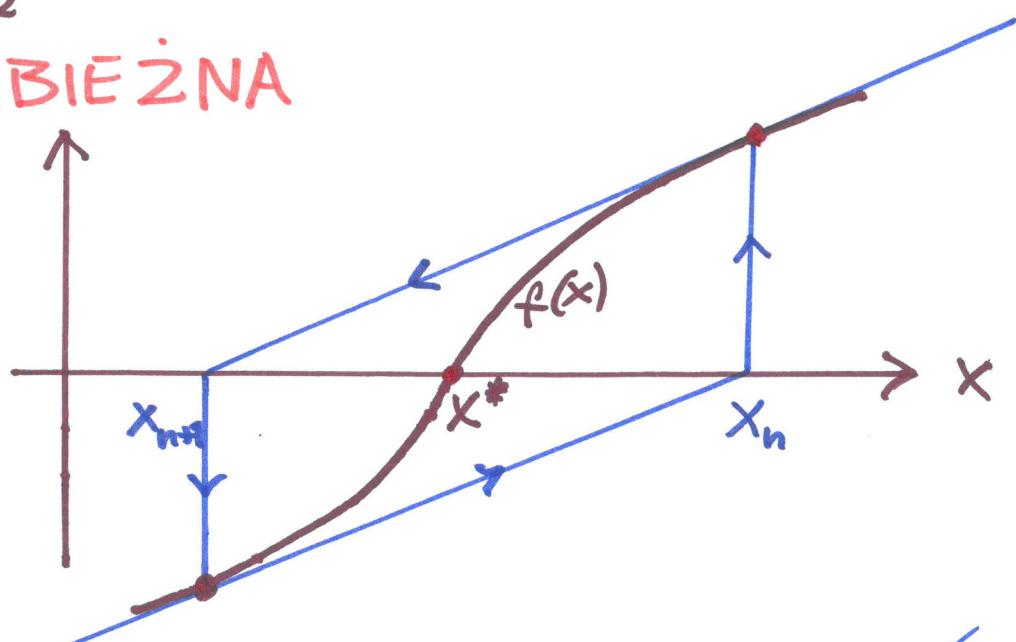
Metoda Newtona może być:

ROZBIEŻNA



lub też

NIEZBIEŻNA

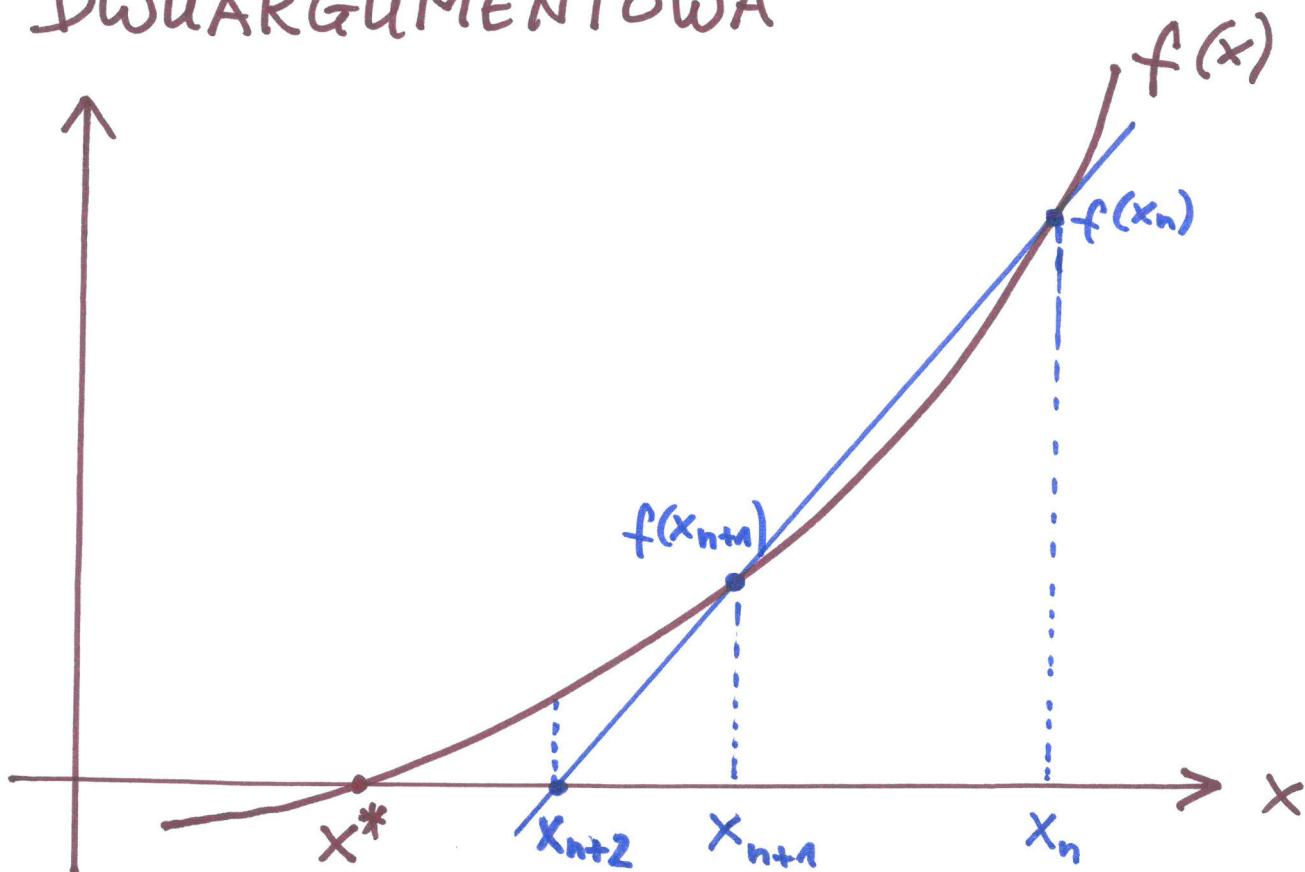


Zaleta: $f(x)$ nie musi
być różniczkowalna

5 Metoda siecznych

podobna do metody Newtona, ale zamiast prostych stycznych stosujemy proste sieczne.

Dlatego jest to metoda iteracyjna
DWUARGUMENTOWA



$$\frac{f(x_n) - 0}{x_n - x_{n+2}} = \frac{f(x_{n+1}) - 0}{x_{n+1} - x_{n+2}}$$



$$x_{n+2} = x_{n+1} - \frac{f(x_{n+1})}{\left[\frac{f(x_{n+1}) - f(x_n)}{x_{n+1} - x_n} \right]}$$

wzór podobny jak
w metodzie Newtona
jeśli przyjąć że to

jest przybliżenie $f'(x_{n+1})$

Kryteria zakończenia iteracji

1. Arbitralne ograniczenie na liczbę iteracji
 $(n \leq n_{\max})$

2. Kryterium dokładności wyznaczenia X_n

$$|e_n| \leq TOLX$$

\uparrow estymator błędu X_n \uparrow zadana tolerancja błędu

Dla bisekcji

$$e_n = \frac{b_n - a_n}{2}$$

Dla pozostałych metod

$$e_n = X_n - X_{n-1}$$

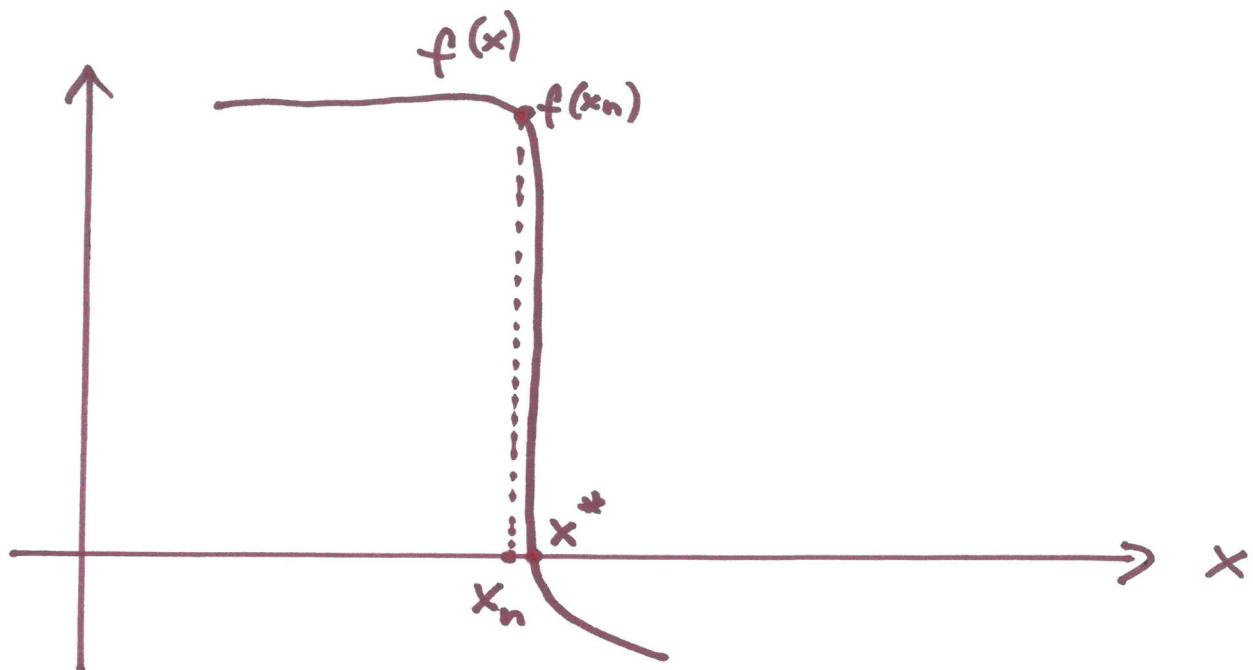
3. Kryterium wiarygodności X_n jako przybliżenia pierwiastka

$$|\underbrace{f(x_n)}_{\text{Reziduum}}| \leq TOLF$$

\uparrow zadana tolerancja reziduum.

równania nieliniowego
(idealnie chcemy uzyskać $f(x_n) = 0$)

Przykład sytuacji, w której kryterium 3 jest niezbędne



Mamy x_n bardzo bliskie x^* , więc wydaje się że x_n jest dobrym przybliżeniem pierwiastka, ale $|f(x_n)|$ jest bardzo odległe od zera

!

Układy algebraicznych równań nieliniowych

$$\vec{f}(\vec{x}) = \vec{0}$$

$$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}, \quad \vec{f}(\vec{x}) = \begin{bmatrix} f_1(\vec{x}) \\ \vdots \\ f_N(\vec{x}) \end{bmatrix} = \begin{bmatrix} f_1(x_1 \dots x_N) \\ \vdots \\ f_N(x_1 \dots x_N) \end{bmatrix}, \quad \vec{0} = \begin{bmatrix} 0 \\ \vdots \\ 0 \end{bmatrix}$$

Uogólniona metoda Newtona dla układów

Linearizacja wokół \vec{x}_n :

$$\vec{y} \approx \vec{f}(\vec{x}_n) + \bar{\bar{J}}(\vec{x}_n)(\vec{x} - \vec{x}_n), \quad \text{gdzie}$$

$$\bar{\bar{J}}(\vec{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \dots & \frac{\partial f_1}{\partial x_N} \\ \vdots & & \vdots \\ \frac{\partial f_N}{\partial x_1} & \dots & \frac{\partial f_N}{\partial x_N} \end{bmatrix} \quad \text{macierz Jacobiego}$$

$$\vec{y} = \vec{0} \Rightarrow \vec{f}(\vec{x}_n) + \bar{\bar{J}}(\vec{x}_n)(\vec{x}_{n+1} - \vec{x}_n) = \vec{0} \quad \bar{\bar{J}}^{-1}(\vec{x}_n) \cdot \dots$$

$$\bar{\bar{J}}^{-1}(\vec{x}_n) \vec{f}(\vec{x}_n) + (\vec{x}_{n+1} - \vec{x}_n) = \vec{0}$$

$$\vec{x}_{n+1} = \vec{x}_n - \underbrace{\bar{\bar{J}}^{-1}(\vec{x}_n) \vec{f}(\vec{x}_n)}_{\text{oznaczmy to}}$$

macierz odwrotna do $\bar{\bar{J}}(\vec{x}_n)$

przez $\bar{\bar{J}}_{n+1}$. Jest to poprawka, jaką odejmiemy od \vec{x}_n żeby uzyskać \vec{x}_{n+1} .

$$\vec{X}_{n+1} = \vec{X}_n - \vec{\delta}_{n+1}$$

Wystarczy tylko obliczyć $\vec{\delta}_{n+1}$. Nie musimy liczyć macierzy odwrotnej $\bar{\bar{g}}^{-1}(\vec{X}_n)$. Wystarczy rozwiązać układ równań algebraicznych liniowych:

$$\bar{\bar{g}}(\vec{X}_n) \vec{\delta}_{n+1} = \vec{f}(\vec{X}_n)$$

Stąd uzyskujemy $\vec{\delta}_{n+1}$

Rozwiązywanie
układów
algebraicznych równań
liniowych

Dodatek matematyczny : Normy wektorów i macierzy

① Normy wektorów

Załóżmy $x \in \mathbb{R}^N$

$\|x\|$ NORMA WEKTORA

Jest miarą "rozmiaru" wektora.

Spłnia aksjomaty:

1. $\|x\| > 0$ dla $x \neq 0$, $\|x\| = 0$ dla $x = 0$
2. $\|c \cdot x\| = |c| \cdot \|x\|$, $c \in \mathbb{R}$ lub $c \in \mathbb{C}$
3. $\|x+y\| \leq \|x\| + \|y\|$ NIERÓWNOŚĆ TRÓJKATA

wektor
zerowy
↓

liczby
zespółone
←

Przykład : Normy "p" (dla $p = 1, 2, \dots, \infty$)

$$\|x\|_p = \left[\sum_{i=1}^N |x_i|^p \right]^{1/p}$$

$$p=1 \Rightarrow \text{norma pierwsza} \quad \|x\|_1 = \sum_{i=1}^N |x_i|$$

$p=2 \Rightarrow$ norma druga (Euklidesowa)

$$\|x\|_2 = \sqrt{\sum_{i=1}^N |x_i|^2}$$

$p \rightarrow \infty \Rightarrow$ norma maksimum $\|x\|_\infty = \max_i \{ |x_i| \}$

② Normy macierzy

A macierz $N \times M$, $a_{ij} \in \mathbb{R}$

$\|A\|$ NORMA MACIERZY

Jest miarą "rozmiaru" macierzy.

Spełnia aksjomaty:

- $\|A\| > 0$ dla $A \neq 0$ i $\|A\| = 0$ dla $A = 0$
- $\|c \cdot A\| = |c| \cdot \|A\|$ dla $c \in \mathbb{R}$ lub $c \in \mathbb{C}$
- $\|A + B\| \leq \|A\| + \|B\|$
- $\|AB\| \leq \|A\| \cdot \|B\|$

Normy "p" macierzy - przykłady

$$\|A\|_1 = \max_j \sum_{i=1}^M |a_{ij}|$$

summa w kolumnie

$$\|A\|_\infty = \max_i \sum_{j=1}^N |a_{ij}|$$

summa we wierszu

$$\begin{bmatrix} 1 & \dots \\ \vdots & \infty \end{bmatrix}$$

$$\|A\|_2 = \sqrt{\rho(A^H A)}, \text{ gdzie } A^H = \text{macierz Hermitowsko sprzyjona z macierzą } A$$

$$\rho(B) = \max_{i=1, \dots, N} |\lambda_i| = \text{promień spektralny macierzy } B$$

wartości własne macierzy B

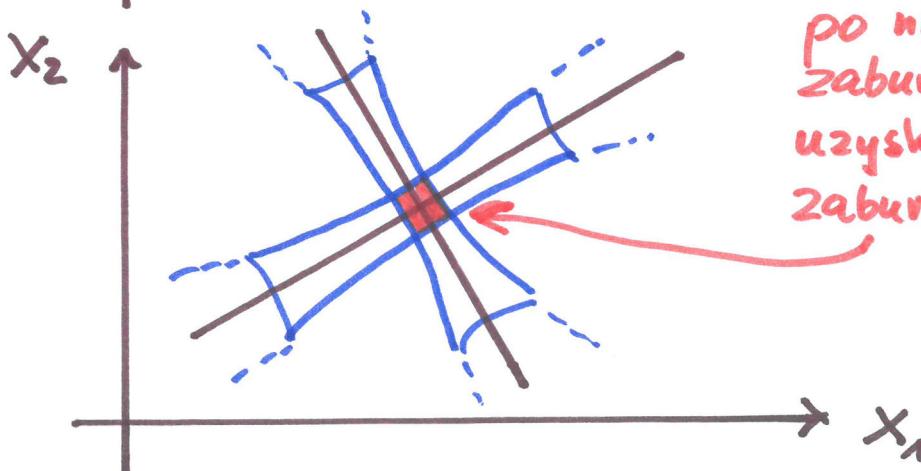
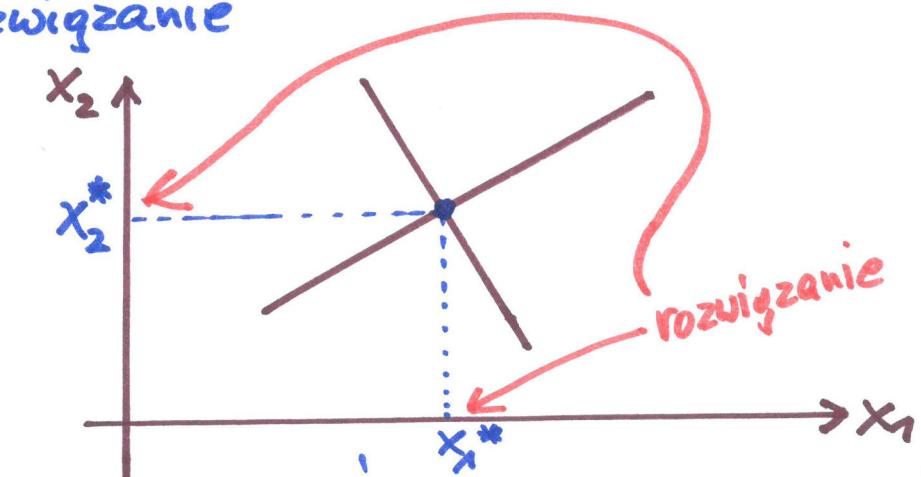
Uwarunkowanie zadania rozwiązywania układu równań liniowych $Ax=b$

Przykład: 2 równania z 2 niewiadomymi

$$\begin{cases} a_{11}x_1 + a_{12}x_2 = b_1 \\ a_{21}x_1 + a_{22}x_2 = b_2 \end{cases}$$

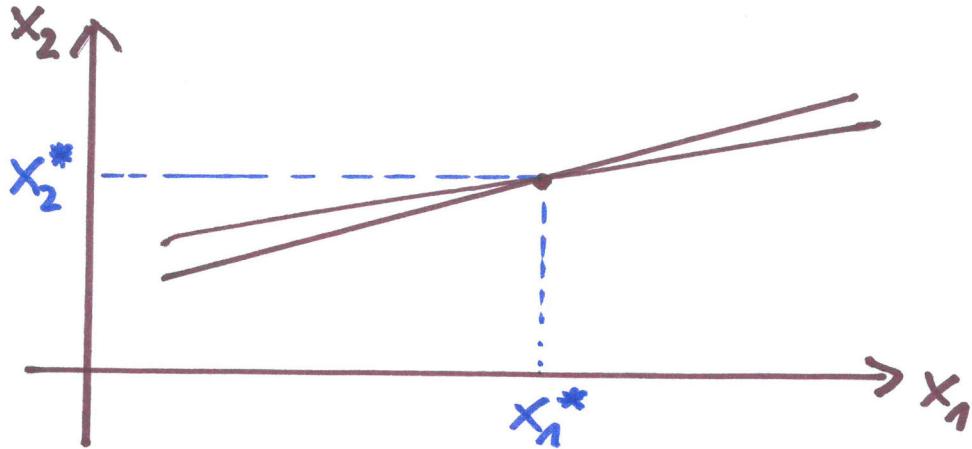
$$\underbrace{\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}}_A \underbrace{\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}}_x = \underbrace{\begin{bmatrix} b_1 \\ b_2 \end{bmatrix}}_b$$

Jeżeli A jest nieosobliwa, to istnieje dokładnie jedno rozwiązanie

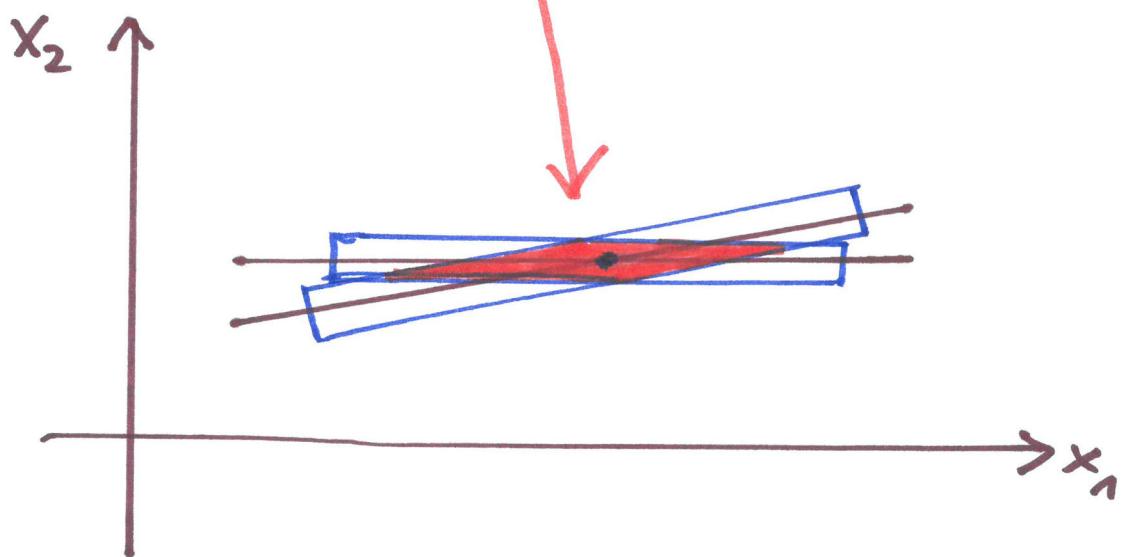


po nieznacznym zaburzeniu A lub b uzyskamy niewielkie zaburzenie rozwiązania

Jeżeli A jest prawie osobliwa, to proste są prawie równoległe



Nieznaczne zaburzenie A lub b powoduje DUŻE zaburzenie rozwiązań



Wniosek: Układ jest najlepiej uwarunkowany dla macierzy nieosobliwych, oraz źle uwarunkowany dla macierzy bliskich osobliwym

Scista analiza pokazuje, że:

jeśli δA — zaburzenie A

δb — zaburzenie b

δx — zaburzenie X

to

$$\frac{\|\delta x\|}{\|x\|} \leq \text{cond}(A) \cdot \frac{\|\delta b\|}{\|b\|}$$

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\text{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|}}{1 - \text{cond}(A) \cdot \frac{\|\delta A\|}{\|A\|}}$$

gdzie

$$\text{cond}(A) = \|A\| \cdot \|A^{-1}\|$$

Wskaźnik uwarunkowania macierzy / układu
(condition number)

Metody rozwiązywania układów równań liniowych

- 1) "szkolne" (wzory Cramera) — nieprzydatne dla $N > 4$ (duży koszt i błędy maszynowe)
- 2) Metody bezpośrednie (direct) zwane "dokładnymi"]
 - dla macierzy pełnych
 - dla macierzy rzadkich
- 3) Metody iteracyjne, przybliżone
- 4) Inne (np. metody Monte Carlo)

Odwiniemy metody 2 i 3, ale najpierw rozważymy kilka układów bardzo łatwych do rozwiązywania

Układy Tatwe do rozwiązyania

1) Z macierzą przekątniową (diagonalną)

$$\begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \quad a_{ii} \neq 0$$

$$x = \begin{bmatrix} b_1/a_{11} \\ \vdots \\ b_N/a_{NN} \end{bmatrix}$$

2) Z macierzą trójkątną dolną

$$\begin{bmatrix} l_{11} & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & 0 \\ l_{N1} & \dots & l_{N,N-1} & l_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_N \end{bmatrix}, \quad l_{ii} \neq 0$$

$$x_1 = b_1 / l_{11}$$

$$x_2 = (b_2 - l_{21}x_1) / l_{22}$$

⋮

$$x_i = (b_i - \sum_{j=1}^{i-1} l_{ij}x_j) / l_{ii}$$

⋮

$$x_N = (b_N - \sum_{j=1}^{N-1} l_{Nj}x_j) / l_{NN}$$

3) Z macierzą trójkątną górną

$$\begin{bmatrix} u_{11} & \dots & u_{1N} \\ 0 & \ddots & \vdots \\ 0 & \dots & u_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}, \quad u_{ii} \neq 0$$

$$\left\{ \begin{array}{l} x_N = b_N / u_{NN} \\ x_{N-1} = (b_{N-1} - u_{N-1,N} x_N) / u_{N-1,N} \\ \vdots \\ x_i = (b_i - \sum_{j=i+1}^N u_{ij} x_j) / u_{ii} \\ \vdots \\ x_1 = (b_1 - \sum_{j=2}^N u_{1j} x_j) / u_{11} \end{array} \right.$$

4) Z macierzą ortogonalną (ma ortogonalne kolumny)
(tzn. taka, że $A^T A = I$)

$$A x = b \quad / A^T \cdot \dots$$

$$\underbrace{A^T A}_{I} x = A^T b$$

$$x = A^T b$$

Układ z macierzą pełną – eliminacja Gaussa

Przekształcamy układ

$$\begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{NN} & \cdots & a_{NN} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ \vdots \\ b_N \end{bmatrix}$$

do postaci układu z macierzą trójkątną górną – poprzez kolejne kroki redukcji, w których zerujemy elementy w k -tej kolumnie macierzy, poniżej przekątnej ($k = \text{numer kroku}$).

Krok 1:

od równań i -tych ($i = 2, \dots, N$) odcinajemy równanie pierwsze pomnożone przez a_{i1}/a_{11} :

$$\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1N} \\ a_{21} - a_{11} \frac{a_{21}}{a_{11}} & a_{22} - a_{12} \frac{a_{21}}{a_{11}} & \cdots & a_{2N} - a_{1N} \frac{a_{21}}{a_{11}} \\ \vdots & \vdots & & \vdots \\ b_1 & b_2 - b_1 \frac{a_{21}}{a_{11}} & \cdots & b_N - b_1 \frac{a_{N1}}{a_{11}} \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}$$

tu powstają zera

W następnym, drugim kroku powtarzamy ten sam zabieg na układzie $N-1$ równań (zaznaczonym na zielono) i.t.d. aż do uzyskania jednego równania z jedną niewiadomą.

Ostatecznie uzyskujemy układ:

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & \dots & a_{1N} & x_1 \\ a_{22}^{(1)} & \dots & a_{2N}^{(1)} & & b_1 \\ a_{33}^{(2)} & \dots & a_{3N}^{(2)} & \vdots & b_2^{(1)} \\ \ddots & \ddots & \ddots & x_N & \vdots \\ a_{NN}^{(N-1)} & & & & b_N^{(N-1)} \end{array} \right]$$

Co pozwala wyliczyć rozwiązanie:

$$x_N = b_N^{(N-1)} / a_{NN}^{(N-1)}$$

$$x_i = \left(b_i^{(i-1)} - \sum_{j=i+1}^N a_{ij}^{(i-1)} x_j \right) / a_{ii}^{(i-1)}$$

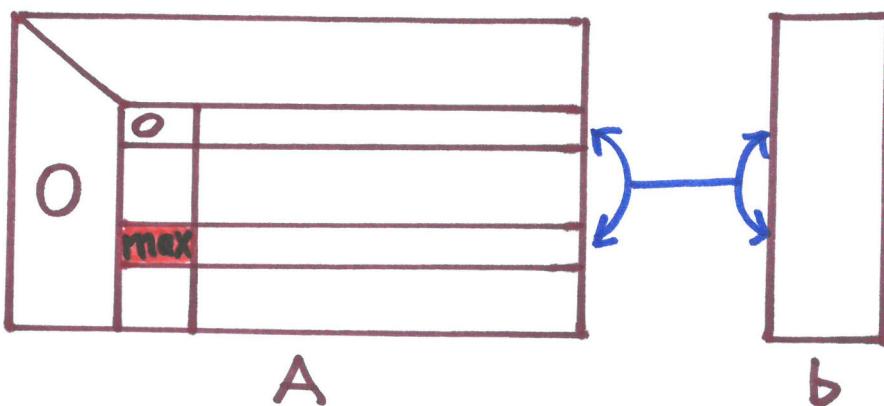
$$(i = N-1, \dots, 1)$$

Problem : dzielimy przez $a_{ii}^{(i-1)}$, które może być równe zero, lub bliskie zero.

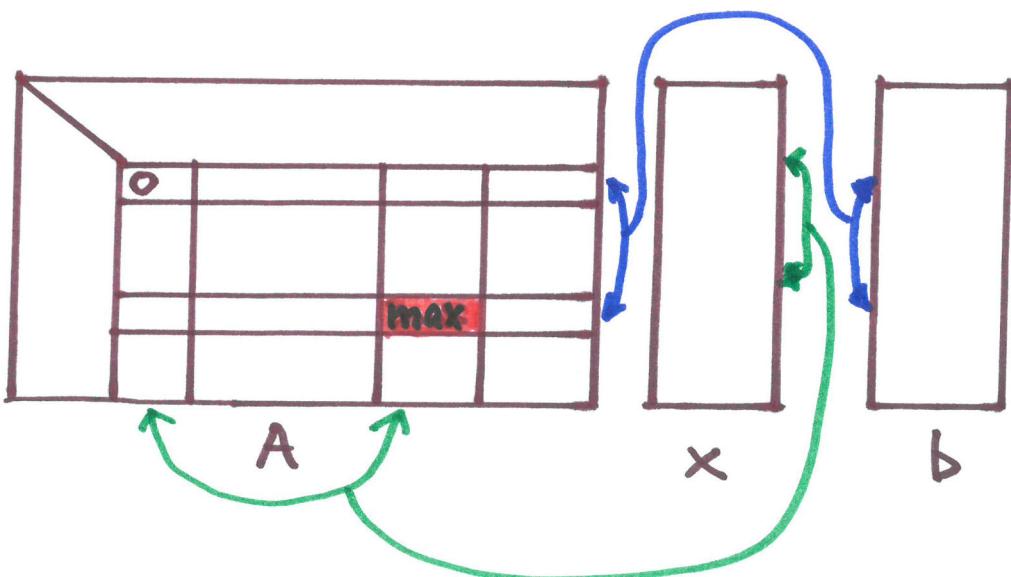
Element podstawowy (pivot)

W takich sytuacjach stosujemy procedury
CZĘŚCIOWEGO lub PEŁNEGO
WYBORU ELEMENTU PODSTAWOWEGO
(partial pivoting, full pivoting)

Wybór częściowy
(zmiana kolejności wierszy)



Wybór pełny
(zmiana kolejności wierszy i kolumn)



max to element dla którego $|a_{kl}^{(i-1)}|$ ma maksymalną wartość

Przykład

$$\left| \begin{array}{cccc|c} 6 & 2 & 2 & 4 & 1 \\ -1 & 2 & 2 & -3 & -1 \\ 0 & 1 & 1 & 4 & 2 \\ 1 & 0 & 2 & 3 & 1 \end{array} \right|$$

odejmujemy 1 wiersz $\times \left(-\frac{1}{6}\right)$
 --- $1 - 1 \times \left(\frac{0}{6}\right)$
 --- $1 - 1 \times \left(\frac{1}{6}\right)$

$$\left| \begin{array}{cccc|c} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 1 & 1 & 4 & 2 \\ 0 & -\frac{1}{3} & \frac{5}{3} & \frac{7}{3} & \frac{5}{6} \end{array} \right|$$

odejmujemy 2 wiersz $\times \left(\frac{3}{7}\right)$
 odejmujemy 2 wiersz $\times \left(-\frac{1}{7}\right)$

$$\left| \begin{array}{cccc|c} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \\ 0 & 0 & 2 & 2 & \frac{5}{7} \end{array} \right|$$

częściowy wybór elementu podstawowego

$$\left| \begin{array}{cccc|c} 6 & 2 & 2 & 4 & 1 \\ 0 & \frac{7}{3} & \frac{7}{3} & -\frac{7}{3} & -\frac{5}{6} \\ 0 & 0 & 2 & 2 & \frac{5}{7} \\ 0 & 0 & 0 & 5 & \frac{33}{14} \end{array} \right|$$

odejmujemy 3 wiersz $\times \left(\frac{0}{2}\right) \Rightarrow$
 nic się nie zmieni

Procedury wyboru elementu podstawowego
nie są potrzebne m. in. gdy

1) A jest $\begin{cases} \text{diagonalnie} \\ \text{diagonalnie silnie} \end{cases}$ dominujące, tzn.

$$|a_{ii}| \left\{ \begin{array}{l} \geq \\ > \end{array} \right\} \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}| \quad \text{dla } i=1,2,\dots,N$$

2) A jest $\begin{cases} \text{diagonalnie} \\ \text{diagonalnie silnie} \end{cases}$ dominujące kolumnowo, tzn.

$$|a_{ii}| \left\{ \begin{array}{l} \geq \\ > \end{array} \right\} \sum_{\substack{j=1 \\ j \neq i}}^N |a_{ij}| \quad \text{dla } i=1,2,\dots,N$$

Inny problem: co zrobić, gdy mamy do rozwiązyania wiele układów z tą samą macierzą:

$$A x_1 = b_1$$

$$A x_2 = b_2$$

:

Trzeba jakś rozdzielić przekształcanie A od przekształcania b , aby zminimalizować koszty.

Metoda dekompozycji LU

Jeśli $A = L U$ gdzie L - macierz trójkątna dolna
 U - -" - -" - górną

to $A x = b$

$$L U x = b$$

$$L \underbrace{(U x)}_{\text{oznaczmy jako } y} = b$$

oznaczmy jako y

$$\begin{cases} L y = b \Rightarrow \text{stąd wyliczamy } y \\ \text{ale} \\ U x = y \Rightarrow \text{stąd wyliczamy } x \end{cases}$$

A zatem, dekompozycję A na iloczyn $L U$ wykonujemy tylko raz, a obliczenia x powtarzamy dla każdego wektora b .

Okazuje się, że macierz U to dokładnie ta sama macierz, jaką uzyskujemy w opisanej eliminacji Gaussa.

Natomiast L to macierz współczynników, przez które mnożyliśmy wiersze macierzy A

Na przekątnej zawiera jedynki, a poniżej przekątnej współczynniki

Dekompozycja LU dla macierzy trójdagonalnej
(przykład macierzy rzadkiej) i rozwiązywanie układu

Algorytm Thomasa (bez wyboru elementów podstawowych)

$$\begin{bmatrix} d_1 & u_1 & & & & \\ l_2 & d_2 & u_2 & & & \\ l_3 & d_3 & u_3 & & & \\ \dots & \dots & \dots & & & \\ & & & l_{N-1} & d_{N-1} & u_{N-1} \\ & & & l_N & d_N & & \end{bmatrix} \begin{matrix} O \\ O \end{matrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_{N-1} \\ x_N \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_{N-1} \\ b_N \end{bmatrix}$$

Eliminacja Gaussa wymaga przekształcenia tylko jednego wiersza w każdym kroku redukcji, ponieważ pozostałe elementy w kolumnach są już zerami.

Wprowadźmy współczynniki γ_i na głównej przekątnej macierzy U , oraz elementy r_i przekształconego wektora b

W pierwszym kroku redukcji $\gamma_1 = d_1$, $r_1 = b_1$, natomiast

$$\gamma_2 = d_2 - l_2 \gamma_1^{-1} u_1, \quad r_2 = b_2 - l_2 \gamma_1^{-1} r_1$$

$$\begin{bmatrix} \gamma_1 & u_1 & & & & \\ 0 & \gamma_2 & u_2 & & & \\ l_3 & d_3 & u_3 & & & \\ \dots & \dots & \dots & & & \\ & & & & & \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ b_3 \\ \vdots \end{bmatrix}$$

Powtarzając ten zabieg w kolejnych krokach redukcji dostajemy algorytm:

$$\gamma_1 = d_1$$

$$r_1 = b_1$$

dla $i = 2, \dots, N$

$$\begin{cases} \gamma_i = d_i - l_i \gamma_{i-1}^{-1} u_{i-1} \\ r_i = b_i - l_i \gamma_{i-1}^{-1} r_{i-1} \end{cases}$$

to najlepiej rozdzielić na dwie procedury, z których jedna działa tylko na macierz A , a druga na wektor b

co daje

$$\begin{bmatrix} \gamma_1 & u_1 & 0 & & & \\ \gamma_2 & u_2 & & & & \\ \vdots & \ddots & & & & \\ 0 & & u_{N-1} & X_{N-1} & & \\ & & \gamma_N & X_N & & \end{bmatrix} = \begin{bmatrix} r_1 \\ r_2 \\ \vdots \\ r_{N-1} \\ r_N \end{bmatrix}$$

Druga procedura
taczymy z

Następnie wyznaczamy rozwiążanie:

$$x_N = \gamma_N^{-1} r_N$$

dla $i = N-1, \dots, 1$

$$x_i = \gamma_i^{-1} (r_i - u_i x_{i+1})$$

Powyższy algorytm nie ulegnie (formalnie) zmianie, jeśli zamiast pojedynczych liczb, symbole $d_i, l_i, u_i, x_i, b_i, \gamma_i, r_i$ oznaczać będą niewielkie macierze i wektory. Wtedy γ_i^{-1} oznaczać będzie macierze odwrotne.

W takim przypadku mówimy o macierzy A blokowo-trójdziagonalnej

Nadokresione
układy liniowych równań
algebraicznych

Nadokreślone układy równań liniowych,
liniowe zadanie najmniejszych kwadratów

$$Ax = b$$

gdzie $A_{m \times n}$, $m > n = \text{rank}(A)$

(więcej równań niż niewiadomych)

Układu nie da się rozwiązać ścisłe, tzn. nie da się uzyskać

$$r = b - Ax = 0$$

↗
wektor rezidualny

Można jednak szukać takiego x aby uzyskać

$$\|r\| = \|b - Ax\| = \min$$

Jeżeli zastosujemy normę drugą $\|r\|_2$

to mamy LINIOWE ZADANIE NAJMIEJSZYCH KWADRATÓW

x nazywamy wówczas PSEUDO ROZWIAZANIEM

Twierdzenie:

Rozwiązyaniem zadania najmniejszych kwadratów, jest wektor x spełniający tzw. układ równań

NORMALNYCH:

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

(w zwykłym sensie)

Dw.

$$\|b - Ax\|_2^2 = \sum_{i=1}^m \left(b_i - \sum_{j=1}^n a_{ij} x_j \right)^2$$

$$\|b - Ax\|_2^2 = \min \Leftrightarrow \frac{\partial}{\partial x_k} \sum_{i=1}^m \left(b_i - \sum_{j=1}^n a_{ij} x_j \right)^2 = 0$$

$$\text{dla } k = 1, \dots, n$$

$$\sum_{i=1}^m 2 \left(b_i - \sum_{j=1}^n a_{ij} x_j \right) a_{ik} = 0$$

$$\sum_{i=1}^m a_{ik} \left(b_i - \sum_{j=1}^n a_{ij} x_j \right) = 0$$

$$A^T (b - Ax) = 0$$

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

Przykład

$$\begin{cases} x+y=2 \\ x-y=0 \\ x-2y=-2 \end{cases}$$

3 równania niezależne
2 niewiadome

przekształcamy do postaci $Ax-b=0$

$$\begin{cases} x+y-2=0 \\ x-y=0 \\ x-2y+2=0 \end{cases}$$

$$\|Ax-b\|_2^2 = (x+y-2)^2 + (x-y)^2 + (x-2y+2)^2$$

$$\frac{\partial}{\partial x} \|Ax-b\|_2^2 = 2(x+y-2) + 2(x-y) + 2(x-2y+2) = 0$$

$$\frac{\partial}{\partial y} \|Ax-b\|_2^2 = 2(x+y-2) - 2(x-y) - 4(x-2y+2) = 0$$

$$\begin{cases} 3x-2y=0 \\ -x+3y-3=0 \end{cases}$$

$$\begin{cases} x=\frac{6}{7} \\ y=\frac{3}{7} \end{cases}$$

pseudorozwiązywanie

Układ równań normalnych należy rozwiązywać za pomocą specjalnych metod, gdyż bywa źle uwarunkowany

Mnożenie macierzy w arytmetyce zmienoprzecinkowej może zmienić rangę macierzy i spowodować, że stanie się osobliwa.

Przykład:

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 \\ e & e & e & e \end{bmatrix}_{5 \times 4} \quad A^T = \begin{bmatrix} 1 & e & e & e \\ 1 & e & e & e \\ 1 & e & e & e \\ 1 & e & e & e \end{bmatrix}_{4 \times 5}$$

$$A^T A = \begin{bmatrix} 1+e^2 & 1 & 1 & 1 \\ 1 & 1+e^2 & 1 & 1 \\ 1 & 1 & 1+e^2 & 1 \\ 1 & 1 & 1 & 1+e^2 \end{bmatrix}$$

Jeżeli $e^2 < 2$ to $\text{fl}(1+e^2) = 1 \Rightarrow \text{rank}(\text{fl}(A^T A)) = 1$
 natomiast $\text{rank}(A) = 4$ gdy tylko $e > 0$.

Dlatego stosuje się rozkład QR:

$$A^T A = QR$$

macierz ortogonalna $(Q^T Q = I)$

macierz trójkątna górska

W takim wypadku

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

$$QR x = A^T b \quad | Q^T \dots$$

$$\underbrace{Q^T Q}_{I} R x = Q^T A^T b$$

$$R x = Q^T A^T b$$

co łatwo rozwiązać

Do przeprowadzenia rozkładu QR
stosuje metodę Householdera

Metody iteracyjne
do rozwiązywania
układów liniowych
równan algebraicznych

$$Ax = b$$

Tworzymy ciąg kolejnych przybliżeń rozwiązania x :

$$x^{(0)}, x^{(1)}, \dots$$

Które obliczamy za pomocą wzoru

$$x^{(n)} = M^{(n)} x^{(n-1)} + c^{(n)}$$

macierz zależna
od metody
(i od A, b)

wektor zależny
od metody
(i od A, b)

Jest to liniowy odpowiednik $x^{(n)} = \Phi(x^{(n-1)})$

$$M^{(n)} = M, \text{ (nie zależy od } n \text{)} \Rightarrow \text{METODY STACJONARNE}$$

$$c^{(n)} = c$$

$$M^{(n)}, c^{(n)} \text{ zależy od } n \Rightarrow \text{METODY NIESTACJONARNE}$$

Omówimy wytypcznie metody STACJONARNE
Główna trudność: jak skonstruować M, c
aby uzyskać szybką zbieżność.

① Metoda Richardsona (odpowiednik metody Picarda)

$$Ax = b$$

$$Ax + x - x = b$$

$$x + (Ax - x) = b$$

$$x - (I - A)x = b$$

$$x = (I - A)x + b$$

$$x^{(n)} = \underbrace{(I - A)}_M x^{(n-1)} + \underbrace{b}_c$$

② Metoda Jacobiego

$$AX = b$$

rozkładamy $A = L + D + U$

Uwaga: nie mylić z rozkładem $L U$, tu nie ma mnożenia!

L

D

U

$$(L + D + U)x = b$$

$$(L + U)x + Dx = b$$

$$Dx = -(L + U)x + b \quad / D^{-1} \dots$$

$$x = -D^{-1}(L + U)x + D^{-1}b$$

$$x^{(n)} = \underbrace{-D^{-1}(L + U)x^{(n-1)}}_M + \underbrace{D^{-1}b}_C$$

③ Metoda Gaussa - Seidela

$$Ax = b$$

$$(L+D+U)x = b$$

$$(L+D)x + Ux = b$$

$$(L+D)x = -Ux + b$$

wzór operacyjny:

$$(L+D)x^{(n)} = -Ux^{(n-1)} + b$$

$$(U+D)x + Lx = b$$

$$(U+D)x = -Lx + b$$

$$(U+D)x^{(n)} = -Lx^{(n-1)} + b$$

wzór teoretyczny:

$$x^{(n)} = \underbrace{-(L+D)^{-1}Ux^{(n-1)}}_M + \underbrace{(L+D)^{-1}b}_C$$

$$x^{(n)} = \underbrace{-(U+D)^{-1}Lx^{(n-1)}}_M + \underbrace{(U+D)^{-1}b}_C$$

Wariant 1

Wariant 2

④ Metoda sukcesywnej nadrelaksacji (SOR)

Younga i Frankela
(successive over-relaxation)

$$Ax = b$$

$$(L + D + U)x = b$$

$$\left[L + \frac{1}{\omega} D + \left(1 - \frac{1}{\omega}\right) D + U \right] x = b$$

ω - parametr
dobierany tak
aby uzyskać
przyspieszenie
zbieżności

$$\left(L + \frac{1}{\omega} D \right) x + \left[\left(1 - \frac{1}{\omega}\right) D + U \right] x = b$$

$$\left(L + \frac{1}{\omega} D \right) x = - \left[\left(1 - \frac{1}{\omega}\right) D + U \right] x + b$$

wzór operacyjny:

$$\left(L + \frac{1}{\omega} D \right) x^{(n)} = - \left[\left(1 - \frac{1}{\omega}\right) D + U \right] x^{(n-1)} + b$$

wzór teoretyczny:

$$x^{(n)} = \underbrace{- \left(L + \frac{1}{\omega} D \right)^{-1} \left[\left(1 - \frac{1}{\omega}\right) D + U \right] x^{(n-1)}}_M + \underbrace{\left(L + \frac{1}{\omega} D \right)^{-1} b}_C$$

Zbieżność stacjonarnych metod iteracyjnych

Z tw. Banacha o kontrakcji, aby iteracje były zbieżne, odwzorowanie $\phi(x)$ musi być zwężające, tzn.

$$\bigvee_{\lambda \in [0,1)} \bigwedge_{x,y} \|\phi(x) - \phi(y)\| \leq \lambda \|x-y\|$$

W naszym przypadku $\phi(x) = Mx + c$

$$\begin{aligned} \|\phi(x) - \phi(y)\| &= \|(Mx + c) - (My + c)\| \\ &= \|M(x-y)\| \leq \underbrace{\|M\|}_{\text{pełni rolę } \lambda} \|x-y\| \end{aligned}$$

Zatem warunkiem dostatecznym zbieżności jest $\|M\| < 1$

Można też udowodnić, że

Warunkiem koniecznym i wystarczającym zbieżności jest $\rho(M) < 1$

$$\rho(M) = \max_{1 \leq i \leq N} |\lambda_i| = \text{promień spektralny } M$$

Mozna też pokazać, że w metodach Jacobiego, Gaussa-Seidela i SOR warunek $\|M\|_\infty < 1$ będzie spełniony jeżeli A jest diagonalnie silnie dominującą, przy czym w SOR musi być też spełniony warunek $0 < \omega < 2$!

Iteracyjne poprawianie rozwiązań w metodach bezpośrednich

$$Ax = b$$

Metody bezpośrednie dają $x^{(0)}$ z błędem maszynowym $e^{(0)}$, tzn.

$$x = x^{(0)} + e^{(0)}$$

$$b - Ax^{(0)} = r^{(0)} \neq 0$$

↑
residuum

Zauważmy, że

$$Ax = Ax^{(0)} + Ae^{(0)} \quad | - b$$

$$Ax - b = Ax^{(0)} - b + Ae^{(0)}$$

$\underbrace{\quad}_{\text{``0"}}$ $\underbrace{\quad}_{\text{``-r^{(0)}"}}$

$Ae^{(0)} = r^{(0)} \Rightarrow$ wyliczamy $e^{(0)}$ i dodajemy

$$\text{do } x^{(0)} : \quad x^{(1)} = x^{(0)} + e^{(0)}$$

↑ poprawione rozwiązań

$x^{(1)}$ możemy poprawić w analogiczny sposób

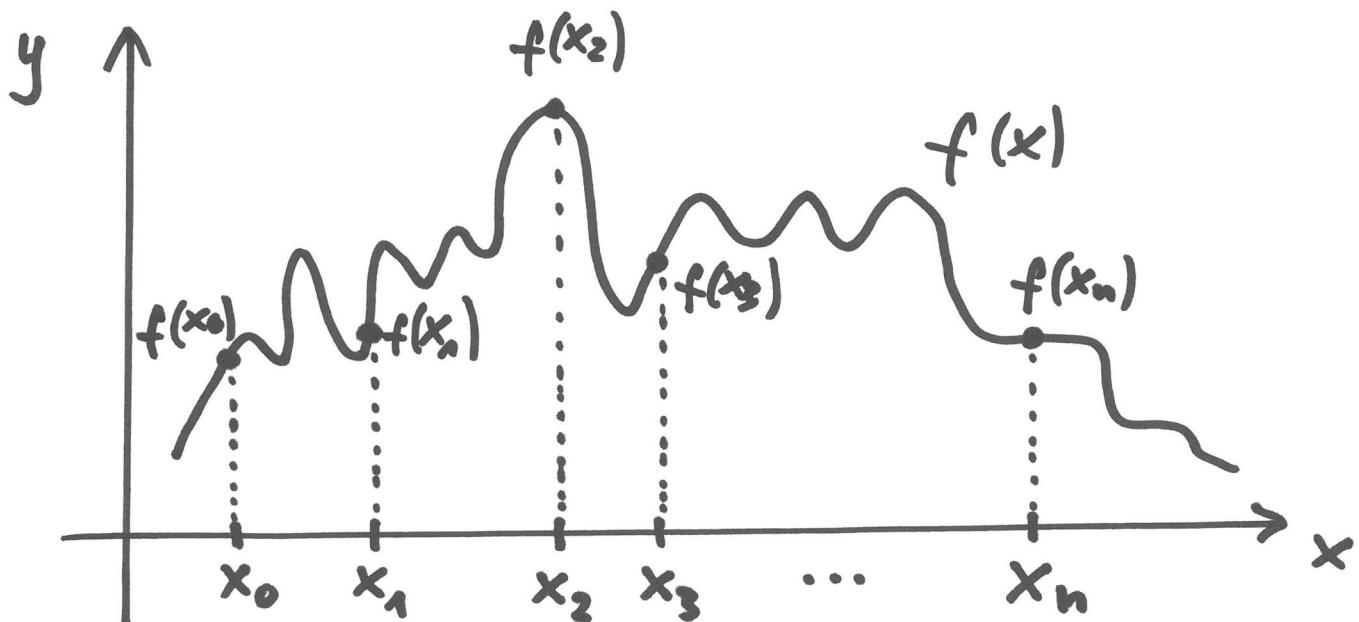
Obliczenia te wymagają jednak podwyższonej precyzji

Podstawy Metod różnicowych

Prybliżenia różnicowe pochodnych funkcji

Postawienie problemu:

Jak obliczyć przybliżone wartości pochodnych funkcji $y = f(x)$, mając jedynie zadane wartości funkcji w punktach x_0, x_1, \dots, x_n ?

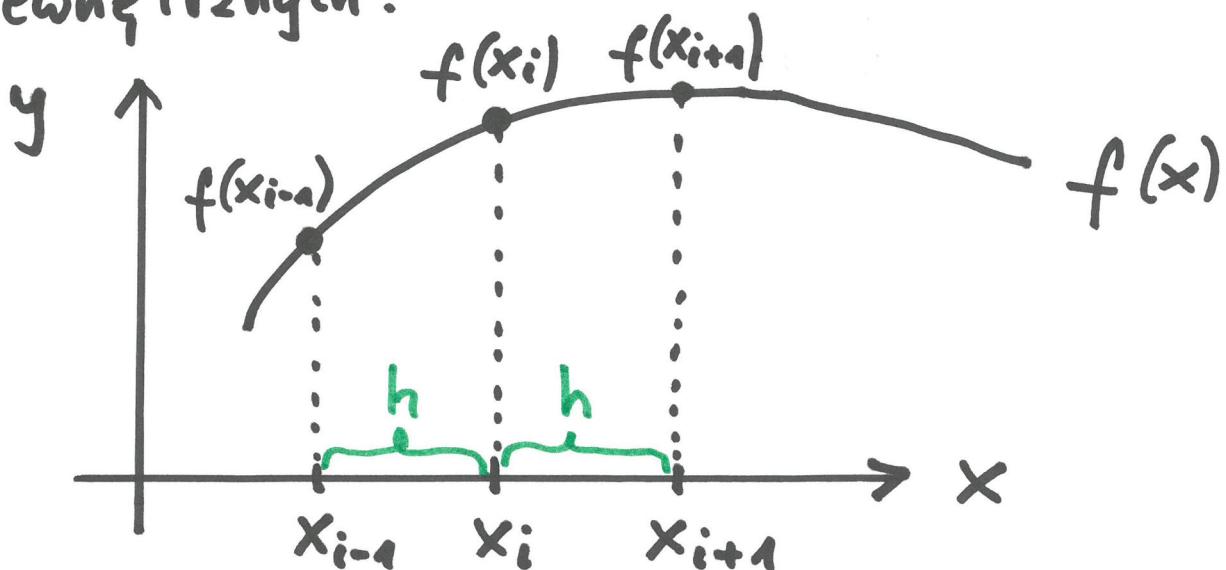


Widac że jest to trudne zadanie, bo nie wiemy o zachowaniu się funkcji pomiędzy punktami x_i

Punkty x_0, x_1, \dots, x_n nazywamy
WEZŁAMI SIATKI DYSKRETNIEJ

Najprostsze przybliżenia pochodnych: wzory DWU i TRZY PUNKTOWE na siatce jednorodnej o kroku h

1 Wzory na pochodne w węzłach wewnętrznych:



$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_i)}{h}$$

RÓŻNICA PROGRESYWNA
(forward difference)

$$f'(x_i) \approx \frac{f(x_i) - f(x_{i-1})}{h}$$

RÓŻNICA WSTECZNA
(backward difference)

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_{i-1})}{2h}$$

RÓŻNICA CENTRALNA
(central difference)

$$f''(x_i) \approx \frac{\frac{f(x_{i+1}) - f(x_i)}{h} - \frac{f(x_i) - f(x_{i-1})}{h}}{h} =$$

$$= \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2}$$

RÓŻNICA CENTRALNA (DLA 2 POCZ.)
 (central difference)

Błąd przybliżeń różnicowych analizujemy
Korzystając z rozwinięcia funkcji w szereg
(zakładając istnienie niezbędnych pochodnych!)

$$f(x_{i+1}) = f(x_i) + f'(x_i)h + \frac{1}{2}f''(x_i)h^2 + \frac{1}{6}f'''(x_i)h^3 + \frac{1}{24}f^{(4)}(x_i)h^4 + \dots$$

$$f(x_{i-1}) = f(x_i) - f'(x_i)h + \frac{1}{2}f''(x_i)h^2 - \frac{1}{6}f'''(x_i)h^3 + \frac{1}{24}f^{(4)}(x_i)h^4 - \dots$$

Stąd uzyskamy:

$$\frac{f(x_{i+1}) - f(x_i)}{h} = f'(x_i) + \frac{1}{2}f''(x_i)h + \frac{1}{6}f'''(x_i)h^2 + \dots$$

Błąd obliczenia = $O(h)$

$$\frac{f(x_i) - f(x_{i-1})}{h} = f'(x_i) - \frac{1}{2}f''(x_i)h + \frac{1}{6}f'''(x_i)h^2 - \dots$$

Błąd obliczenia = $O(h)$

$$\frac{f(x_{i+1}) - f(x_{i-1})}{2h} = f'(x_i) + \frac{1}{6} f''(x_i) h^2 + \dots$$

Błąd obcięcia = $O(h^2)$

$$\frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2} = f''(x_i) + \frac{1}{12} f^{(4)}(x_i) h^2 + \dots$$

Błąd obcięcia = $O(h^2)$

Ogólne biorąc, błąd obcięcia dla przybliżeń różnicowych wyraża się wzorem:

RZĄD DOKŁADNOŚCI PRZYBLIŻENIA

$$T = A h^P + B h^{P+1} + C h^{P+2} + \dots$$

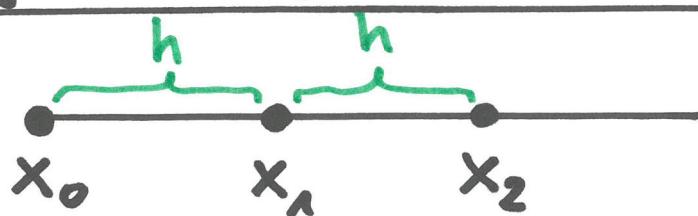
wyraz dominujący
błędu obcięcia

współczynniki zależne od
wyższych pochodnych $f(x)$

Zazwyczaj dalsze wyrazy pomijamy:

$$T \approx A h^P = O(h^P)$$

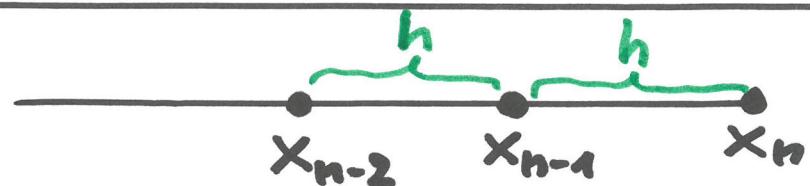
Wzory dwu- i trzypunktowe dla pochodnych w węzłach końcowych siatki



$$f'(x_0) = \frac{f(x_1) - f(x_0)}{h} + O(h)$$

$$f'(x_0) = \frac{-\frac{3}{2}f(x_0) + 2f(x_1) - \frac{1}{2}f(x_2)}{h} + O(h^2)$$

$$f''(x_0) = \frac{f(x_0) - 2f(x_1) + f(x_2)}{h^2} + O(h)$$



$$f'(x_n) = \frac{f(x_n) - f(x_{n-1})}{h} + O(h)$$

$$f'(x_n) = \frac{\frac{1}{2}f(x_{n-2}) - 2f(x_{n-1}) + \frac{3}{2}f(x_n)}{h} + O(h^2)$$

$$f''(x_n) = \frac{f(x_{n-2}) - 2f(x_{n-1}) + f(x_n)}{h^2} + O(h)$$

Analiza błędów z uwzględnieniem wpływu błędów maszynowych

Na przykładzie różnicy progresywnej

$$\frac{[f(x_{i+1})(1+e_{i+1}) - f(x_i)(1+e_i)](1+\delta_0)}{h} (1+\delta_D) \approx f'(x_i) + \frac{1}{2} f''(x_i) h$$

Pomijamy błąd reprezentacji h .

Jeśli założyć że $|e_i| \gg |\delta_0|, |\delta_D|$ to

$$\frac{f(x_{i+1}) - f(x_i)}{h} + \frac{f(x_{i+1})e_{i+1} - f(x_i)e_i}{h} \approx$$

$$f'(x_i) + \frac{1}{2} f''(x_i) h$$

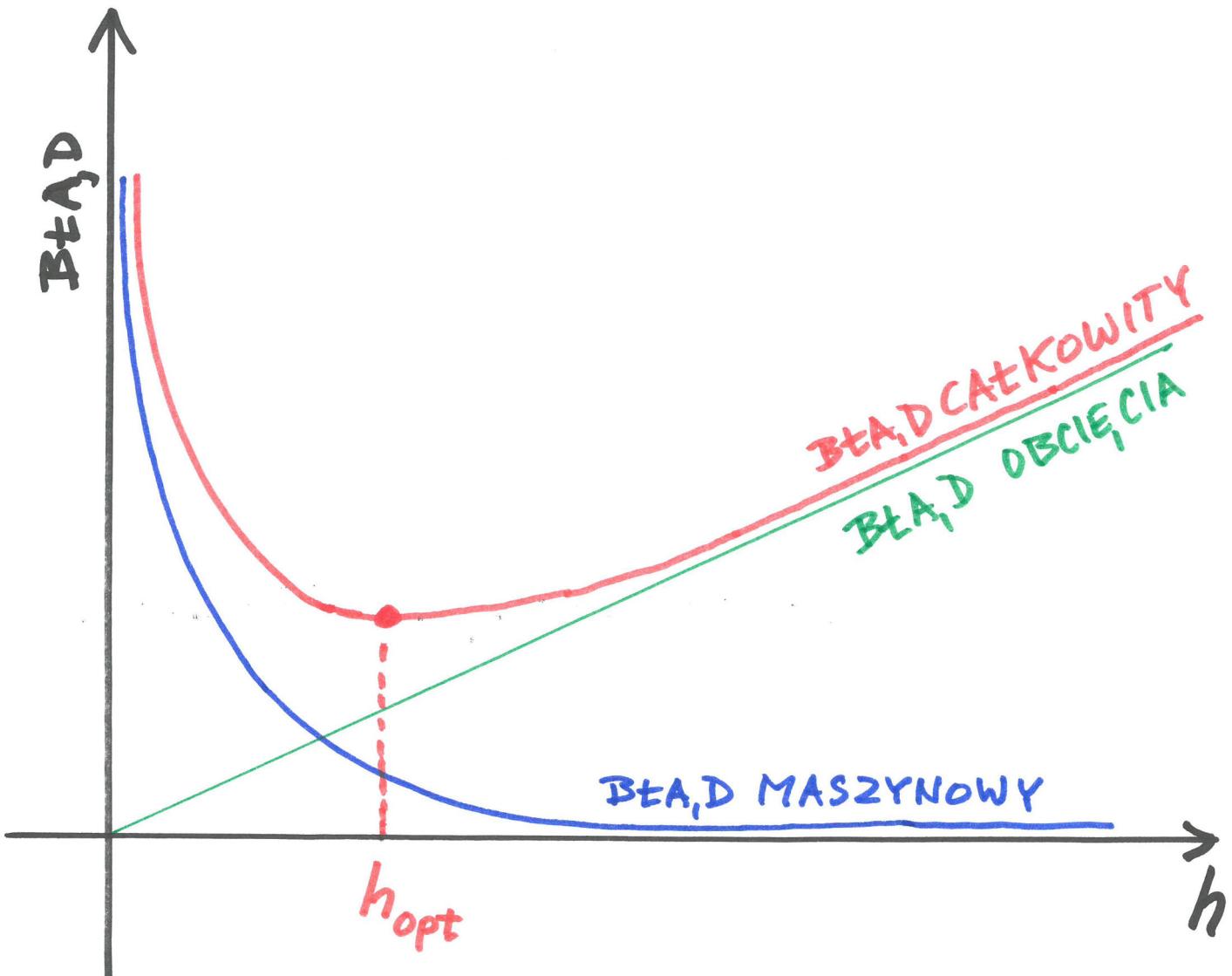
$$\left| \frac{f(x_{i+1}) - f(x_i)}{h} - f'(x_i) \right| \approx$$

$$\left| -\frac{f(x_{i+1})e_{i+1} - f(x_i)e_i}{h} + \frac{1}{2} f''(x_i) h \right|$$

$$\leq \underbrace{\frac{|f(x_{i+1})| \cdot |e_{i+1}| + |f(x_i)| \cdot |e_i|}{h}}_{\text{Błąd maszynowy}} + \underbrace{\frac{1}{2} |f''(x_i)| h}_{\text{Błąd obcięcia}}$$

błąd maszynowy

błąd obcięcia



h_{opt} to krok "optymalny" w tym sensie że odpowiada on najmniejszemu błądowi całkowitemu.

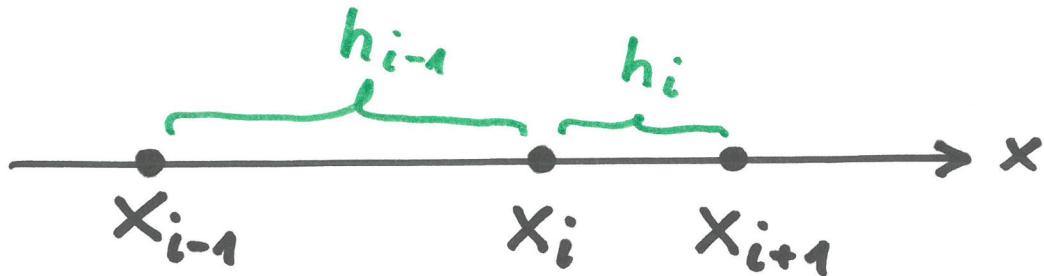
Jednakże w obliczeniach należy zawsze używać $h \gg h_{opt}$

Nyprawdżanie wzorów na pochodne w ogólnym przypadku siatek niejednorodnych

Dwie alternatywne metody:

- 1) Rozwinięcia w szereg + warunki zgodności z pochodną + minimalizacja błędu obciążenia.
- 2) Różniczkowanie analityczne wzorów interpolacyjnych Lagrange'a
(tą metodę omówimy później)

Metoda 1 : wyprowadzenie wzoru trzypunktowego na pierwszą pochodną - w wewnętrznym węźle siatki niejednorodnej -



$$f(x_{i+1}) = f(x_i) + f'(x_i)h_i + \frac{1}{2}f''(x_i)h_i^2 + \frac{1}{6}f'''(x_i)h_i^3 + \dots$$

$$f(x_{i-1}) = f(x_i) - f'(x_i)h_{i-1} + \frac{1}{2}f''(x_i)h_{i-1}^2 - \frac{1}{6}f^{(3)}(x_i)h_{i-1}^3 + \dots$$

żądamy :

$$f'(x_i) = a \cdot f(x_{i-1}) + b \cdot f(x_i) + c \cdot f(x_{i+1}) + T$$

Współczynniki do wyznaczenia

Możliwie najmniejszy

$$a \left[f(x_i) - f'(x_i)h_{i-1} + \frac{1}{2}f''(x_i)h_{i-1}^2 - \frac{1}{6}f^{(3)}(x_i)h_{i-1}^3 + \dots \right]$$

$$+ b \cdot f(x_i)$$

$$+ c \left[f(x_i) + f'(x_i)h_i + \frac{1}{2}f''(x_i)h_i^2 + \frac{1}{6}f^{(3)}(x_i)h_i^3 + \dots \right]$$

$$= f'(x_i) - T$$

grupujemy wyrazy z kolejnymi pochodnymi:

$$\begin{aligned} & (a+b+c) \cdot f(x_i) \\ & + (-ah_{i-1} + ch_i) \cdot f'(x_i) \\ & + \left(\frac{1}{2}ah_{i-1}^2 + \frac{1}{2}ch_i^2\right) \cdot f''(x_i) \\ & + \left(-\frac{1}{6}ah_{i-1}^3 + \frac{1}{6}ch_i^3\right) f^{(3)}(x_i) \\ & + \dots \\ & = f'(x_i) - T \end{aligned}$$

stąd

$$\begin{cases} a+b+c=0 \\ -ah_{i-1}+ch_i=1 \\ \frac{1}{2}ah_{i-1}^2+\frac{1}{2}ch_i^2=0 \end{cases} \begin{array}{l} \text{zgodność z pochodnymi} \\ \text{minimalizacja } T \end{array}$$

$$T = -\left(-\frac{1}{6}ah_{i-1}^3 + \frac{1}{6}ch_i^3\right) f^{(3)}(x_i) + \dots$$

po rozwiązyaniu układu:

$$a = -\frac{h_i}{h_{i-1}(h_{i-1}+h_i)}$$

$$b = \frac{h_i - h_{i-1}}{h_i h_{i-1}}$$

$$c = \frac{h_{i-1}}{h_i(h_{i-1}+h_i)}$$

$$T = -\frac{1}{6}h_i h_{i-1} f^{(3)}(x_i) + \dots$$

Stąd ostatecznie

$$f'(x_i) = -\frac{h_i}{h_{i-1}(h_{i-1}+h_i)} f(x_{i-1}) + \frac{h_i - h_{i-1}}{h_i h_{i-1}} f(x_i) + \frac{h_{i-1}}{h_i(h_{i-1}+h_i)} f(x_{i+1}) - T$$

Dla siatki jednorodnej $h_{i-1} = h_i = h$
a wzór redukuje się do różnicy centralnej:

$$f'(x_i) \approx \frac{f(x_{i+1}) - f(x_{i-1})}{2h}$$

Podobnie dla drugiej pochodnej,
żądając

$$f''(x_i) = a \cdot f(x_{i-1}) + b \cdot f(x_i) + c \cdot f(x_{i+1}) - T$$

uzyskamy (zadanie domowe !)

$$f''(x_i) = \frac{2}{h_{i-1}(h_{i-1} + h_i)} f(x_{i-1}) - \frac{2}{h_{i-1} h_i} f(x_i) \\ + \frac{2}{h_i(h_{i-1} + h_i)} f(x_{i+1}) - T,$$

$$T = \frac{1}{3} (h_{i-1} - h_i) f^{(3)}(x_i) + \dots$$

co w przypadku siatki jednorodnej redukuje się do

$$f''(x_i) \approx \frac{f(x_{i-1}) - 2f(x_i) + f(x_{i+1})}{h^2}$$

Metoda 2: Wyprowadzenie wzorów na pochodne poprzez różniczkowanie wielomianu interpolacyjnego Lagrange'a

$$f(x) = \sum_{i=0}^n f(x_i) L_i(x) + \frac{1}{(n+1)!} f^{(n+1)}(\xi(x)) \prod_{i=0}^n (x - x_i)$$

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

wielomian interpolacyjny w bazie Lagrange'a

$$f'(x) \approx \sum_{i=0}^n f(x_i) L'_i(x)$$

$$f''(x) \approx \sum_{i=0}^n f(x_i) L''_i(x)$$

Dla 3 węzłów interpolacji

$$f(x) \approx f(x_{i-1}) \frac{(x-x_i)(x-x_{i+1})}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})}$$

$$+ f(x_i) \frac{(x-x_{i-1})(x-x_{i+1})}{(x_i-x_{i-1})(x_i-x_{i+1})}$$

$$+ f(x_{i+1}) \frac{(x-x_{i-1})(x-x_i)}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)}$$

$$f'(x) \approx f(x_{i-1}) \frac{2x-x_i-x_{i+1}}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})}$$

$$+ f(x_i) \frac{2x-x_{i-1}-x_{i+1}}{(x_i-x_{i-1})(x_i-x_{i+1})}$$

$$+ f(x_{i+1}) \frac{2x-x_{i-1}-x_i}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)}$$

$$f''(x) \approx f(x_{i-1}) \frac{2}{(x_{i-1}-x_i)(x_{i-1}-x_{i+1})}$$

$$+ f(x_i) \frac{2}{(x_i-x_{i-1})(x_i-x_{i+1})}$$

$$+ f(x_{i+1}) \frac{2}{(x_{i+1}-x_{i-1})(x_{i+1}-x_i)}$$

W szczególności, biorąc $X = X_i$
i uwzględniając $h_{i-1} = x_i - x_{i-1}$

$$h_i = x_{i+1} - x_i$$

Otrzymujemy

$$f'(x_i) \approx -\frac{h_i}{h_{i-1}(h_{i-1} + h_i)} f(x_{i-1})$$

$$+ \frac{h_i - h_{i-1}}{h_{i-1} h_i} f(x_i)$$

$$+ \frac{h_{i-1}}{h_i(h_{i-1} + h_i)} f(x_{i+1})$$

$$f''(x_i) \approx \frac{2}{h_{i-1}(h_{i-1} + h_i)} f(x_{i-1})$$

$$- \frac{2}{h_{i-1} h_i} f(x_i)$$

$$+ \frac{2}{h_i(h_{i-1} + h_i)} f(x_{i+1})$$

Równania różniczkowe (dodatek matematyczny)

Równanie różniczkowe (RR) jest to równanie zawierające nieznaną (szukaną) funkcję oraz jej pochodne.

Rozwiązywanie RR polega na wyznaczeniu tej nieznanej funkcji (której wówczas nazywamy rozwiązyaniem RR)

RR e

ZWYCZAJNE

(Szukana funkcja jest funkcją jednej zmiennej niezależnej)

CZĄSTKOWE

(szukana funkcja jest funkcją kilku zmiennych niezależnych, a w RR występują pochodne cząstkowe względem różnych zmiennych)

Rzg RR to stopień najwyższej pochodnej występującej w RR.

RR zwykle rozwija się w pewnym obszarze zmiennych niezależnych.

Aby rozwijać RR, nie wystarczy znajomość samego RR. Potrzebne są dodatkowe równania zwane warunkami dodatkowymi (brzegowymi lub początkowymi), określone na brzegu obszaru rozwiązywania, i w liczbie równej rzędowi RR.

Za pomocą metod obliczeniowych nie da się ścisłe wyznaczyć rozwiązywania RR. Można tylko obliczyć przybliżone wartości szukanej funkcji na pewnej siatce dyskretnej, lub skonstruować przybliżone wzory na szukanej funkcji.

Zastosowanie
metod różnicowych
do przybliżonego
rozwijania
zagadnienia z warunkami
brzegowymi dla
RR zwykłego 2-go rzędu

Liniowe RR zwyczajne 2-go rzędu

$$p(x) \cdot y''(x) + q(x) \cdot y'(x) + r(x) \cdot y(x) + s(x) = 0$$

$$x \in [a, b]$$

$y(x) \rightarrow$ szukana funkcja

$p(x)$
 $q(x)$
 $r(x)$
 $s(x)$

$x \rightarrow$ zmienna niezależna

$\left. \begin{array}{c} p(x) \\ q(x) \\ r(x) \\ s(x) \end{array} \right\} \rightarrow$ zadane funkcje

↑
obszar
rozwijania

Warunki brzegowe

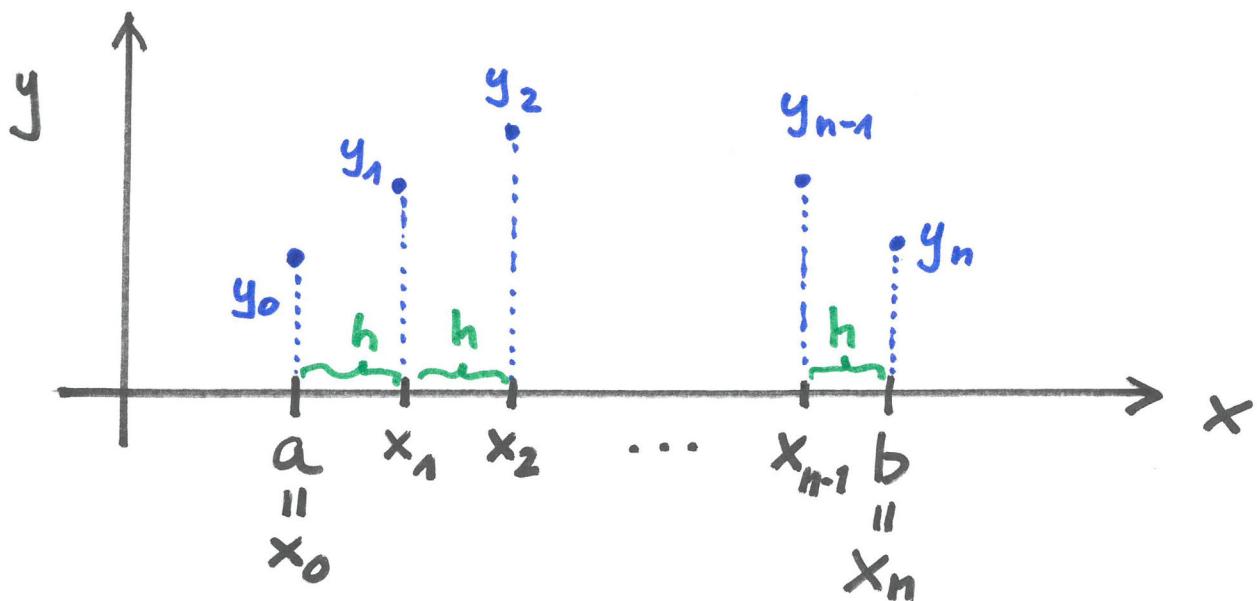
(liniowe, mieszane) :

$$\alpha \cdot y'(a) + \beta \cdot y(a) + \gamma = 0$$

$$\varphi \cdot y'(b) + \psi \cdot y(b) + \Theta = 0$$

$\alpha, \beta, \gamma, \varphi, \psi, \Theta \rightarrow$ zadane współczynniki

Zasady dyskretyzacji różnicowej -



x_0, x_1, \dots, x_n siatka jednorodna o kroku h
(dla ułatwienia)

y_0, y_1, \dots, y_n przybliżone wartości szukanej
funkcji w węzłach

UWAGA: Należy odróżnić
 y_i od $y(x_i)$!

↑
to są przybliżone
wartości

↑
to są ścisłe
wartości

W każdym węźle siatki zastępujemy RR lub warunki brzegowe przez odpowiednie równanie różnicowe stosując następujące podstawienia:

$$x \leftarrow x_i$$

$$y(x) \leftarrow y_i$$

$y'(x) \leftarrow$ przybliżenie różnicowe w punkcie x_i

$y''(x) \leftarrow$ przybliżenie różnicowe w punkcie x_i

Na przykład, stosując przybliżenia dwupunktowe na pierwsze pochodne w warunkach brzegowych, oraz trzypunktowe centralne w RR uzyskamy:

$$\left\{ \begin{array}{l} \alpha \frac{y_1 - y_0}{h} + \beta y_0 + \gamma = 0 \\ \vdots \\ p(x_i) \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} + q(x_i) \frac{y_{i+1} - y_{i-1}}{2h} + r(x_i) y_i + s(x_i) = 0 \\ \vdots \\ \phi \frac{y_n - y_{n-1}}{h} + \psi y_n + \theta = 0 \end{array} \right.$$

dla $i = 1, \dots, n-1$

Jest to układ $n+1$ liniowych równań algebraicznych
> niewiadomymi y_0, \dots, y_n

W postaci wektorowo-macierzowej:

$$A \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} -\gamma \\ -s(x_1) \\ \vdots \\ -s(x_{n-1}) \\ -\theta \end{bmatrix}$$

gdzie

$$\begin{bmatrix} (\beta - \frac{\alpha}{h}) & \frac{\alpha}{h} \\ & \end{bmatrix}$$

$$A = \begin{bmatrix} \dots & \dots & \dots \\ \left[\frac{p(x_i)}{h^2} - \frac{q(x_i)}{2h} \right] & \left[r(x_i) - \frac{2p(x_i)}{h^2} \right] & \left[\frac{p(x_i)}{h^2} + \frac{q(x_i)}{2h} \right] \\ \dots & \dots & \dots \\ & & \begin{bmatrix} -\frac{\phi}{h} & \left(\frac{\phi}{h} + \psi \right) \end{bmatrix} \end{bmatrix}$$

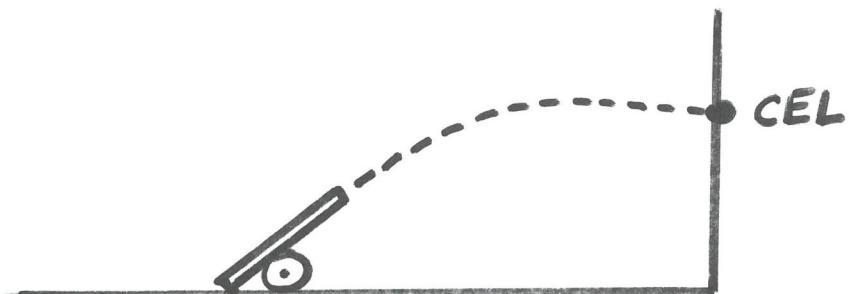
Jest to macierz TRÓJDIAGONALNA

Mozemy zastosowac algorytm Thomasa

Alternatywny sposób wyznaczenia

y_0, y_1, \dots, y_n : Metoda strzałów

Analogia do strzelania z armaty:



Aby trafić w cel należy odpowiednio ustawić lufę armaty (kąt nachylenia lufy).

Procedura:

- 1) Zakładamy dowolną wartość $p = y'(a)$.
- 2) Mamy zatem

$$\begin{cases} \alpha p + \beta y_0 + \gamma = 0 \\ \frac{y_1 - y_0}{h} = p \end{cases} \quad \Rightarrow \begin{array}{l} \text{stąd wyliczamy} \\ y_0, y_1 \end{array}$$

- 3) Z postaci dyskretnej RR wyliczamy po kolei y_2, y_3, \dots, y_n . Jest to możliwe, bo dla danego "i" znamy już y_{i-1} oraz y_i . Jedynie nie znamy y_{i+1}

4) Obliczamy reziduum drugiego warunku
brzegowego:

$$R(p) = \phi \frac{y_n - y_{n-1}}{h} + \psi y_n + \Theta$$

5) Musimy teraz tak dobrac p , aby uzyskac

$$R(p) = 0$$

Jest to rownanie nieliniowe, do którego
stosujemy dowolna metoda iteracyjna.

Zastosowanie
metod różnicowych
do przybliżonego rozwiązywania
zagadnienia z warunkiem
początkowym dla
RR zwykłego 1-go rzędu

RR pierwszego rzędu (postać ogólna)

$$y'(t) - f(t, y(t)) = 0$$

$y(t)$ → szukana funkcja

$f(t, y)$ → zadana funkcja

t → zmienna niezależna (często jest to czas)

Warunek początkowy

$$y(0) = \alpha$$

α — zadana wartość

Wyróżniamy kilka typów RR pierwszego rzędu:

y, f rzeczywiste i $\frac{\partial f}{\partial y} < 0 \Rightarrow$ RR typu "rozpadu"

y, f rzeczywiste i $\frac{\partial f}{\partial y} > 0 \Rightarrow$ RR typu "wzrostu"

y, f zespolone i $\frac{\partial f}{\partial y}$ urojone \Rightarrow RR typu "osyłajczego"

- Przykłady:

$$y'(t) + \frac{y(t)}{\Theta} = 0$$

RR typu "rozpadu"

opisuje: rozpad promieniotwórczy z okresem połowiecznym zaniku Θ

rozkład substancji chemicznej ze stałą szybkości Θ^{-1}

zanik prądu elektrycznego w obwodzie LR , gdzie $\Theta = L/R$

$$y'(t) - k y(t) = 0$$

RR typu "wzrostu"

opisuje: wzrost stężenia produktu w reakcji chemicznej autokatalitycznej (np. wybuchy) o stałej szybkości k .

równanie oscylatora harmonicznego

$x''(t) + \omega^2 x(t) = 0$ zapisujemy jako układ:

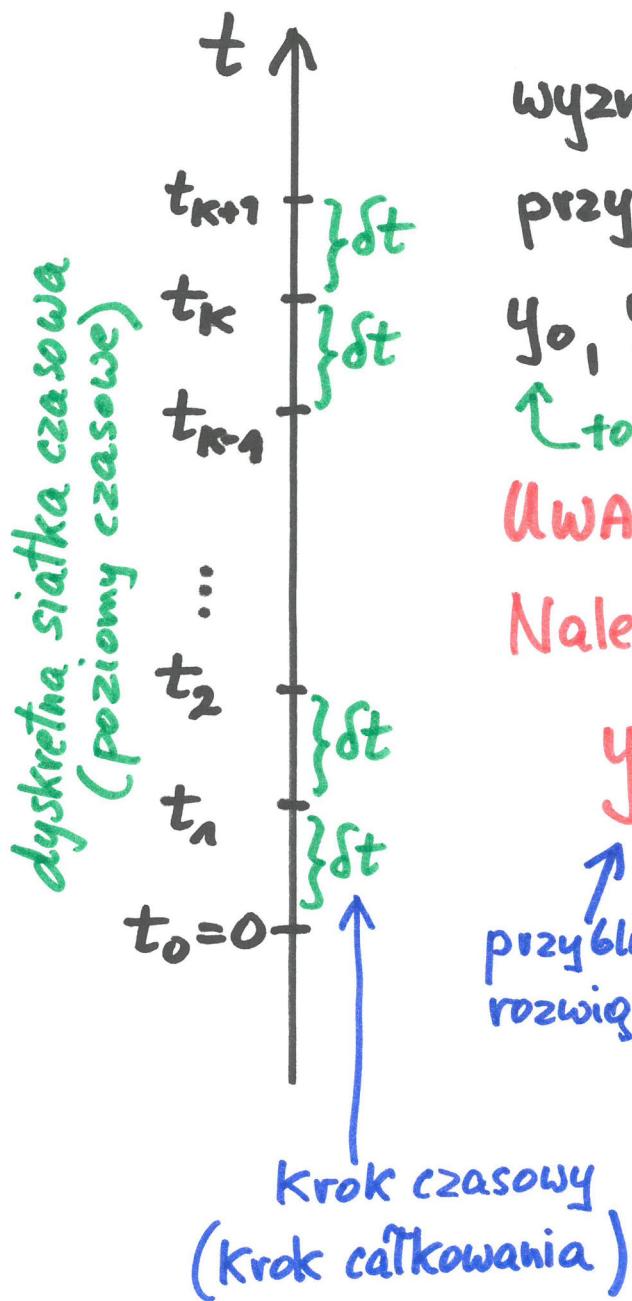
$$\begin{cases} x'(t) - \omega \nu(t) = 0 \\ \nu'(t) + \omega x(t) = 0 \end{cases} \quad \text{gdzie } \nu(t) = \frac{1}{\omega} x'(t)$$

i wprowadzamy $y(t) = x(t) + j \nu(t)$ zespolone co daje

$$y'(t) + j\omega y(t) = 0$$

RR typu "oscylacyjnego"

Koncepcja metod różnicowych dla RR pierwszego rzędu



wyznaczamy kolejne
przybliżone wartości rozwiązań:
 $y_0, y_1, y_2, \dots, y_{K-1}, y_K$, itd.

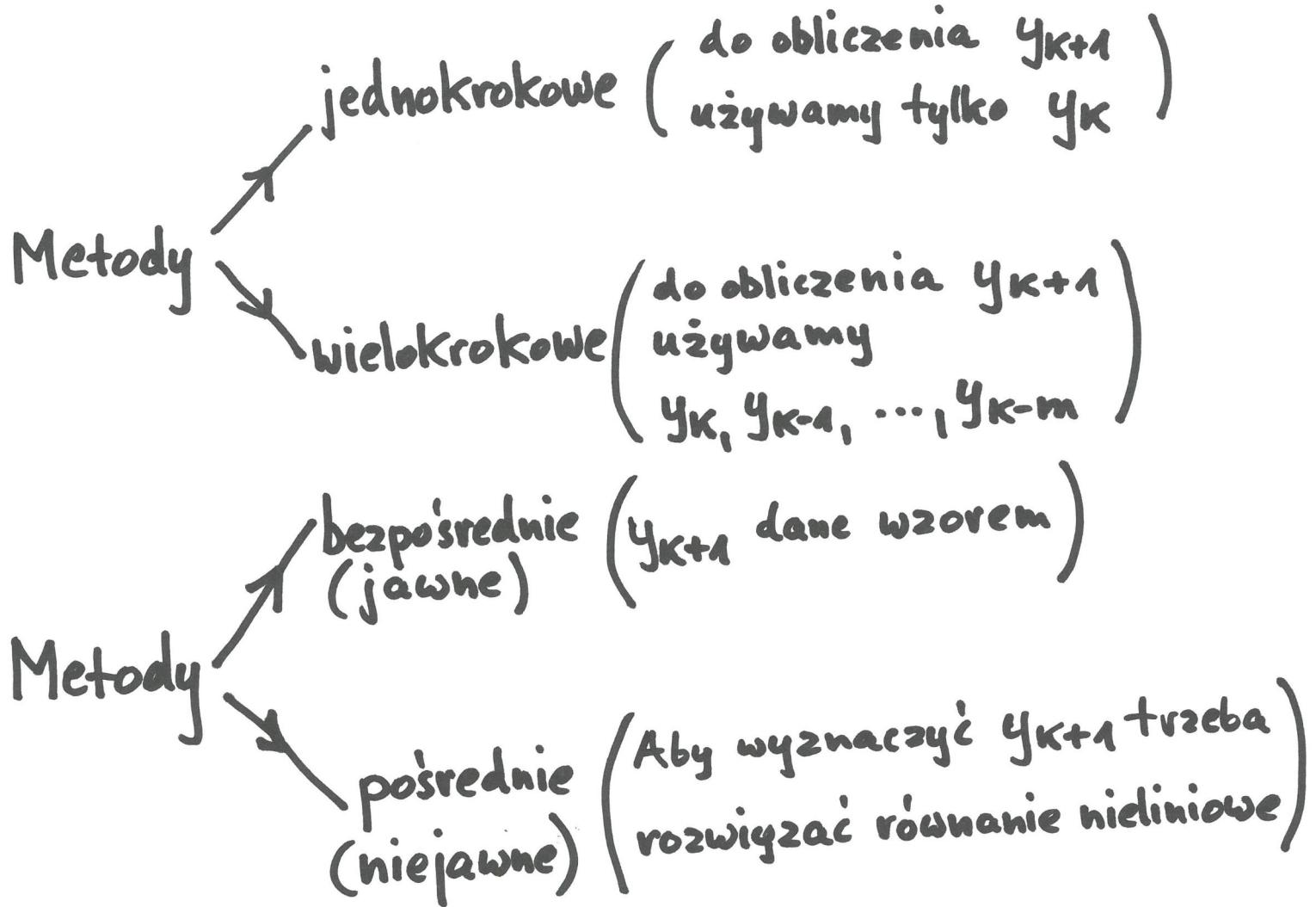
UWAGA:

Należy odróżnić

y_K od $y(t_K)$

↑
przybliżone wartości
rozwiązań

↑
ścisłe wartości
rozwiązań



Metody powinny być:

ZGODNE – Równanie różnicowe staje się identyczne z RR gdy $\delta t \rightarrow 0$

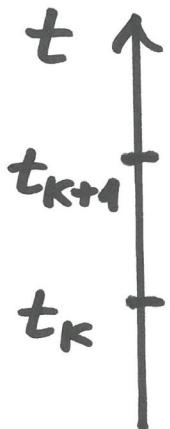
ZBIEŻNE – $y_k \rightarrow y(t_k)$ gdy $\delta t \rightarrow 0$

STABILNE – $\exists t$ gdy przenoszony z poprzednich poziomów czasowych jest tłumioły

DOKŁADNE – y_k jest dobrym przybliżeniem $y(t_k)$

EFEKTYWNE (WYDAJNE) – dają dobrą dokładność niskim kosztem obliczeniowym

① Metoda bezpośrednia Eulera (BME) (jednokrokowa)



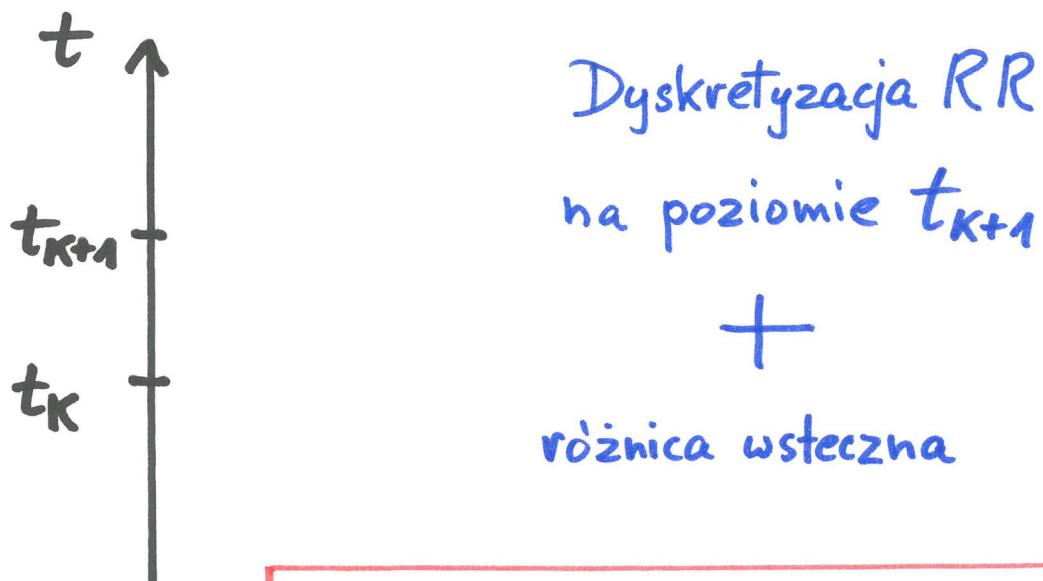
Dyskretyzacja RR
na poziomie t_K
+
różnica progresywna

Równanie
różnicowe:

$$\frac{y_{K+1} - y_K}{\delta t} - f(t_K, y_K) = 0$$

Stąd
jawny wzór: $y_{K+1} = y_K + f(t_K, y_K) \cdot \delta t$

② Metoda pośrednia Eulera (PME) (jednokrokowa)

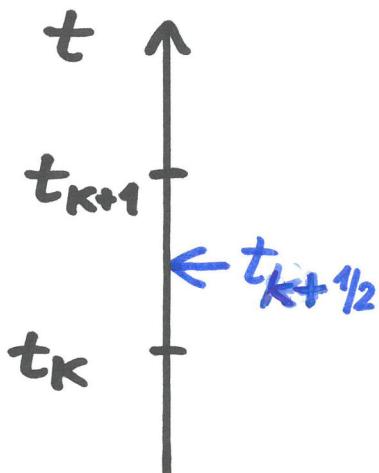


Równanie
różnicowe

$$\frac{y_{K+1} - y_K}{\delta t} - f(t_{K+1}, y_{K+1}) = 0$$

niewiadoma:
trzeba rozwiązać równanie
nieliniowe

③ Metoda pośrednia trapezów (PMT) (jednokrokowa)



Dyskretyzacja RR
w połowie drogi między t_k i t_{k+1} :

$$t_{k+1/2} = \frac{t_k + t_{k+1}}{2}$$

+

różnica centralna

Równanie
różnicowe:

$$\frac{y_{k+1} - y_k}{\delta t} - \frac{f(t_k, y_k) + f(t_{k+1}, y_{k+1})}{2} = 0$$

niewiadoma,
 równanie
 nielinowe

Dokładność metod możemy analizować obliczając tzw. lokalny błąd obiekcia dla równania różnicowego. Obliczamy go podstawiając $y_k \leftarrow y(t_k)$ i stosując rozwinięcia w szereg.

Przykład: BME:

równanie różnicowe

$$\frac{y_{k+1} - y_k}{\delta t} - f(t_k, y_k) = 0$$

lokalny błąd obiekcia $T_{k+1} = \frac{y(t_{k+1}) - y(t_k)}{\delta t} - f(t_k, y(t_k)) =$

Korzystamy z błądu obiekcia dla różnicy progresywnej

$$= y'(t_k) + \frac{1}{2} y''(t_k) \delta t + \dots - f(t_k, y(t_k)) =$$

$= 0$ na podstawie RR

$$T_{k+1} = \frac{1}{2} y''(t_k) \delta t + \dots = O(\delta t)$$

Zatem BME ma dokładność 1-go rzędu i jest zgodna, bo $T_{k+1} \rightarrow 0$ gdy $\delta t \rightarrow 0$

Analiza stabilności wymaga przyjśzenia się błądów rozwiązań, i wyznaczenia współczynnika wzmacniania błędów.

Przykład: BME

$$\begin{cases} y_{k+1} = y_k + f(t_k, y_k) \delta t \\ y(t_{k+1}) = y(t_k) + f(t_k, y(t_k)) \delta t + T_{k+1} \delta t \end{cases}$$

Błąd rozwiązań (globalne) :

$$e_k = y(t_k) - y_k$$

$$e_{k+1} = y(t_{k+1}) - y_{k+1}$$

zatem

$$e_{k+1} = e_k + [f(t_k, y(t_k)) - f(t_k, y_k)] \delta t + T_{k+1} \delta t$$

ale $f(t_k, y_k) \approx f(t_k, y(t_k)) + \frac{\partial f}{\partial y} \bigg|_{\substack{t_k \\ y(t_k)}} \underbrace{(y_k - y(t_k))}_{\text{"}} - e_k$

stąd

$$e_{k+1} = e_k + \frac{\partial f}{\partial y} \bigg|_{\substack{t_k \\ y(t_k)}} e_k \delta t + T_{k+1} \delta t$$

współczynnik wzmacniania błędu g_K

$$e_{K+1} \approx \left[1 + \frac{\partial f}{\partial y} \left. \frac{\delta t}{y(t_K)} \right|_{t_K} \right] e_K + \sum_{k=1}^{K+1} \delta t$$

Błąd przeniesiony
z poziomu t_K

Błąd lokalny
rozwijania
(wytworzony przy
przejściu z t_K na t_{K+1})

$$g_K = 1 + \frac{\partial f}{\partial y} \cdot \delta t$$

współczynnik
wzmacniania
błędu dla BME

Dla stabilności numerycznej musi być

$$|g_K| \leq 1$$

Dla równania rozpadu $\frac{\partial f}{\partial y} < 0$

$$\left| 1 + \frac{\partial f}{\partial y} \delta t \right| \leq 1$$

$$-1 \leq 1 + \frac{\partial f}{\partial y} \delta t \leq 1 \quad /-1$$

$$-2 \leq \frac{\partial f}{\partial y} \delta t \leq 0 \quad / \cdot (-1)$$

zawsze spełniona bo $\frac{\partial f}{\partial y} < 0$

$$2 \geq \left(-\frac{\partial f}{\partial y} \right) \delta t$$

$$\delta t \leq \frac{2}{\left(-\frac{\partial f}{\partial y} \right)}$$

warunek stabilności
dla równania
rozpadu

Wniosek: Dla równania rozpadu

BME jest WARUNKOWO stabilna: tzn.
tylko gdy δt jest odpowiednio
mały.

Uwaga: Dla każdej KONKRETNEJ
wartości δt BME jest
albo stabilna albo niestabilna.

Dla równania wzrostu $\frac{\partial f}{\partial y} > 0$

a zatem

$$g_K = 1 + \frac{\partial f}{\partial y} \delta t > 1 \Rightarrow |g_K| > 1$$

a więc BME jest BEZWARUNKOWO niestabilna
(niezależnie od wyboru δt)

Dla równania oscylacyjnego $\frac{df}{dy} = jA$

$$g_K = 1 + jA \delta t$$

$$|g_K| = \sqrt{1 + (A \delta t)^2} > 1 \Rightarrow$$

BME jest BEZWARUNKOWO Niestabilna

Analogiczna analiza dokładności i stabilności dla PME

pokazuje, że:

$$T_{K+1} = -\frac{1}{2} y''(t_{K+1}) \delta t + \dots = O(\delta t)$$

więc metoda ma dokładność 1-go rzędu

$$e_{K+1} \approx \underbrace{\frac{1}{1 - \frac{\partial f}{\partial y} \delta t} e_K}_{\text{błąd przeniesiony z poziomu } t_K} + \underbrace{\frac{T_{K+1} \delta t}{1 - \frac{\partial f}{\partial y} \delta t}}_{\text{błąd lokalny rozwiązania}}$$

współczynnik
wzmocnienia błędu

$$g_K = \frac{1}{1 - \frac{\partial f}{\partial y} \delta t}$$

Dla równania rozpadu $\frac{\partial f}{\partial y} < 0$

$$\frac{\partial f}{\partial y} < 0 \Rightarrow 1 - \frac{\partial f}{\partial y} \delta t > 1 \Rightarrow \frac{1}{1 - \frac{\partial f}{\partial y} \delta t} < 1$$

$$\Rightarrow |g_K| < 1$$

PME jest BEZWARUNKOWO stabilna

Dla równania wzrostu $\frac{\partial f}{\partial y} > 0$

$$\frac{\partial f}{\partial y} > 0 \Rightarrow \text{dla małych } \delta t \quad 1 - \frac{\partial f}{\partial y} \delta t < 1 \Rightarrow$$

$$\frac{1}{1 - \frac{\partial f}{\partial y} \delta t} > 1 \Rightarrow |g_K| > 1$$

PME jest BEZWARUNKOWO niestabilna

Dla równania oscylacyjnego $\frac{\partial f}{\partial y} = jA$

$$|g_K| = \left| \frac{1}{1 - jA \delta t} \right| = \frac{1}{\sqrt{1 + (A \delta t)^2}} < 1$$

PME jest BEZWARUNKOWO stabilna

Analogiczna analiza dokładności i stabilności dla PMT

pokazuje, że

$$T'_{K+1} = O(\delta t^2)$$

więc metoda ma dokładność 2-go rzędu

$$e_{K+1} = \frac{1 + \frac{\partial f}{\partial y} \frac{\delta t}{2}}{1 - \frac{\partial f}{\partial y} \frac{\delta t}{2}} e_K + \frac{T_{K+1} \delta t}{1 - \frac{\partial f}{\partial y} \frac{\delta t}{2}}$$

błąd przeniesiony

z poziomu t_K

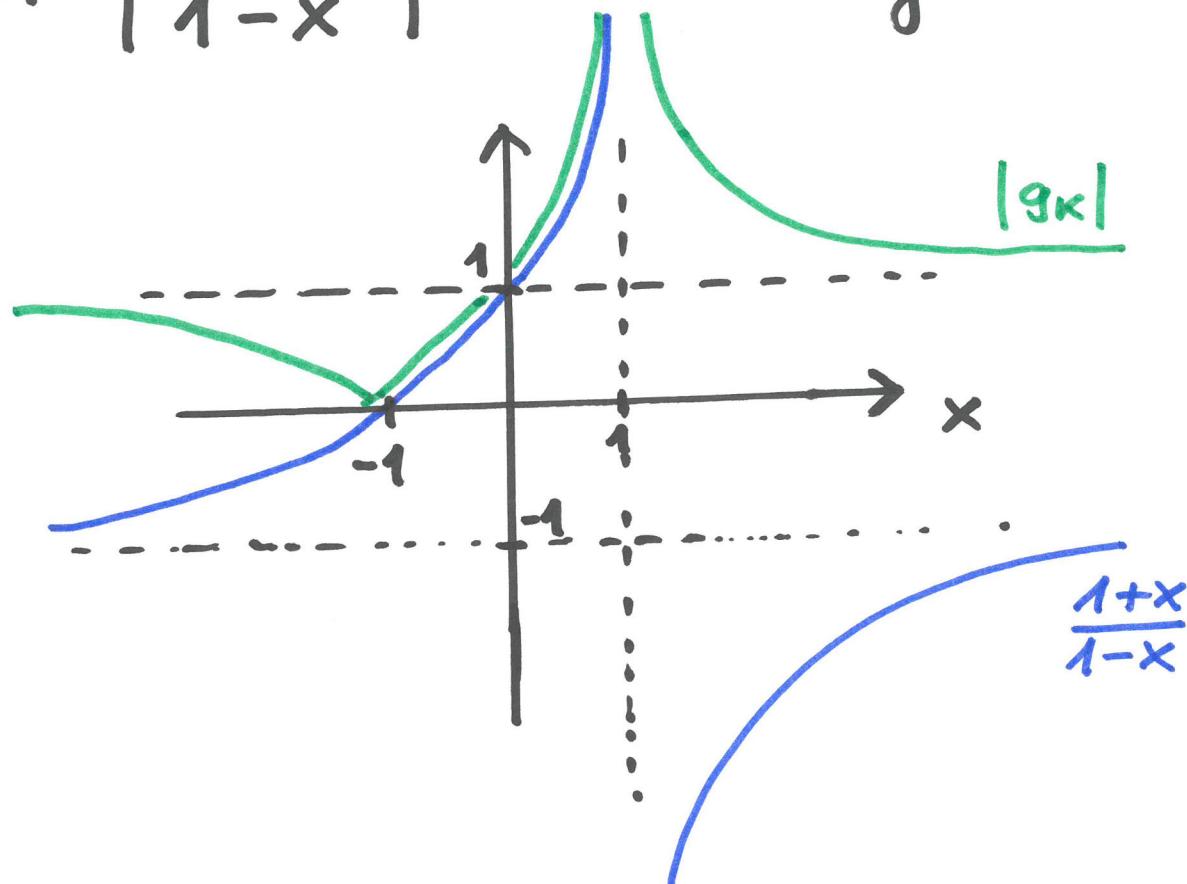
błąd lokalny
rozwiązań

współczynnik
wzmocnienia
błędu

$$g_K = \frac{1 + \frac{\partial f}{\partial y} \frac{\delta t}{2}}{1 - \frac{\partial f}{\partial y} \frac{\delta t}{2}}$$

Dla równania rozpadu $\frac{\partial f}{\partial y} < 0$

$$|g_K| = \left| \frac{1+x}{1-x} \right| \quad \text{gdzie } x = \frac{\partial f}{\partial y} \frac{\delta t}{2}$$



$$\frac{\partial f}{\partial y} < 0 \Rightarrow x < 0 \Rightarrow |g_K| < 1 \Rightarrow$$

PMT jest BEZWARUNKOWO stabilna

Dla równania wzrostu $\frac{\partial f}{\partial y} > 0$

$$\frac{\partial f}{\partial y} > 0 \Rightarrow x > 0 \Rightarrow |g_K| > 1 \Rightarrow$$

PMT jest BEZWARUNKOWO niestabilna

Dla równania oscylacyjnego $\frac{\partial f}{\partial y} = jA$

$$|g_K| = \left| \frac{1 + jA \frac{\delta t}{2}}{1 - jA \frac{\delta t}{2}} \right| = \frac{\sqrt{1 + (A \frac{\delta t}{2})^2}}{\sqrt{1 + (A \frac{\delta t}{2})^2}} = 1 \Rightarrow$$

PMT jest stabilna, ale na granicy stabilności

Podsumowanie wniosków dotyczących

stabilności numerycznej -

Metoda	TYP RR	rozpadu	wzrostu	osyłacyjne
BME		stabilna tylko gdy $\delta t \leq -\frac{2}{\partial f / \partial y}$	niestabilna	niestabilna
PME		zawsze stabilna	niestabilna	zawsze stabilna
PMT		zawsze stabilna	niestabilna	stabilna, ale na granicy stabilności

Równania różniczkowe cząstkowe
drugiego rzędu, zależne od dwóch
zmiennych niezależnych,
całkowicie liniowe
(dodatek matematyczny)

Ogólna postać takich RR:

$$a \frac{\partial^2 u}{\partial x^2} + b \frac{\partial^2 u}{\partial x \partial y} + c \frac{\partial^2 u}{\partial y^2}$$

$$+ d \frac{\partial u}{\partial x} + e \frac{\partial u}{\partial y} + f u + g = 0$$

$x, y \rightarrow$ zmienne niezależne

$u = u(x, y) \rightarrow$ szukana funkcja

$$a = a(x, y)$$

$$b = b(x, y)$$

$$c = c(x, y)$$

$$d = d(x, y)$$

$$e = e(x, y)$$

$$f = f(x, y)$$

$$g = g(x, y)$$

} → zadane funkcje

Wyróżnik równania

$$\Delta \stackrel{\text{df}}{=} b^2 - 4ac$$

Klasyfikacja równań:

$\Delta < 0 \Rightarrow$ ELIPTYCZNE

$\Delta = 0 \Rightarrow$ PARABOLICZNE

$\Delta > 0 \Rightarrow$ HIPERBOLICZNE

Jeżeli a, b, c zależą od x, y to
równanie może być różnego typu dla
różnych wartości x, y .

Przykłady -

① Jednowymiarowe równanie dyfuzji (lub transportu ciepła)

$$\frac{\partial u(x,t)}{\partial t} = D \frac{\partial^2 u(x,t)}{\partial x^2}$$

współczynnik dyfuzji

(rolę y pełni zmienna t) x - współrzędna przestrzenna
 t - czas

$$D \frac{\partial^2 u}{\partial x^2} + 0 \cdot \frac{\partial^2 u}{\partial x \partial t} + 0 \cdot \frac{\partial^2 u}{\partial t^2} - \frac{\partial u}{\partial t} = 0$$

$\begin{matrix} \parallel & \parallel & \parallel \\ a & b & c \end{matrix}$

$$\Delta = b^2 - 4ac = 0 - 4 \cdot 0 \cdot D = 0$$

RÓWNAŃE PARABOLICZNE

② Dwuwymiarowe równanie Poissona
na rozkład potencjału pola elektrostatycznego

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} = -\rho(x,y)$$

x, y - współrzędne przestrzenne

↑
gęstość ładunków elektrycznych

$$1 \cdot \frac{\partial^2 u}{\partial x^2} + 0 \cdot \frac{\partial^2 u}{\partial x \partial y} + 1 \cdot \frac{\partial^2 u}{\partial y^2} + \rho(x,y) = 0$$

" " " "

a b c

$$\Delta = b^2 - 4ac = 0 - 4 \cdot 1 \cdot 1 = -4 < 0$$

RÓWNANIE ELIPTYCZNE

③ Jednowymiarowe równanie falowe (np. drgania struny)

$$\frac{\partial^2 u(x,t)}{\partial t^2} - \omega^2 \frac{\partial^2 u(x,t)}{\partial x^2} = 0$$

$$\omega^2 = \frac{\gamma}{m} \leftarrow \begin{array}{l} \text{napiącie struny} \\ \text{masa na jednostkę} \end{array}$$

$\frac{\gamma}{m}$ ← $\begin{array}{l} \text{długość} \\ x - \text{współrzędna przestrzenna} \end{array}$

(rolę y pełni zmienność t) t - czas

$$-\omega^2 \frac{\partial^2 u}{\partial x^2} + 0 \cdot \frac{\partial^2 u}{\partial x \partial t} + 1 \cdot \frac{\partial^2 u}{\partial t^2} = 0$$

a b c

$$\Delta = b^2 - 4ac = 0^2 - 4 \cdot (-\omega^2) \cdot 1 = 4\omega^2 > 0$$

RÓWNAŃIE HIPERBOLICZNE

Zastosowanie
metod różnicowych do
przybliżonego
rozwiązywania
równania dyfuzji
z warunkiem początkowym
i warunkami brzegowymi

Zagadnienie do rozwiązyania

$$\frac{\partial u(x,t)}{\partial t} = D \frac{\partial^2 u(x,t)}{\partial x^2}$$

$$x \in [a, b]$$

$$t \in [0, \infty)$$

$$u(x, 0) = u^*(x) \quad \leftarrow \quad \text{WARUNEK POCZĄTKOWY}$$

$$\left. \alpha(t) \frac{\partial u(x,t)}{\partial x} \right|_{x=a} + \beta(t) u(a,t) + \gamma(t) = 0$$

$$\left. \varphi(t) \frac{\partial u(x,t)}{\partial x} \right|_{x=b} + \psi(t) u(b,t) + \theta(t) = 0$$

} WARUNKI BRZEGOWE

$u(x,t) \rightarrow$ szukana funkcja

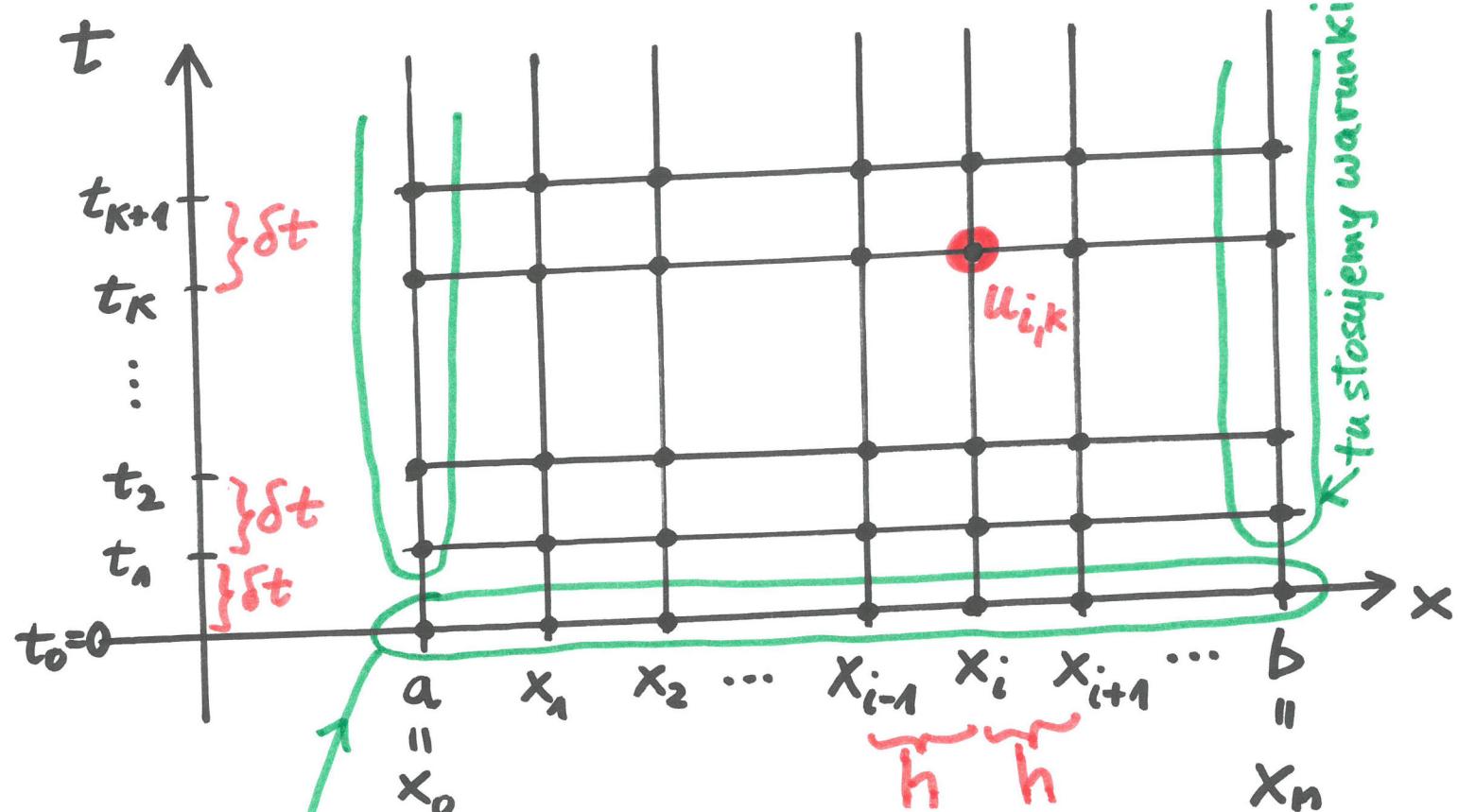
(MIESZANE,
LINIOWE)

$u^*(x) \rightarrow$ zadana funkcja

$\alpha(t)$
 $\beta(t)$
 $\gamma(t)$
 $\varphi(t)$
 $\psi(t)$
 $\theta(t)$

zadane funkcje

Zasady dyskretyzacji różnicowej



tu stosujemy warunek początkowy

Siatka czasowo-przestrzenna:

$x_0, x_1, \dots, x_n \rightarrow$ jednorodna siatka przestrzenna
o kroku h

$t_0, t_1, t_2, \dots \rightarrow$ jednorodna siatka czasowa
o kroku δt

$u_{i,k} \rightarrow$ przybliżone rozwiązanie.
w węźle x_i i na poziomie
czasowym t_k

UWAGA: Należy odróżnić $u_{i,k}$ od $u(x_i, t_k)$

wartości przybliżone →
wartości ścisłe →

Dyskretyzacja warunków brzegowych

Najprościej zastosować przybliżenia dwupunktowe na pierwsze pochodne przestrzenne:

$$\alpha(t_{K+1}) \frac{u_{1,K+1} - u_{0,K+1}}{h} + \beta(t_{K+1}) u_{0,K+1} + \gamma(t_{K+1}) = 0$$

$$\phi(t_{K+1}) \frac{u_{n,K+1} - u_{n-1,K+1}}{h} + \psi(t_{K+1}) u_{n,K+1} + \theta(t_{K+1}) = 0$$

UWAGA: Zawsze robimy dyskretyzację na poziomie czasowym t_{K+1}
(tak jest najwygodniej)

Dyskretyzacja równania dyfuzji

① Klasyczna metoda bezpośrednia (KMB)

Trzypunktowe
przybliżenie
centralne
na $\frac{\partial^2 u}{\partial x^2}$

+ MBE

$$\frac{u_{i+1,k} - u_{i,k}}{\delta t} = D \frac{u_{i-1,k} - 2u_{i,k} + u_{i+1,k}}{h^2}$$

Stąd

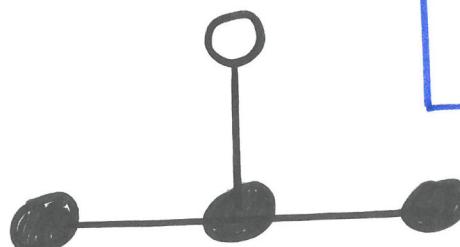
$$u_{i,k+1} = \gamma u_{i-1,k} + (1-2\gamma)u_{i,k} + \gamma u_{i+1,k}$$

gdzie

$$\gamma = D \frac{\delta t}{h^2}$$

UWAGA:
Metoda stabilna
tylko gdy $\gamma \leq \frac{1}{2}$

"Molekuła obliczeniowa"



Dokładność:
 $T = O(\delta t, h^2)$

Po obliczeniu $u_{i,k+1}$ dla $i = 1, \dots, n-1$
obliczamy $u_{0,k+1}$ oraz $u_{n,k+1}$ z warunków brzegowych

② Metoda Laasonen (ML)

Trzypunktowe
przybliżenie
centralne
 $\frac{\partial^2 u}{\partial x^2}$

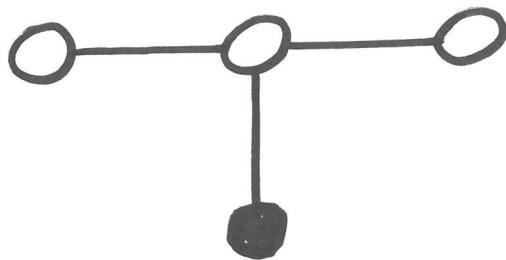
+ PME

$$\frac{u_{i,K+1} - u_{i,K}}{\delta t} = D \frac{u_{i-1,K+1} - 2u_{i,K+1} + u_{i+1,K+1}}{h^2}$$

$$\lambda u_{i-1,K+1} - (1+2\lambda)u_{i,K+1} + \lambda u_{i+1,K+1} = -u_{i,K}$$

niewiadome

"Molekuła obliczeniowa"



Metoda bezwarunkowo
stabilna

(λ może być dowolne) !

Dokładność: $T = O(\delta t, h^2)$

Obliczenie $u_{i,k+1}$ dla $i = 0, 1, \dots, n$
 wymaga rozwiązyania układu liniowych
 równań algebraicznych:

$$A \begin{bmatrix} u_{0,k+1} \\ u_{1,k+1} \\ \vdots \\ u_{n-1,k+1} \\ u_{n,k+1} \end{bmatrix} = \begin{bmatrix} -\gamma(t_{k+1}) \\ -u_{1,k} \\ \vdots \\ -u_{n-1,k} \\ -\Theta(t_{k+1}) \end{bmatrix} \quad \text{gdzie}$$

$$A = \begin{bmatrix} -\frac{\alpha(t_{k+1})}{h} + \beta(t_{k+1}) & \frac{\alpha(t_{k+1})}{h} & & & \\ & \ddots & \ddots & \ddots & \\ & \lambda & -(1+2\lambda) & \lambda & \dots \\ & \dots & \dots & \dots & \dots \\ & & & -\frac{\varphi(t_{k+1})}{h} & \frac{\varphi(t_{k+1})}{h} + \psi(t_{k+1}) \end{bmatrix}$$

Jest to macierz TRÓJDIAGONALNA

Mozemy zastosowac algorytm Thomasa

③ Metoda Cranka-Nicolson (MCN)

Trzypunktowe
przybliżenia
centralne
na $\frac{\partial^2 u}{\partial x^2}$

+

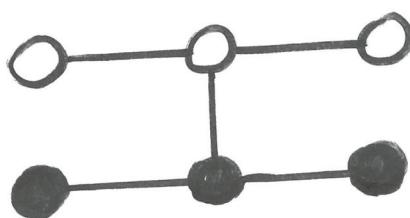
PMT

$$\frac{u_{i,K+1} - u_{i,K}}{\delta t} = \frac{1}{2} \left[D \frac{u_{i-1,K+1} - 2u_{i,K+1} + u_{i+1,K+1}}{h^2} + D \frac{u_{i-1,K} - 2u_{i,K} + u_{i+1,K}}{h^2} \right]$$

$$\frac{\lambda}{2} u_{i-1,K+1} - (1+\lambda) u_{i,K+1} + \frac{\lambda}{2} u_{i+1,K+1} = - \left[\frac{\lambda}{2} u_{i-1,K} + (1-\lambda) u_{i,K} + \frac{\lambda}{2} u_{i+1,K} \right]$$

niewiadome

"Molekuła obliczeniowa"



Metoda bezwarunkowo stabilna !

(λ może być dowolne)

!

Dokładność: $T = O(\delta t^2, h^2)$

Obliczenie $u_{i,k+1}$ dla $i = 0, 1, \dots, n$
 wymaga rozwiązyania układu liniowych
 równań algebraicznych:

$$A \begin{bmatrix} u_{0,k+1} \\ u_{1,k+1} \\ \vdots \\ u_{n-1,k+1} \\ u_{n,k+1} \end{bmatrix} = \begin{bmatrix} -\gamma(t_{k+1}) \\ \vdots \\ -\left[\frac{\lambda}{2}u_{i-1,k} + (1-\lambda)u_{i,k} + \frac{\lambda}{2}u_{i+1,k}\right] \\ \vdots \\ -\theta(t_{k+1}) \end{bmatrix}$$

gdzie

$$A = \begin{bmatrix} -\frac{\alpha(t_{k+1})}{h} + \beta(t_{k+1}) & \frac{\alpha(t_{k+1})}{h} & & & \\ & \ddots & \ddots & \ddots & \\ & \frac{\lambda}{2} & -(1+\lambda) & \frac{\lambda}{2} & \\ & & \ddots & \ddots & \ddots \\ & & & -\frac{\phi(t_{k+1})}{h} & \frac{\phi(t_{k+1})}{h} + \psi(t_{k+1}) \end{bmatrix}$$

Jest to macierz TRÓJDIAGONALNA

Mozemy zastosowac algorytm Thomasa

Analiza stabilności
numerycznej
metod różnicowych dla
równania dyfuzji
(metoda von Neumanna)

Założenia:

$$u_{i,k} = u(x_i, t_k) - e_{i,k}$$

↑ błąd dyskretyzacji w węźle
 x_i , na poziomie czasowym t_k

$x_i \in [a, b]$, jednorodna sieć przestrzenna
 o kroku h .

Rozwijamy $e_{i,k}$ w szereg Fouriera:

$$e_{i,k} = \sum_{j=1}^n a_k^{(j)} \exp\left(\frac{2\pi x_i}{L} \nu j\right)$$

na poziomie t_k

↑
 amplituda składowej
 Fouriera błędu

↑
 $L = b - a$

jednostka urojona $j = \sqrt{-1}$

podobnie na poziomie t_{k+1} :

$$e_{i,k+1} = \sum_{j=1}^n a_{k+1}^{(j)} \exp\left(\frac{2\pi x_i}{L} \nu j\right)$$

Ponieważ równanie dyfuzji jest liniowe, możemy badać zachowanie się poszczególnych składowych Fouriera błędów niezależnie od siebie.

Klasyczna metoda bezpośrednią

$$u_{i,K+1} = \gamma u_{i-1,K} + (1-2\gamma) u_{i,K} + \gamma u_{i+1,K}$$

równanie dyfuzji jest liniowe, więc również

$$c_{i,K+1} = \gamma c_{i-1,K} + (1-2\gamma) c_{i,K} + \gamma c_{i+1,K} .$$

Podstawiamy sktadową Fouriera:

$$\begin{aligned} a_{K+1}^{(s)} \exp\left(\frac{2\pi x_i}{L} \nu j\right) &= \gamma a_K^{(s)} \exp\left(\frac{2\pi x_{i-1}}{L} \nu j\right) \\ &+ (1-2\gamma) a_K^{(s)} \exp\left(\frac{2\pi x_i}{L} \nu j\right) \\ &+ \gamma a_K^{(s)} \exp\left(\frac{2\pi x_{i+1}}{L} \nu j\right) . \end{aligned}$$

Dzielimy przez $\exp\left(\frac{2\pi x_i}{L} \nu j\right)$, co daje

$$a_{K+1}^{(s)} = a_K^{(s)} \left[\gamma \exp\left(-\frac{2\pi h}{L} \nu j\right) + (1-2\gamma) + \gamma \exp\left(\frac{2\pi h}{L} \nu j\right) \right]$$

Współczynnik wzmacniania błędu możemy zdefiniować jako

$$g = \left| \frac{a_{K+1}^{(s)}}{a_K^{(s)}} \right|$$

Zatem

$$g = \left| \gamma \left[\exp\left(-\frac{2\pi h}{L} \nu j\right) + \exp\left(\frac{2\pi h}{L} \nu j\right) \right] + (1-2\gamma) \right| =$$
$$\left| 1 - 4\gamma \sin^2\left(\frac{\pi h}{L} \nu\right) \right|$$

Dla stabilności wymagamy $|g| \leq 1$, czyli

$$\left| 1 - 4\gamma \sin^2\left(\frac{\pi h}{L}\gamma\right) \right| \leq 1$$

Warunek musi być spełniony dla dowolnego γ , a więc
również gdy $\sin^2(\dots)$ jest największe ($=1$)

$$|1 - 4\gamma| \leq 1$$

$$-1 \leq 1 - 4\gamma \leq 1$$

\Downarrow

spełniona zawsze

$$\boxed{\gamma \leq \frac{1}{2}}$$

Warunek stabilności
metody KMB

Metoda KMB jest
warunkowo stabilna!

Metoda Laasonen

$$\lambda u_{i-1, K+1} - (1+2\lambda)u_{i, K+1} + \lambda u_{i+1, K+1} = -u_{i, K}$$

równanie dyfuzji jest liniowe, więc również

$$\lambda e_{i-1, K+1} - (1+2\lambda)e_{i, K+1} + \lambda e_{i+1, K+1} = -e_{i, K}$$

Podstawiamy sktadówkę Fouriera:

$$\begin{aligned} \lambda \alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_{i-1}}{L} \gamma_j\right) - (1+2\lambda) \alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_i}{L} \gamma_j\right) \\ + \lambda \alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_{i+1}}{L} \gamma_j\right) = -\alpha_K^{(2)} \exp\left(\frac{2\pi x_i}{L} \gamma_j\right). \end{aligned}$$

Dzielimy przez $\exp\left(\frac{2\pi x_i}{L} \gamma_j\right)$, co daje:

$$\alpha_{K+1}^{(2)} \left[\lambda \exp\left(-\frac{2\pi h}{L} \gamma_j\right) - (1+2\lambda) + \lambda \exp\left(\frac{2\pi h}{L} \gamma_j\right) \right] = -\alpha_K^{(2)}$$

stąd współczynnik wzmocnienia błędu:

$$g = \left| \frac{\alpha_{K+1}^{(2)}}{\alpha_K^{(2)}} \right| = \frac{1}{\left| \lambda \left[\exp\left(-\frac{2\pi h}{L} \gamma_j\right) + \exp\left(\frac{2\pi h}{L} \gamma_j\right) \right] - (1+2\lambda) \right|} =$$

$$= \frac{1}{\left| 1 + 4\lambda \sin^2\left(\frac{\pi h}{L} \gamma_j\right) \right|}$$

Jak widać, $g \leq 1$ dla dowolnego γ oraz λ , więc metoda Laasonen jest bezogólnie stabilna!

Metoda Cranka-Nicolson

$$\frac{1}{2}u_{i-1, k+1} - (1+\lambda)u_{i, k+1} + \frac{1}{2}u_{i+1, k+1} = \\ = - \left[\frac{\lambda}{2}u_{i-1, k} + (1-\lambda)u_{i, k} + \frac{\lambda}{2}u_{i+1, k} \right]$$

równanie dyfuzji jest liniowe, więc również

$$\frac{1}{2}e_{i-1, k+1} - (1+\lambda)e_{i, k+1} + \frac{1}{2}e_{i+1, k+1} = \\ = - \left[\frac{\lambda}{2}e_{i-1, k} + (1-\lambda)e_{i, k} + \frac{\lambda}{2}e_{i+1, k} \right].$$

Podstawiamy składową Fouriera:

$$\frac{1}{2}\alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_{i-1}}{L} \gamma j\right) - (1+\lambda)\alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_i}{L} \gamma j\right) + \frac{\lambda}{2}\alpha_{K+1}^{(2)} \exp\left(\frac{2\pi x_{i+1}}{L} \gamma j\right) \\ = - \left[\frac{\lambda}{2}\alpha_K^{(2)} \exp\left(\frac{2\pi x_{i-1}}{L} \gamma j\right) + (1-\lambda)\alpha_K^{(2)} \exp\left(\frac{2\pi x_i}{L} \gamma j\right) + \frac{\lambda}{2}\alpha_K^{(2)} \exp\left(\frac{2\pi x_{i+1}}{L} \gamma j\right) \right].$$

Dzielimy przez $\exp\left(\frac{2\pi x_i}{L} \gamma j\right)$, co daje:

$$\alpha_{K+1}^{(2)} \left[\frac{\lambda}{2} \exp\left(-\frac{2\pi h}{L} \gamma j\right) - (1+\lambda) + \frac{\lambda}{2} \exp\left(\frac{2\pi h}{L} \gamma j\right) \right] \\ = -\alpha_K^{(2)} \left[\frac{\lambda}{2} \exp\left(-\frac{2\pi h}{L} \gamma j\right) + (1-\lambda) + \frac{\lambda}{2} \exp\left(\frac{2\pi h}{L} \gamma j\right) \right].$$

Stąd współczynnik wzmacniania będzie:

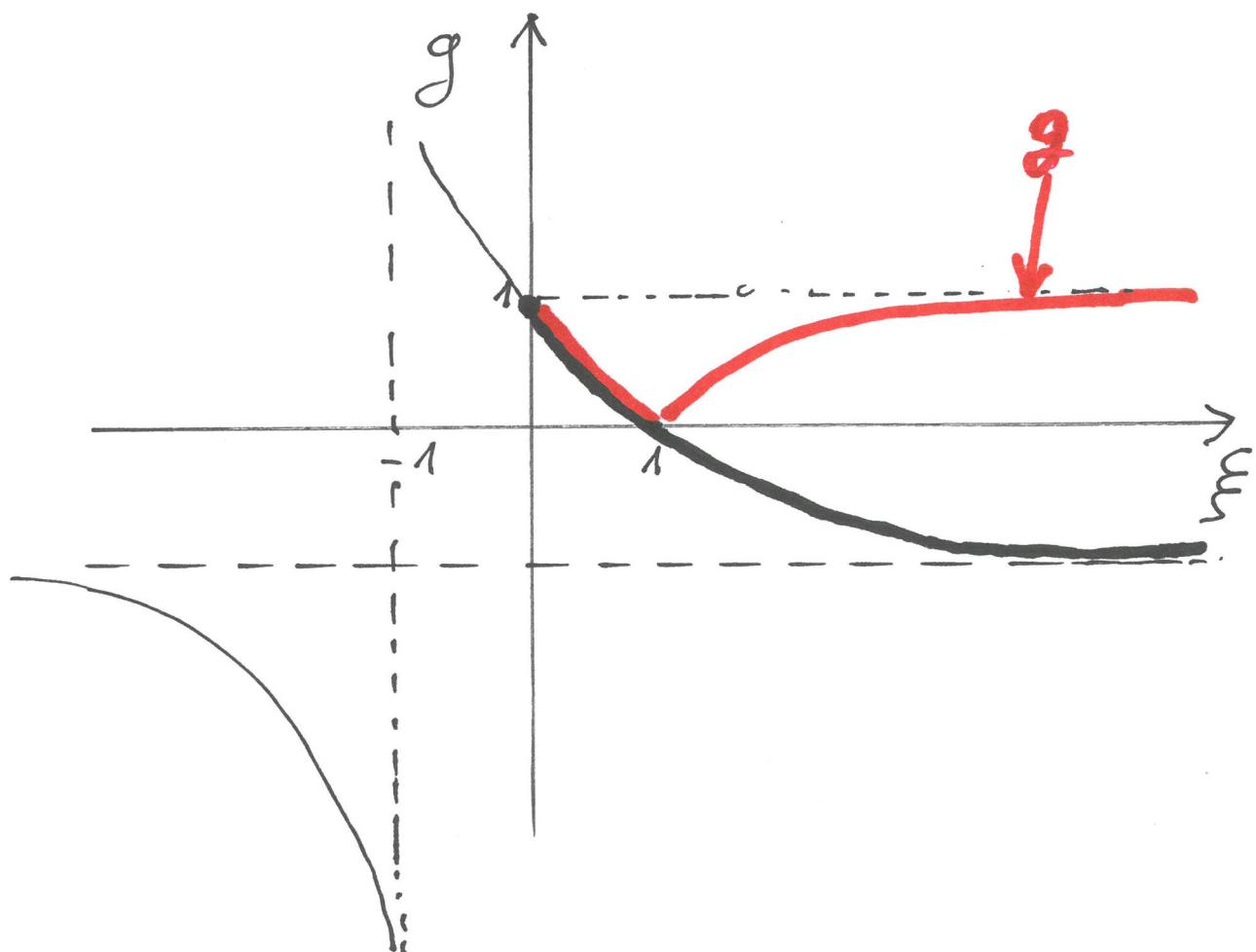
$$g = \left| \frac{\alpha_{K+1}^{(2)}}{\alpha_K^{(2)}} \right| = \left| \frac{\frac{1}{2} \left[\exp\left(-\frac{2\pi h}{L} \gamma j\right) + \exp\left(\frac{2\pi h}{L} \gamma j\right) \right] + (1-\lambda)}{\frac{1}{2} \left[\exp\left(-\frac{2\pi h}{L} \gamma j\right) + \exp\left(\frac{2\pi h}{L} \gamma j\right) \right] - (1+\lambda)} \right|$$

$$g = \left| \frac{1 - 2\lambda \sin^2\left(\frac{\pi h}{L} \nu\right)}{1 + 2\lambda \sin^2\left(\frac{\pi h}{L} \nu\right)} \right|$$

Oznaczmy

$$\xi = 2\lambda \sin^2\left(\frac{\pi h}{L} \nu\right)$$

Mamy zatem $g = \left| \frac{1-\xi}{1+\xi} \right|$ gdzie $\xi \geq 0$



A zatem $g \leq 1$ czyli metoda Cranka-Nicolson jest bezwzględnie stabilna (niezależnie od wartości ν oraz λ)

Zastosowanie
metod różnicowych do
przybliżonego
rozwiązańia
równania Poissona
z warunkami brzegowymi

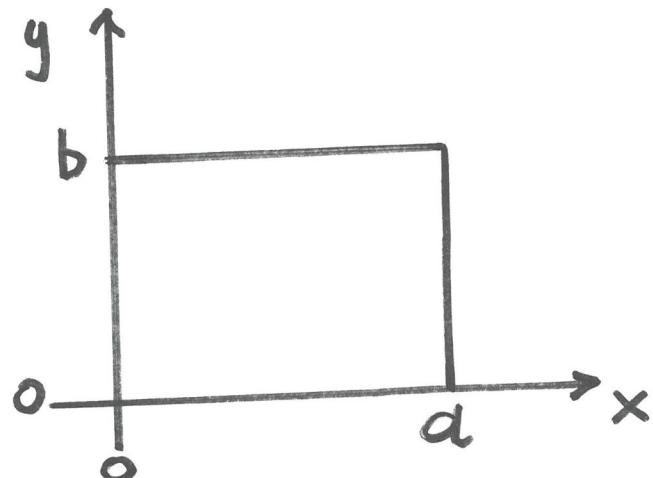
Zagadnienie do rozwiązania

$$\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2} + p(x,y) = 0$$

$$x \in [0, a]$$

$$y \in [0, b]$$

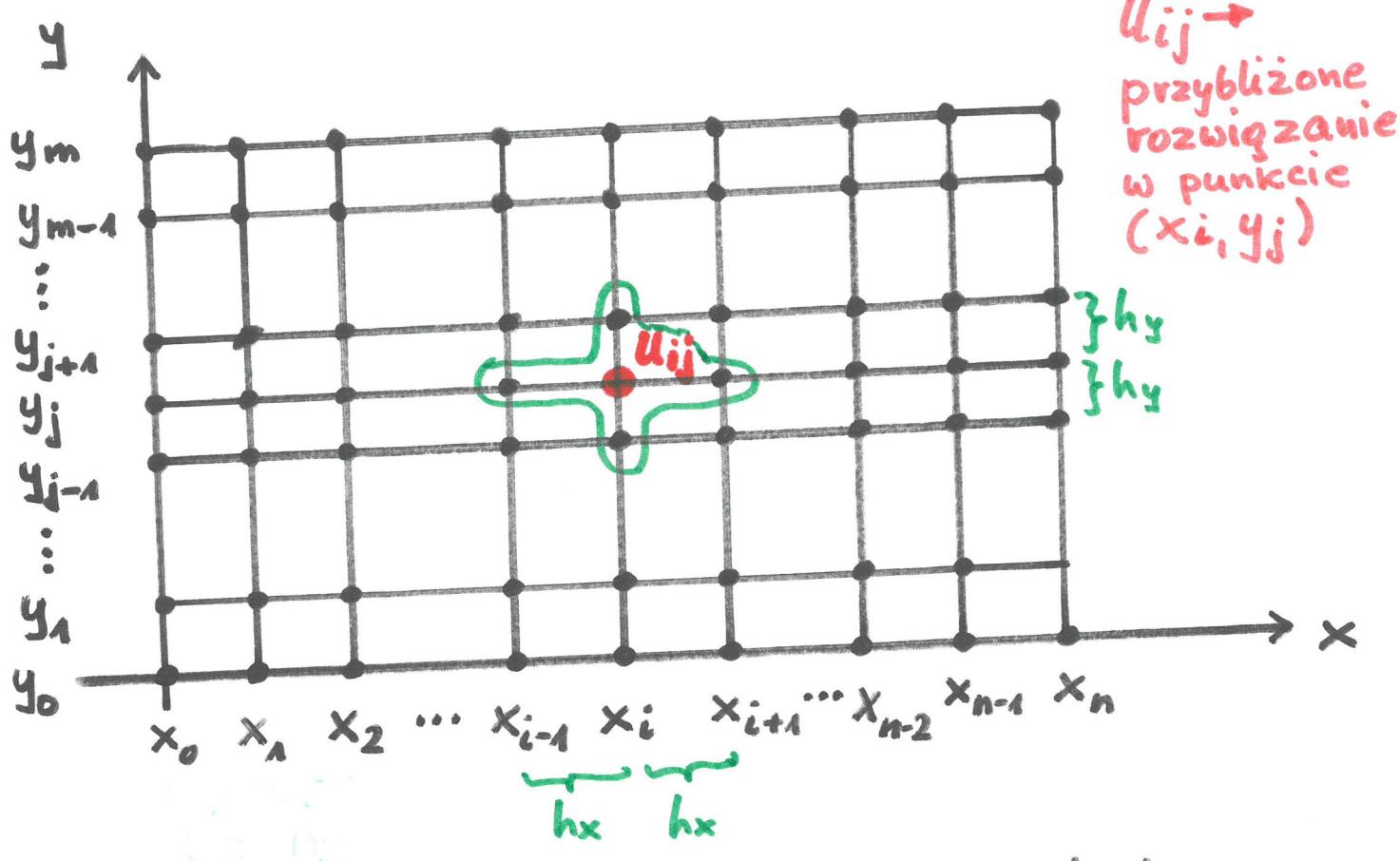
$$\left. \begin{array}{l} u(0,y) = 0 \\ u(a,y) = 0 \\ u(x,0) = 0 \\ u(x,b) = 0 \end{array} \right\} \text{warunki brzegowe}$$



$u(x,y) \rightarrow$ Szukana funkcja

$p(x,y) \rightarrow$ zadana funkcja

siatka przestrzenna dwuwymiarowa



$u_{i,j} \rightarrow$
przybliżone
rozwiązywanie
w punkcie
(x_i, y_j)

Stosujemy przybliżenia trzypunktowe centralne
na drugie pochodne:

$$\left\{ \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h_x^2} + \frac{u_{i,j-1} - 2u_{i,j} + u_{i,j+1}}{h_y^2} + \rho(x_i, y_j) = 0 \right. \\ \left. \text{dla } i = 1, \dots, n-1; j = 1, \dots, m-1 \right.$$

$$u_{i,j} = 0 \text{ dla } i = 0; j = 0, \dots, m \\ i = n; j = 0, \dots, m \\ i = 1, \dots, n-1; j = 0 \\ i = 1, \dots, n-1; j = m$$

Jest to układ liniowych równań algebraicznych, z którego wyliczamy u_{ij}

UWAGA: Należy odróżnić

u_{ij} od $u(x_i, y_i)$

wartość przybliżona
rozwiązańia

wartość ścisła
rozwiązańia

!

jeśli $h_x = h_y = h$ to

$$u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j} = -h^2 g(x_i, y_i)$$

i układ równań algebraicznych przybiera postać

$$A \vec{u} = \vec{w}, \text{ gdzie}$$

$$\vec{u} = \begin{bmatrix} u_{1,1} \\ \vdots \\ \underline{u_{1,m-1}} \\ \vdots \\ u_{2,1} \\ \vdots \\ \underline{u_{2,m-1}} \\ \vdots \\ \vdots \\ \underline{u_{n-1,1}} \\ \vdots \\ u_{n-1,m-1} \end{bmatrix}$$

$$\vec{w} = \begin{bmatrix} -h^2 g(x_1, y_1) \\ \vdots \\ -h^2 g(x_1, y_{m-1}) \\ \hline -h^2 g(x_2, y_1) \\ \vdots \\ -h^2 g(x_2, y_{m-1}) \\ \hline \vdots \\ -h^2 g(x_{n-1}, y_1) \\ \vdots \\ -h^2 g(x_{n-1}, y_{m-1}) \end{bmatrix}$$

Jest to macierz rzadka, BLOKOWO-TRÓJDIAGONALNA.

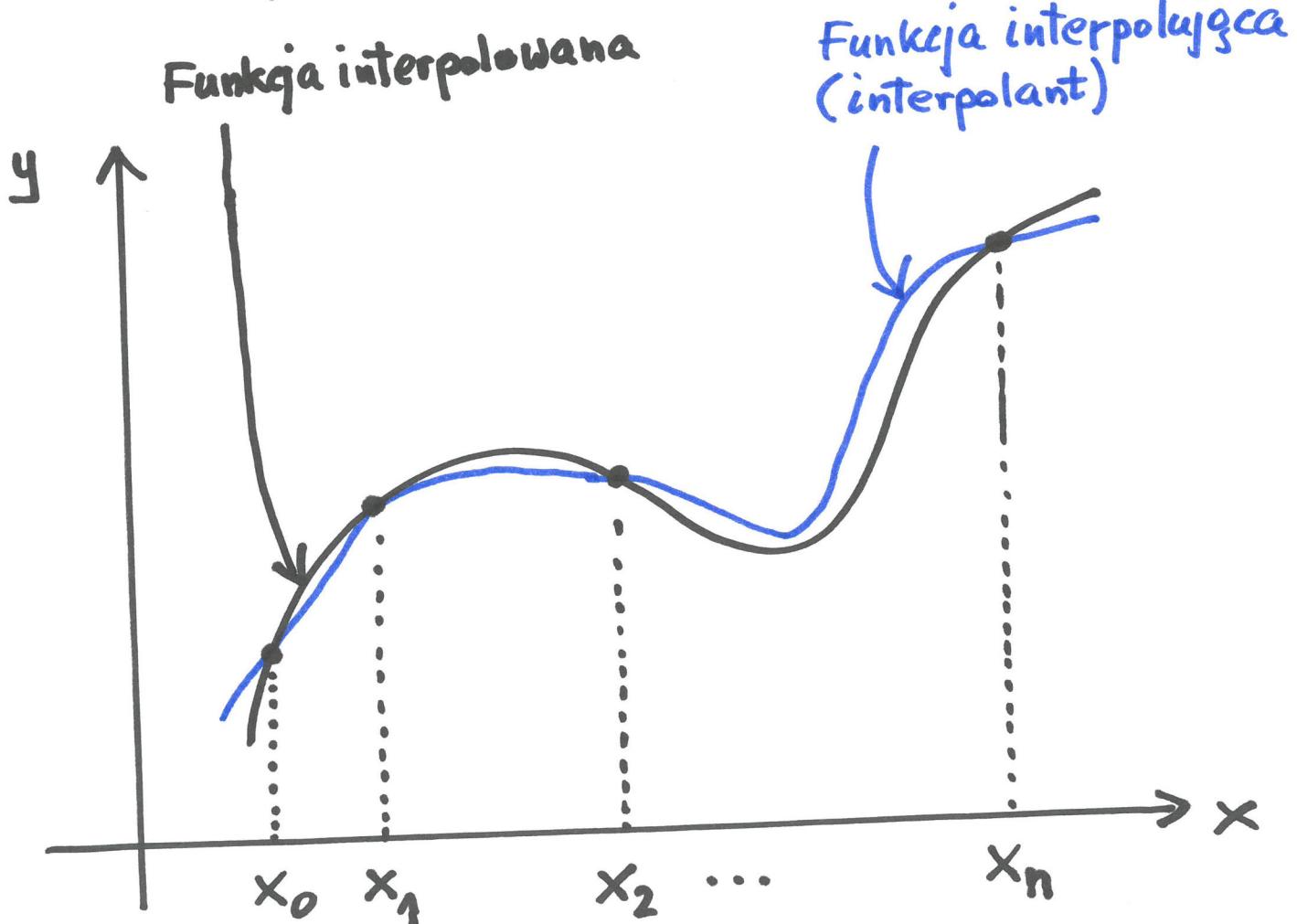
Stosujemy metody iteracyjne (ab bezpośrednie
(np. ogólniony algorytm Thomasa))

Interpolacja
funkcji
jednej zmiennej

Zadanie interpolacji

DANE: Wartości funkcji $f(x)$ i ewentualnie jej pochodnych, w węzłach x_0, x_1, \dots, x_n pewnej siatki

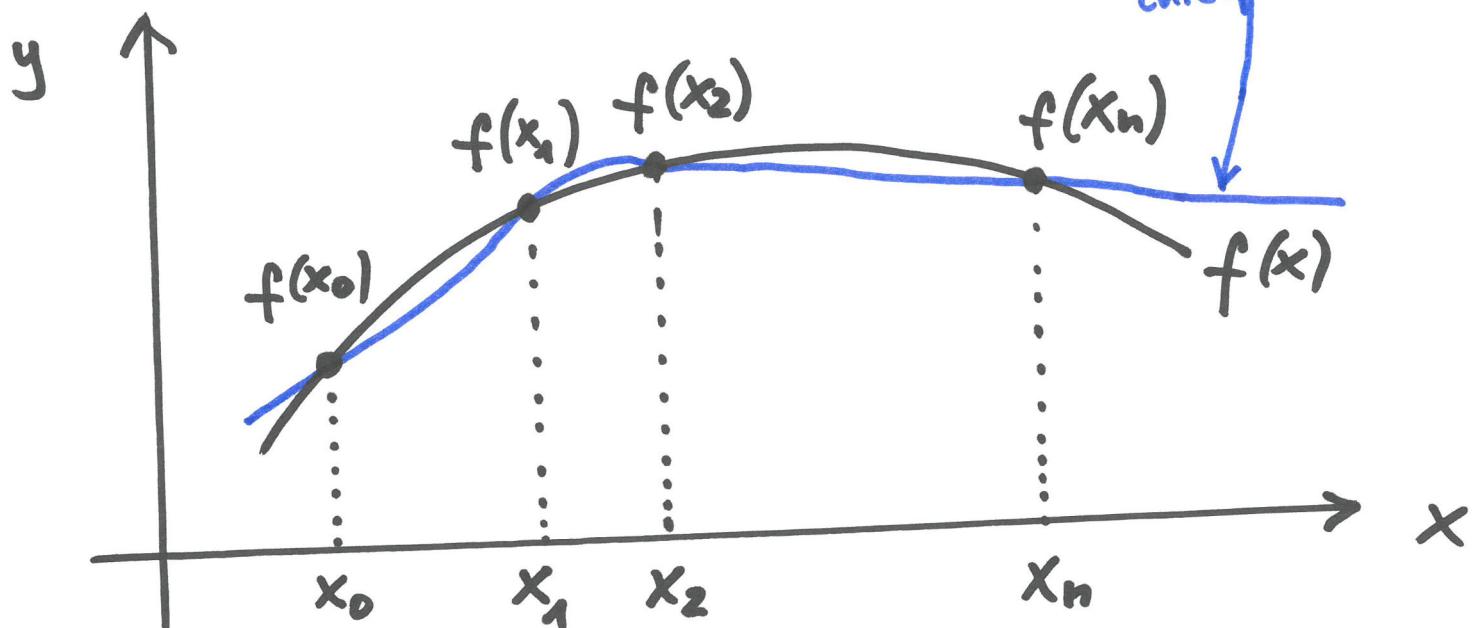
SZUKANE: Inna funkcja, która przyjmuje te same wartości (funkcji i pochodnych) w węzłach, a pomiędzy węzłami stanowi możliwe dobre przybliżenie $f(x)$.



Rodzaje interpolacji

- 1) Wielomianowa
 - Lagrange'a
(tylko wartości $f(x_i)$)
 - Hermite'a
(wartości $f(x_i)$ i pochodnych
 $f'(x_i), f''(x_i), \text{etc.}$)
 - 2) Za pomocą funkcji sklejanych
(splines)
 - 3) Za pomocą funkcji wymiernych
(interpolacja Padé)
 - 4) Za pomocą funkcji wykładniczych
- ⋮ I inne

Interpolacja wielomianowa Lagrange'a



$$\Pi_n = \left\{ P_n(x) = a_n x^n + \dots + a_1 x + a_0 \right\}$$

przestrzeń liniowa wielomianów stopnia co najwyżej n .

Twierdzenie o istnieniu i jednoznaczności zadania interpolacji Lagrange'a:

Dla danej funkcji $f: D \rightarrow \mathbb{R}$ istnieje dokładnie jeden wielomian $P_n(x) \in \Pi_n$ interpolujący $f(x)$ w węzłach x_i ; $i = 0, \dots, n$

Procedura konstrukcji wielomianu interpolującego zależy od wyboru **BAZY** wielomianów w przestrzeni Π_n

Zastosowanie bazy-potęgowej (naturalnej)

$$\left. \begin{array}{l} \varphi_0(x) = 1 \\ \varphi_1(x) = x \\ \varphi_2(x) = x^2 \\ \vdots \\ \varphi_n(x) = x^n \end{array} \right\}$$

Wielomiany bazowe

wielomian interpolujący
(interpolacyjny)

$$p_n(x) = a_0 \varphi_0(x) + a_1 \varphi_1(x) + \dots + a_n \varphi_n(x) = \\ a_0 + a_1 x + \dots + a_n x^n$$

Warunki interpolacji: $p_n(x_0) = f(x_0)$

$$p_n(x_1) = f(x_1)$$

$$\vdots$$

$$p_n(x_n) = f(x_n)$$

w postaci macierzowej:

$$\begin{bmatrix} 1 & x_0 & \dots & x_0^n \\ 1 & x_1 & \dots & x_1^n \\ \vdots & & & \\ 1 & x_n & \dots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$$

↑ Macierz Vandermonde'a

Bywa źle uwarunkowana, poza tym koszt rozwiązywania układu jest duży ($\sim n^3$)

Są to wady użycia bazy potęgowej

Jedyna zaleta bazy potęgowej to wygodne obliczanie wartości $p_n(x)$ za pomocą ALGORYTMU HORNERA:

$$p_n(x) = (((a_n x + a_{n-1}) x + a_{n-2}) x + \dots + a_1) x + a_0$$

(minimalizuje liczbę działań arytmetycznych)

Zastosowanie bazy Lagrange'a

Wielomiany bazowe:

$$L_i(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_n)}{(x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_n)}$$
$$= \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j} \quad i=0, \dots, n$$

Wielomian bazowy $L_i(x)$ zeruje się we wszystkich węzłach za wyjątkiem węzła x_i , w którym przyjmuje wartość 1:

$$L_i(x_j) = \delta_{ij}$$

\nwarrow **DELTA KRONECKERA**

Stąd wynika, że

$$p_n(x) = \sum_{i=0}^n f(x_i) \cdot L_i(x)$$

WIELOMIAN
INTERPOLUJĄCY
(INTERPOLACYJNY)

Konstrukcja $p_n(x)$ jest bardzo łatwa w bazie Lagrange'a

Natomiast obliczanie wartości $P_n(x)$ jest raczej kosztowne. Do minimizacji liczby niezbędnych działań arytmetycznych służy ALGORYTM NEVILLE'A:

$$P_{i(i+1)\dots(i+m)} =$$

$$\frac{(x-x_{i+m}) P_{i(i+1)\dots(i+m-1)} - (x-x_i) P_{(i+1)(i+2)\dots(i+m)}}{x_i - x_{i+m}}$$

x_0	$P_0 = f(x_0)$	P_{01}	P_{012}	P_{0123}	
x_1	$P_1 = f(x_1)$	P_{12}			
x_2	$P_2 = f(x_2)$	P_{23}	P_{123}		
x_3	$P_3 = f(x_3)$				
\vdots	\vdots	\vdots	\vdots	\vdots	

↑
W OSTATNIEJ
KOLUMNIE
DOSTAJEMY
WARTOŚĆ
 $P_n(x)$

W szczególnosci:

$$P_{01} = \frac{(x-x_1)p_0 - (x-x_0)p_1}{x_0 - x_1}$$

$$P_{12} = \frac{(x-x_2)p_1 - (x-x_1)p_2}{x_1 - x_2}$$

$$P_{23} = \frac{(x-x_3)p_2 - (x-x_2)p_3}{x_2 - x_3}$$

$$P_{012} = \frac{(x-x_2)p_{01} - (x-x_0)p_{12}}{x_0 - x_2}$$

$$P_{123} = \frac{(x-x_3)p_{12} - (x-x_1)p_{23}}{x_1 - x_3}$$

$$P_{0123} = \frac{(x-x_3)p_{012} - (x-x_0)p_{123}}{x_0 - x_3}$$

itd.

Przykład konstrukcji wielomianu interpolacyjnego w bazie Lagrange'a

Dane:

x_i	5	-7	-6	0
$f(x_i)$	1	-23	-54	-954

$$l_0(x) = \frac{(x+7)(x+6)(x-0)}{(5+7)(5+6)(5-0)}$$

$$l_1(x) = \frac{(x-5)(x+6)(x-0)}{(-7-5)(-7+6)(-7-0)}$$

$$l_2(x) = \frac{(x-5)(x+7)(x-0)}{(-6-5)(-6+7)(-6-0)}$$

$$l_3(x) = \frac{(x-5)(x+7)(x+6)}{(0-5)(0+7)(0+6)}$$

$$P_3(x) = 1 \cdot l_0(x) - 23 \cdot l_1(x) - 54 \cdot l_2(x) - 954 \cdot l_3(x)$$

Zastosowanie bazy Newtona

$$\varphi_0(x) = 1$$

$$\varphi_1(x) = x - x_0$$

$$\varphi_2(x) = (x - x_0)(x - x_1)$$

⋮

$$\varphi_n(x) = (x - x_0)(x - x_1) \cdots (x - x_{n-1})$$

UWAGA: Tu nie pojawia się x_n !

wielomian interpolacyjny:

$$p_n(x) = c_0 \varphi_0(x) + c_1 \varphi_1(x) + \dots + c_n \varphi_n(x)$$

$$= \sum_{i=0}^n c_i \prod_{j=0}^{i-1} (x - x_j)$$

Można wykazać, że współczynniki c_i da się obliczyć jako odpowiednie ILORAZY RÓŻNICOWE, co pokażemy na przykładzie

WIELOMIANY
BAZOWE

Dane:

x_i	5	-7	-6	0
$f(x_i)$	1	-23	-54	-954

Tworzymy tabelkę

ILORAZY RÓŻNICOWE				
x_i	$f(x_i)$	Rzędu 1	Rzędu 2	Rzędu 3
5	1 c_0	$\frac{-23-1}{-7-5} = 2$ c_1		
-7	-23	$\frac{-54+23}{-6+7} = -31$	$\frac{-31-2}{-6-5} = 3$ c_2	$\frac{-17-3}{0-5} = 4$ c_3
-6	-54	$\frac{-954+54}{0+6} = -150$	$\frac{-150+31}{0+7} = -17$	
0	-954			

Wielomian interpolacyjny:

$$\begin{aligned}
 p_3(x) = & \underline{1} \cdot 1 + \underline{2} \cdot (x-5) + \underline{3} \cdot (x-5)(x+7) \\
 & + \underline{4} \cdot (x-5)(x+7)(x+6)
 \end{aligned}$$

Do obliczenia wartości wielomianu interpolacyjnego w bazie Newtona stosujemy modyfikację ALGORYTMU HORNERA:

$$p_n(x) = [\dots [[c_n(x - x_{n-1}) + c_{n-1}] (x - x_{n-2}) + c_{n-2}] \dots] (x - x_0) + c_0$$

Wtedy interpolacji wielomianowej – Lagrange'a

Tw. Jeżeli $f \in C^{n+1} [a, b]$, a wielomian $P_n \in \mathbb{P}_n$ interpoluje f w węzłach x_0, x_1, \dots, x_n z przedziału $[a, b]$, to

$$\begin{array}{c} \wedge \quad \vee \\ x \in [a, b] \quad \xi_x \in (a, b) \end{array} \quad f(x) - P_n(x) = \\ = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^n (x - x_i)$$

OBSERWACJA :

ponieważ przedział $[a, b]$ jest dowolny, więc iloczyn $\prod_{i=0}^n (x - x_i)$ może być

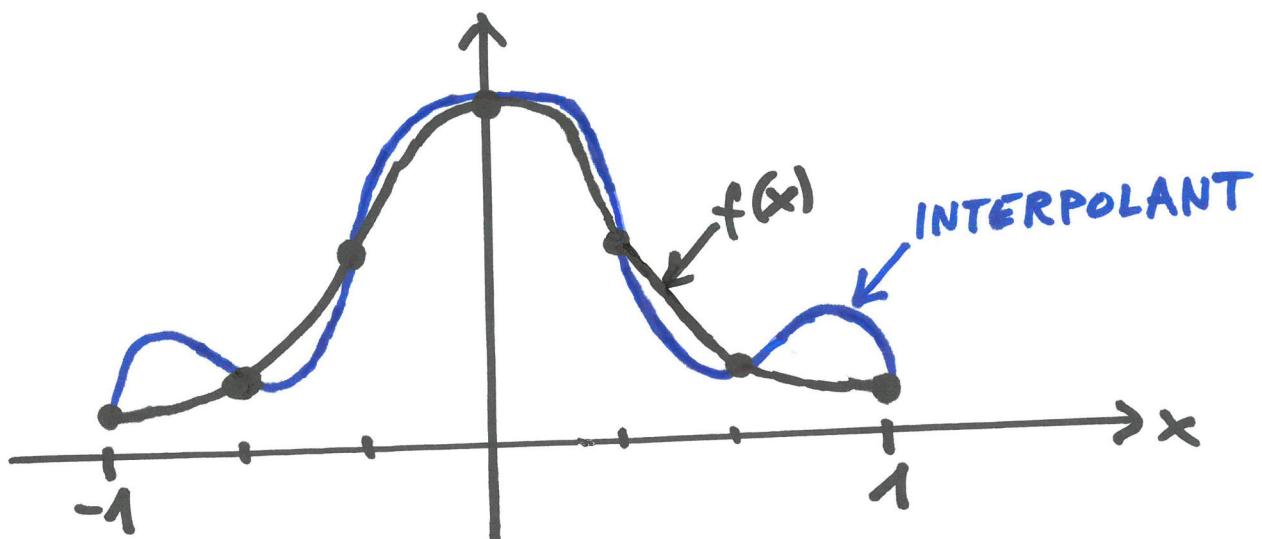
dowolnie duży. Tym samym wtedy może być dowolnie duży!

Zjawisko Rungego (1901 r.)

Interpolacja $f(x) = \frac{1}{1+25x^2}$, $x \in [-1, 1]$
nawęztachn równoodległych

OBSERWACJA:

Ze wzrostem n bieg interpolacji
rośnie, zwłaszcza na końcach przedziału



Wniosek: zaleca się stosowanie $n \leq 5$ lub 6.

Aby zminimalizować błąd interpolacji i zjawisko Rungego, stosujemy też węzły nierównoodległe, zagęszczone na końcach przedziału

WEZŁY CZEBSZEWA:

$$x_i = \frac{b+a}{2} + \frac{b-a}{2} \xi_i$$

$$\xi_i = \cos\left(\frac{2i+1}{2n+2} \pi\right) \quad ; \quad i=0, 1, \dots, n$$

↑
Są to pierwiastki WIELOMIANOW CZEBSZEWA:

$$T_0(x) \stackrel{\text{df}}{=} 1$$

$$T_1(x) \stackrel{\text{df}}{=} x$$

$$T_{k+1}(x) \stackrel{\text{df}}{=} 2xT_k(x) - T_{k-1}(x) \quad ; \quad k \geq 1$$

Interpolacja wielomianowa Hermite'a

Warunki interpolacji w węzle x_i :

$$\left. \begin{array}{l} p_n(x_i) = f(x_i) \\ p_n^{(1)}(x_i) = f^{(1)}(x_i) \\ \vdots \\ p_n^{(k_i-1)}(x_i) = f^{(k_i-1)}(x_i) \end{array} \right\} i = 0, 1, \dots, m$$

$k_i \rightarrow$ "krotność" węzła x_i

Łączna liczba warunków daje

$$k_0 + k_1 + \dots + k_m = n+1$$

↑
Stopień wielomianu
interpolacyjnego

Zastosowanie bazy Newtona

podobnie jak w interpolacji Lagrange'a,
tylko że:

- 1) węzły wielokrotne uwzględniamy wielokrotnie (z kratnościami k_i)
- 2) ilorazy różnicowe rzędu r , oparte na tych samych węzłach zastępujemy przez

$$\frac{f^{(r)}(x_i)}{r!}$$

nie wolno zapominać o $r!$!

Przykład

x_i	$f(x_i)$	$f'(x_i)$	$f''(x_i)$	k_i
1	2	3	-	2
2	6	7	8	3

Tworzymy tabelkę

x_i	$f(x_i)$	ILORAZY RÓŻNICOWE			
		$v_{2\text{gdu}}^1$	$v_{2\text{gdu}}^2$	$v_{2\text{gdu}}^3$	$v_{2\text{gdu}}^4$
1	2	$\frac{3}{1!} = 3$	$\frac{4-3}{2-1} = 1$	$\frac{3-1}{2-1} = 2$	$\frac{1-2}{2-1} = -1$
1	2	$\frac{6-2}{2-1} = 4$	$\frac{7-4}{2-1} = 3$	$\frac{4-3}{2-1} = 1$	
2	6	$\frac{7}{1!} = 7$	$\frac{8}{2!} = 4$		
2	6	$\frac{7}{1!} = 7$			
2	6				

Wielomiany bazowe:

$$\varphi_0(x) = 1$$

$$\varphi_3(x) = (x-1)^2(x-2)$$

$$\varphi_1(x) = x-1$$

$$\varphi_4(x) = (x-1)^2(x-2)^2$$

$$\varphi_2(x) = (x-1)^2$$

wielomian interpolacyjny:

$$P_4(x) = 2 \cdot 1 + 3 \cdot (x-1) + 1 \cdot (x-1)^2 + 2 \cdot (x-1)^2(x-2) - 1 \cdot (x-1)^2(x-2)^2$$

Interpolacja za pomocą funkcji sklejanych

Węzły: x_i ; $i = 0, \dots, n$

Podprzedziały: $[x_i, x_{i+1}]$ o długości

$$h_i = x_{i+1} - x_i$$
$$i = 0, 1, \dots, n-1$$

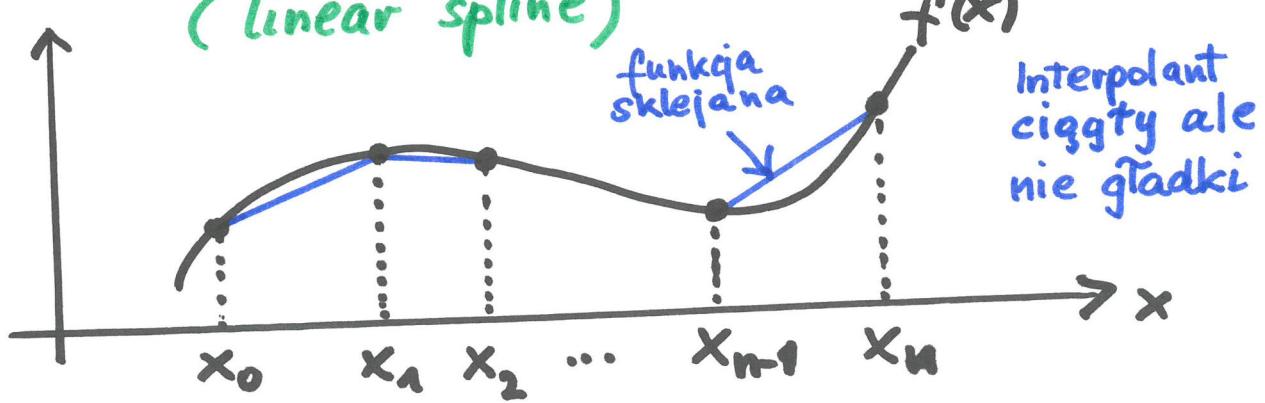
W każdym podprzedziale konstruujemy wielomian możliwie niskiego stopnia i żądamy aby spełnione były

WARUNKI INTERPOLACJI w węzłach

oraz

WARUNKI CIĄGŁOŚCI FUNKCJI INTERPOLUJĄcej i jej pochodnych w węzłach

Przykład: funkcja sklejana 1-go stopnia
(linear spline)



$$S_i(x) = f(x_i) + \frac{f(x_{i+1}) - f(x_i)}{h_i} (x - x_i) \text{ dla } x \in [x_i, x_{i+1}]$$

Interpolacyjna funkcja sklejana 3-go stopnia (cubic spline)

Definicja:

$$S_i(x) \stackrel{\text{df}}{=} a_i + b_i(x - x_i) + c_i(x - x_i)^2 + d_i(x - x_i)^3$$

dla $i = 0, \dots, n-1$

$4n$ niewiadomych współczynników

$$S_i'(x) = b_i + 2c_i(x - x_i) + 3d_i(x - x_i)^2$$

$$S_i''(x) = 2c_i + 6d_i(x - x_i)$$

Warunki interpolacji:

$$S_i(x_i) = f(x_i) \quad ; \quad i = 0, \dots, n-1$$

$$S_{n-1}(x_n) = f(x_n)$$

$n+1$ równan'

Warunki ciągłości:

$$\left. \begin{array}{l} S_i(x_{i+1}) = S_{i+1}(x_{i+1}) \\ S_i'(x_{i+1}) = S_{i+1}'(x_{i+1}) \\ S_i''(x_{i+1}) = S_{i+1}''(x_{i+1}) \end{array} \right\} \begin{array}{l} \text{dla } i = 0, \dots, n-2 \\ 3(n-1) \text{ równan}' \end{array}$$

$$\text{Brakuje } 4n - [(n+1) + 3(n-1)] = 2$$

równan. Potrzebne są dodatkowe dwa
Warunki brzegowe.

Na przykład:

$$\begin{cases} S'_0(x_0) = \alpha \\ S'_{n-1}(x_n) = \beta \end{cases} \quad \begin{matrix} \leftarrow \text{zadane wartości} \\ \leftarrow \end{matrix}$$

albo:

$$\begin{cases} S''_0(x_0) = \gamma \\ S''_{n-1}(x_n) = \delta \end{cases} \quad \begin{matrix} \leftarrow \text{zadane wartości} \\ \leftarrow \end{matrix}$$

$\gamma = \delta = 0 \Rightarrow \text{NATURALNA FUNCJA SKLEJANA}$

albo

$$\begin{cases} S'_0(x_0) = S'_{n-1}(x_n) \\ S''_0(x_0) = S''_{n-1}(x_n) \end{cases}$$

(periodyczne warunki
brzegowe
dla funkcji
okresowych)

Z warunków interpolacji:

$$S_i(x_i) = f(x_i)$$

$$a_i + b_i \underbrace{(x_i - x_i)}_{=0} + c_i \underbrace{(x_i - x_i)}_{=0}^2 + d_i \underbrace{(x_i - x_i)}_{=0}^3 = f(x_i)$$

$$a_i = f(x_i) \quad i = 0, \dots$$

Z warunków ciągłości:

$$\left\{ \begin{array}{l} a_i + b_i \underbrace{(x_{i+1} - x_i)}_{=h_i} + c_i \underbrace{(x_{i+1} - x_i)}_{=0}^2 + d_i \underbrace{(x_{i+1} - x_i)}_{=0}^3 = \\ a_{i+1} + b_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0} + c_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0}^2 \\ + d_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0}^3 \\ b_i + 2c_i \underbrace{(x_{i+1} - x_i)}_{=0} + 3d_i \underbrace{(x_{i+1} - x_i)}_{=0}^2 = \\ b_{i+1} + 2c_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0} + 3d_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0}^2 \\ 2c_i + 6d_i \underbrace{(x_{i+1} - x_i)}_{=0} = 2c_{i+1} + 6d_{i+1} \underbrace{(x_{i+1} - x_{i+1})}_{=0} \end{array} \right.$$

Czyli

$$\begin{cases} a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = a_{i+1} & (*) \\ b_i + 2c_i h_i + 3d_i h_i^2 = b_{i+1} & (**) \\ 2c_i + 6d_i h_i = 2c_{i+1} & (***) \end{cases} \quad \text{dla } i=0, \dots$$

Z równania (***) dostajemy

$$d_i = \frac{c_{i+1} - c_i}{3h_i}; \quad i=0, \dots$$

Wstawiamy d_i do równania (*) :

$$a_i + b_i h_i + c_i h_i^2 + \frac{c_{i+1} - c_i}{3h_i} h_i^3 = a_{i+1}$$

skąd wyliczamy

$$b_i = \frac{a_{i+1} - a_i}{h_i} - \frac{2c_i + c_{i+1}}{3} h_i$$

wskaznik i można zwiększyć o 1, co daje

$$b_{i+1} = \frac{a_{i+2} - a_{i+1}}{h_{i+1}} - \frac{2c_{i+1} + c_{i+2}}{3} h_{i+1}$$

Wstawiamy b_i oraz b_{i+1} do równania (**):

$$\frac{a_{i+1} - a_i}{h_i} - \frac{2(c_i + c_{i+1})}{3} h_i + 2c_i h_i + 3 \frac{c_{i+1} - c_i}{3h_i} h_i^2 =$$

$$\frac{a_{i+2} - a_{i+1}}{h_{i+1}} - \frac{2(c_{i+1} + c_{i+2})}{3} h_{i+1}$$

co jest równoważne równaniu:

$$h_i c_i + 2(h_i + h_{i+1}) c_{i+1} + h_{i+1} c_{i+2} =$$

$$3 \left(\frac{a_{i+2} - a_{i+1}}{h_{i+1}} - \frac{a_{i+1} - a_i}{h_i} \right)$$

dla $i = 0, \dots$

Dla naturalnej funkcji sklejanej
mamy ponadto

$$S_0''(x_0) = 2c_0 + 6d_0 \underbrace{(x_0 - x_0)}_{=0} = 0 \Rightarrow \boxed{c_0 = 0}$$

$$S_{n-1}''(x_n) = 2c_n + 6d_n \underbrace{(x_n - x_n)}_{=0} = 0 \Rightarrow \boxed{c_n = 0}$$

Dostaliśmy układ liniowych równań algebraicznych na współczynniki c_i :

$$A \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{n-1} \\ c_n \end{bmatrix} = \begin{bmatrix} 0 \\ 3\left(\frac{a_2-a_1}{h_1} - \frac{a_1-a_0}{h_0}\right) \\ \vdots \\ 3\left(\frac{a_n-a_{n-1}}{h_{n-1}} - \frac{a_{n-1}-a_{n-2}}{h_{n-2}}\right) \\ 0 \end{bmatrix}$$

gdzie

$$A = \begin{bmatrix} 1 & 0 & & & & \\ h_0 & 2(h_0+h_1) & h_1 & & & \\ & h_1 & 2(h_1+h_2) & h_2 & & \\ & & \ddots & \ddots & \ddots & \\ & & & h_{n-2} & 2(h_{n-2}+h_{n-1}) & h_{n-1} \\ & & & & 0 & 1 \end{bmatrix}$$

Jest to macierz **TRÓJDIAGONALNA**

Stosujemy np. algorytm **Thomasa**

Bląd interpolacji za pomocą funkcji sklejanych 3-go stopnia

Można pokazać, że dla funkcji $f \in C^2_{[x_0, x_n]}$

$$\max_{x \in [x_0, x_n]} |f(x) - S(x)| \leq 5 \cdot M \cdot \max_i (h_i^2)$$

$$\text{gdzie } M = \max_{x \in [x_0, x_n]} |f''(x)|$$

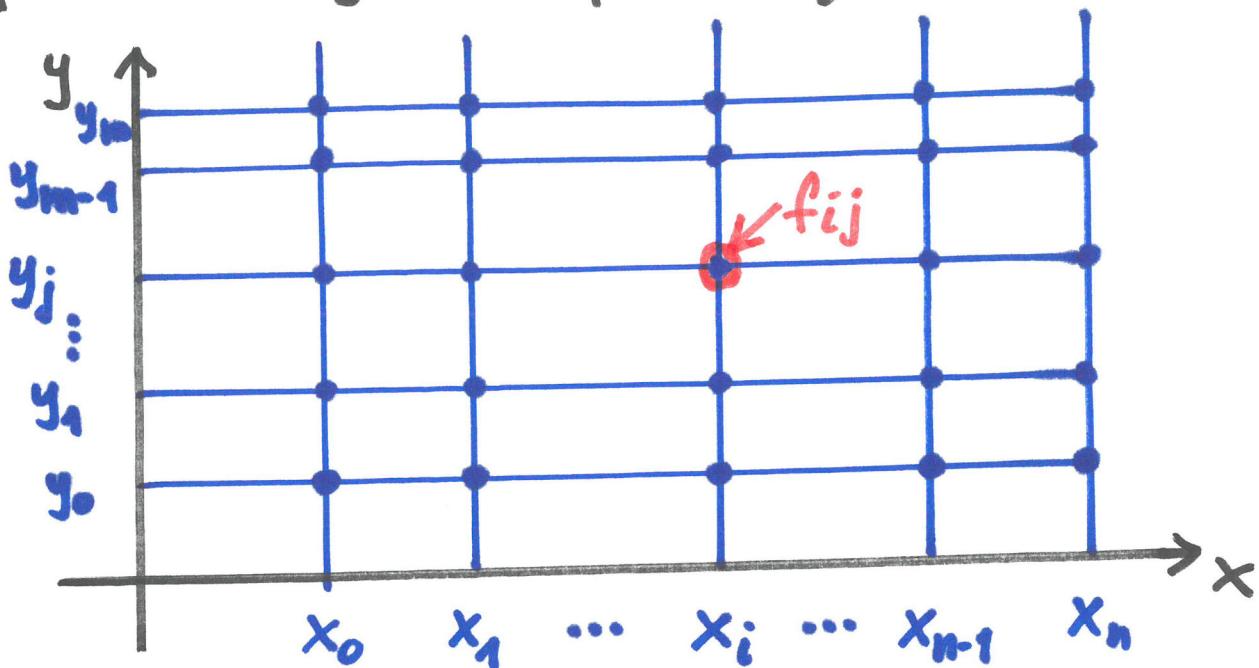
Zatem gdy n rośnie, to h_i maleje
i bląd MAŁEJE !

Co stanowi przewagę nad interpolacją Lagrange'a
bądź Hermite'a.

Przykład
interpolacji
funkcji dwóch zmiennych:
Interpolacja biliniowa

Interpolacja biliniowa w 2D

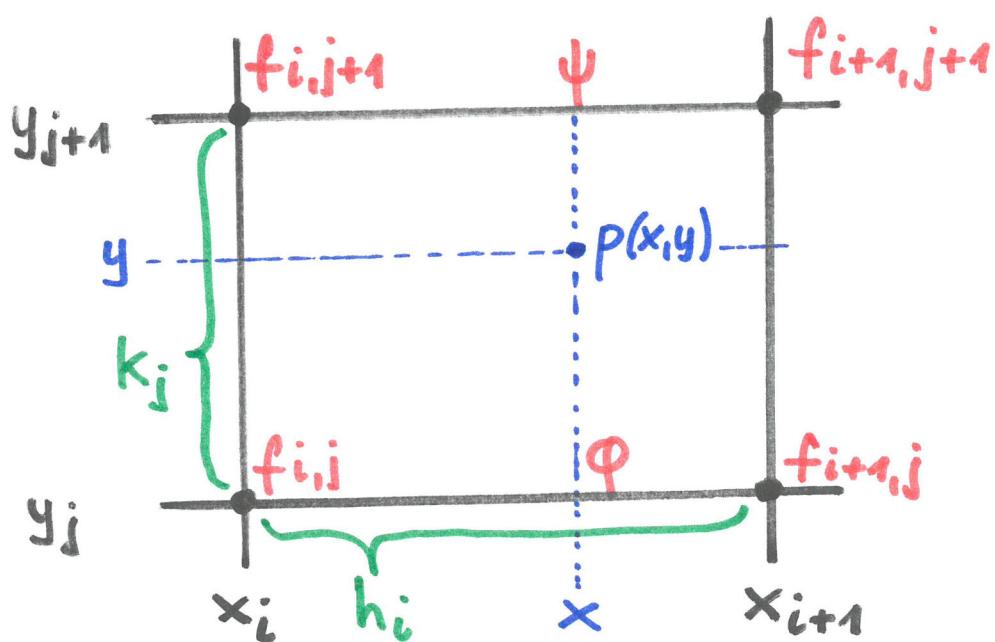
Wprowadzamy siatkę dwuwymiarową:



Tu oznaczamy: $f_{ij} = f(x_i, y_j)$!

$$\begin{array}{l} i = 0, \dots, n \\ j = 0, \dots, m \end{array} \quad \begin{array}{l} h_i = x_{i+1} - x_i \\ k_j = y_{j+1} - y_j \end{array} \quad \left. \begin{array}{l} h_i \\ k_j \end{array} \right\} \text{kroki siatki}$$

Zasada interpolacji:



Najpierw interpolacja liniowa w kierunku x :

$$\varphi = f_{i,j} + \frac{f_{i+1,j} - f_{i,j}}{h_i} (x - x_i) =$$
$$= f_{i,j} \cdot (1 - u_i) + f_{i+1,j} \cdot u_i \quad u_i = \frac{x - x_i}{h_i}$$

$$\psi = f_{i,j+1} + \frac{f_{i+1,j+1} - f_{i,j+1}}{h_i} (x - x_i) =$$
$$= f_{i,j+1} \cdot (1 - u_i) + f_{i+1,j+1} \cdot u_i$$

Potem interpolacja liniowa w kierunku y :

$$p(x,y) = \varphi + \frac{\psi - \varphi}{k_j} (y - y_j) =$$
$$= \varphi (1 - w_j) + \psi w_j \quad w_j = \frac{y - y_j}{k_j}$$

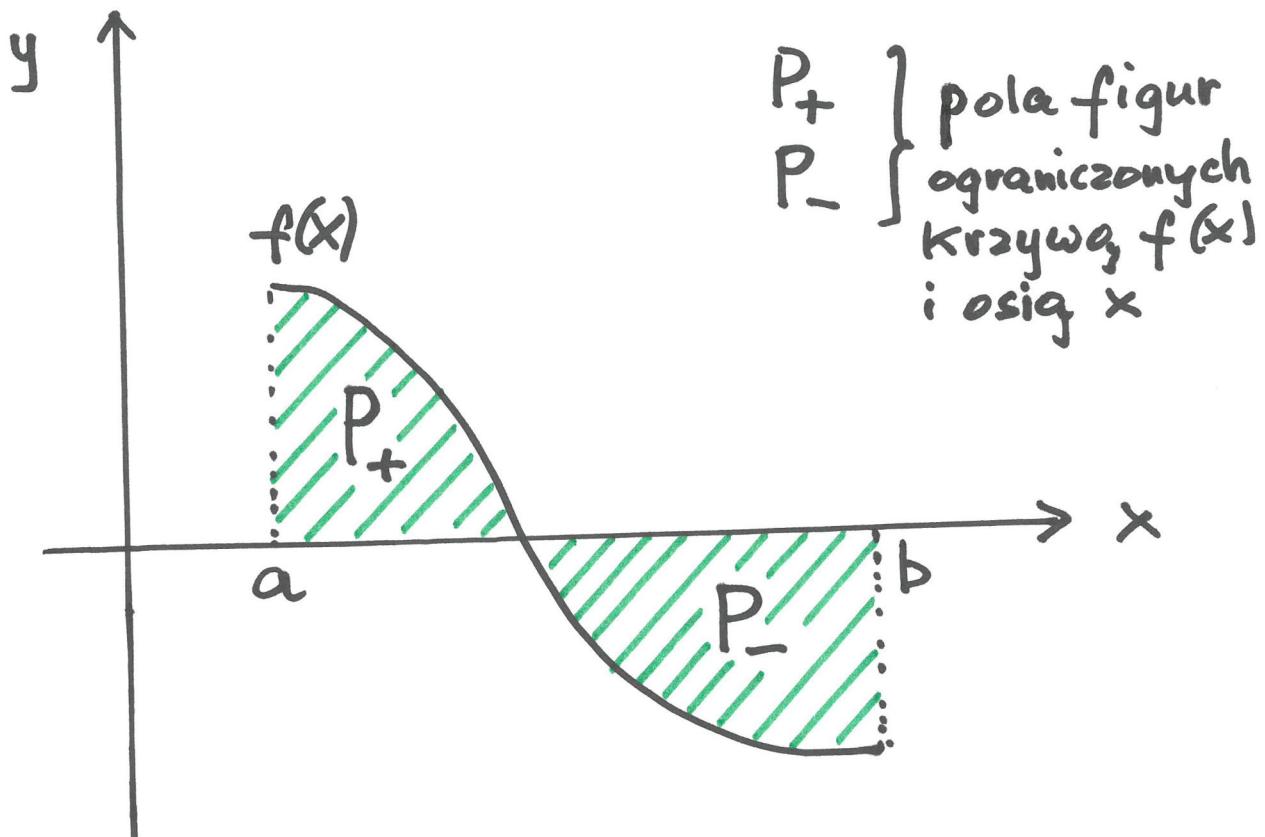
Ostatecznie:

$$p(x,y) = f_{i,j} (1 - u_i) (1 - w_j)$$
$$+ f_{i+1,j} u_i (1 - w_j)$$
$$+ f_{i,j+1} (1 - u_i) w_j$$
$$+ f_{i+1,j+1} u_i w_j$$

UWAGA: $p(x,y)$ nie reprezentuje wycinka otrzymanego!

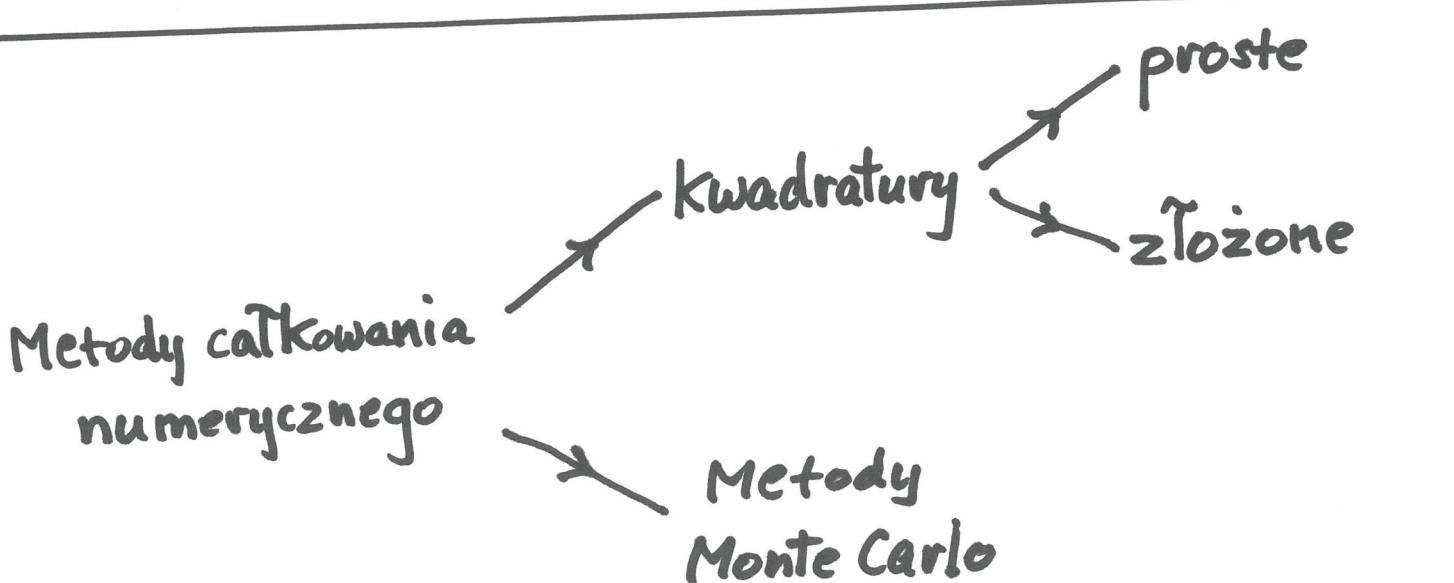
Numeryczne obliczanie
całek oznaczonych
funkcji jednej zmiennej

Zadanie całkowania numerycznego



Chcemy obliczyć

$$I = \int_a^b f(x) dx = P_+ - P_-$$



Kwadratury -

W przedziale $[a, b]$ wybieramy węzły x_0, \dots, x_n w których mamy dane wartości funkcji $f(x_i)$ i ewentualnie jej pochodnych $f^{(1)}(x_i), \dots, f^{(K_i-1)}(x_i)$.

KWADRATURA:

$$Q(f) = \sum_{i=0}^n a_i f(x_i) \quad (\text{lub})$$

$$Q(f) = \sum_{i=0}^n \sum_{j=0}^{K_i-1} a_{ij} f^{(j)}(x_i)$$

$Q(f)$ stanowi przybliżenie całki I

Kwadratury interpolacyjne

Konstruujemy korzystając z wielomianu interpolacyjnego $p(x)$:

$$Q(f) = \int_a^b p(x) dx$$

wielomian
całkujemy
analitycznie

kwadratury interpolacyjne

proste

(stosujemy pojedynczy wielomian w przedziale $[a, b]$)

inne

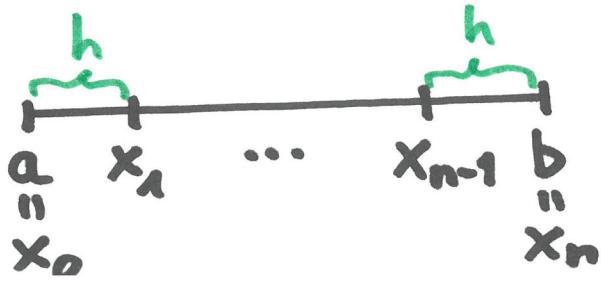
złożone

(Najpierw dzielimy $[a, b]$ na podprzedziały i w nich stosujemy kwadratury proste)

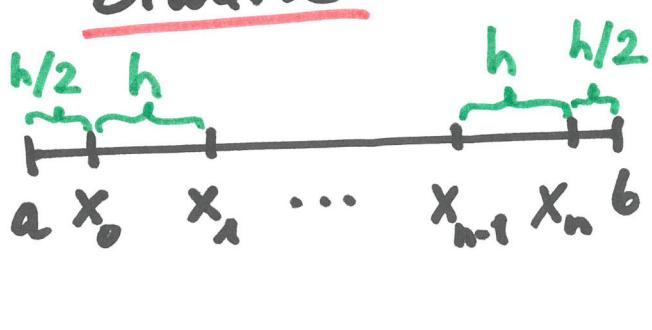
Newtona-Cotesa

(wielomiany interpolacyjne
(Lagrange'a na węzłach)
równoodległych)

zamknięte



otwarte



Blqd kwadratury interpolacyjnej

oceniamy za pomocą wzoru na blqd interpolacji.

$$f(x) = p_n(x) + r_n(x)$$

dla interpolacji Lagrange'a:

$$r_n(x) = \frac{1}{(n+1)!} f^{(n+1)}(\xi_x) \prod_{i=0}^n (x-x_i),$$

$\xi_x \in (a, b)$

stqd

$$I = \int_a^b f(x) dx = \int_a^b p_n(x) dx + \int_a^b r_n(x) dx$$

$\underbrace{\hspace{100pt}}$

$\underbrace{\hspace{100pt}}$

$Q(f)$

$R(f)$

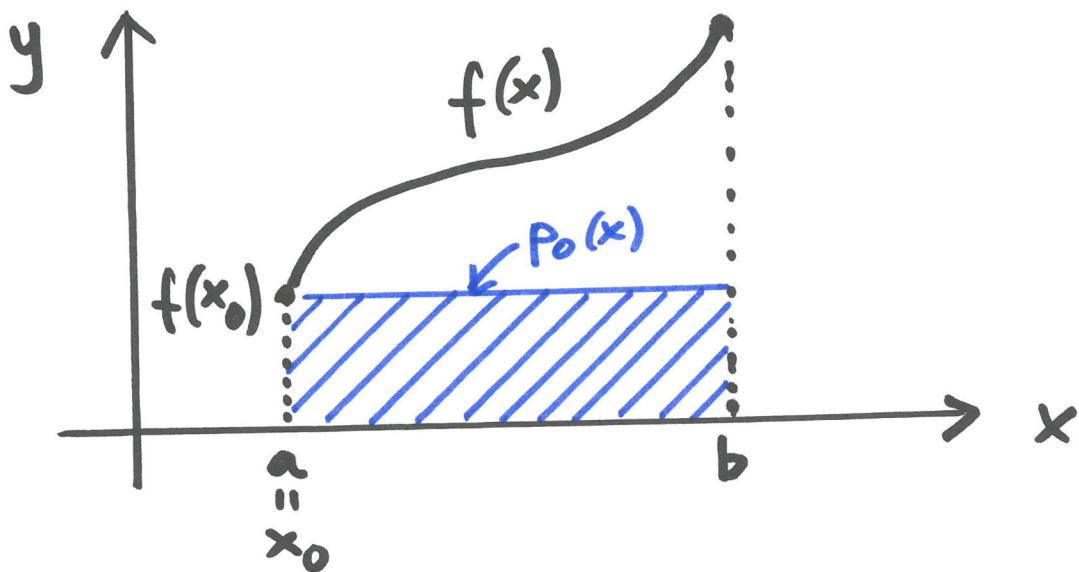
blqd kwadratury

Kwadratury interpolacyjne proste

1 Metoda prostokątów, wariant A

wielomian interpolacyjny 1zgdu 0, z węzłem interpolacji $x_0 = a$

$$p_0(x) = f(x_0) = f(a)$$
$$n_0(x) = f'(\xi_x)(x-x_0), \quad \xi_x \in (a, b)$$



$$Q(f) = \int_a^b f(x_0) dx = f(x_0)(b-a) = f(a)(b-a)$$

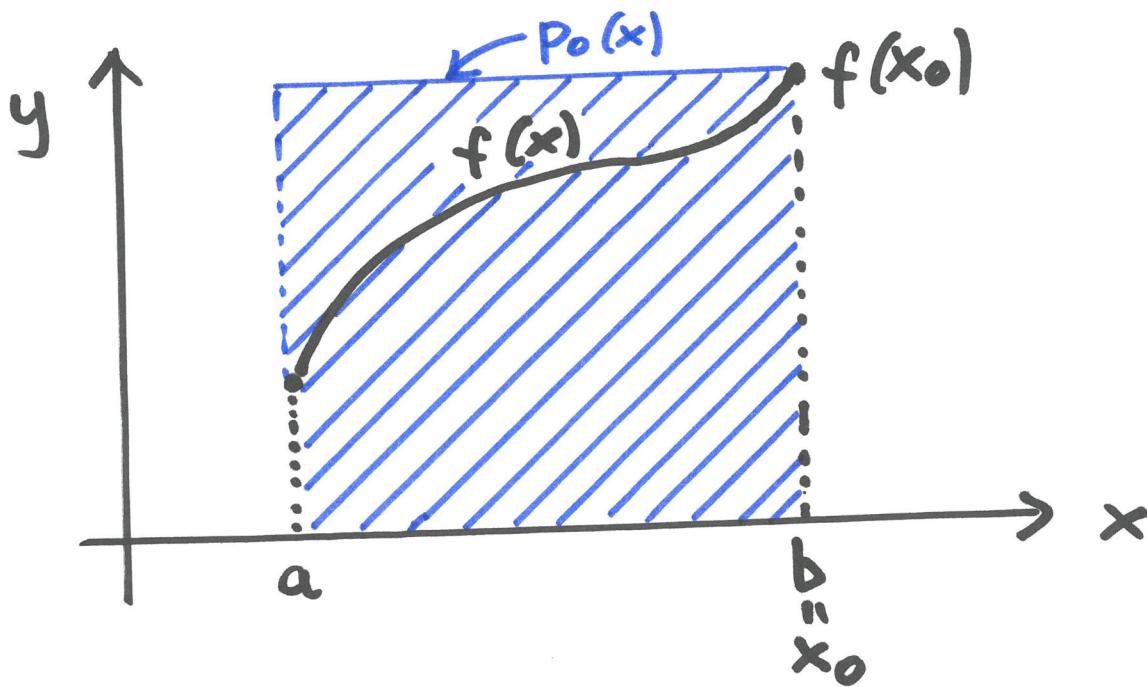
$$R(f) = f'(\xi) \frac{(b-a)^2}{2}, \quad \xi \in (a, b)$$

② Metoda prostokątów, wariant B

wielomian interpolacyjny 1. stopnia P_1 z węzłem interpolacji $x_0 = b$

$$P_0(x) = f(x_0) = f(b)$$

$$r_0(x) = f'(\xi_x)(x - x_0), \quad \xi_x \in (a, b)$$



$$Q(f) = \int_a^b f(x_0) dx = f(x_0)(b-a) = f(b)(b-a)$$

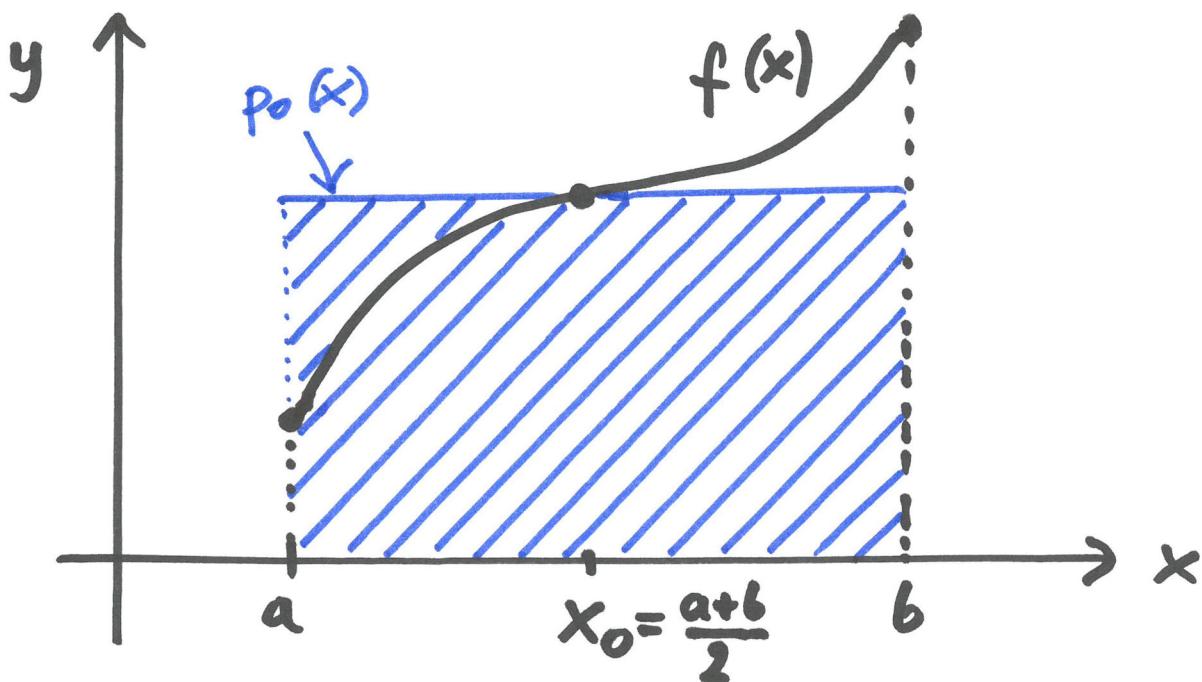
$$R(f) = -f'(\xi) \frac{(b-a)^2}{2}, \quad \xi \in (a, b)$$

3) Metoda prostokątów, wariant C

wielomian interpolacyjny rzędu 0, z węzłem interpolacji $x_0 = \frac{a+b}{2}$

$$p_0(x) = f(x_0) = f\left(\frac{a+b}{2}\right)$$

$$r_0(x) = f'(\xi_x)(x - x_0), \quad \xi_x \in (a, b)$$



$$Q(f) = \int_a^b f(x_0) dx = f(x_0)(b-a) = f\left(\frac{a+b}{2}\right)(b-a)$$

$$R(f) = f''(\xi) \frac{(b-a)^3}{24}, \quad \xi \in (a, b)$$

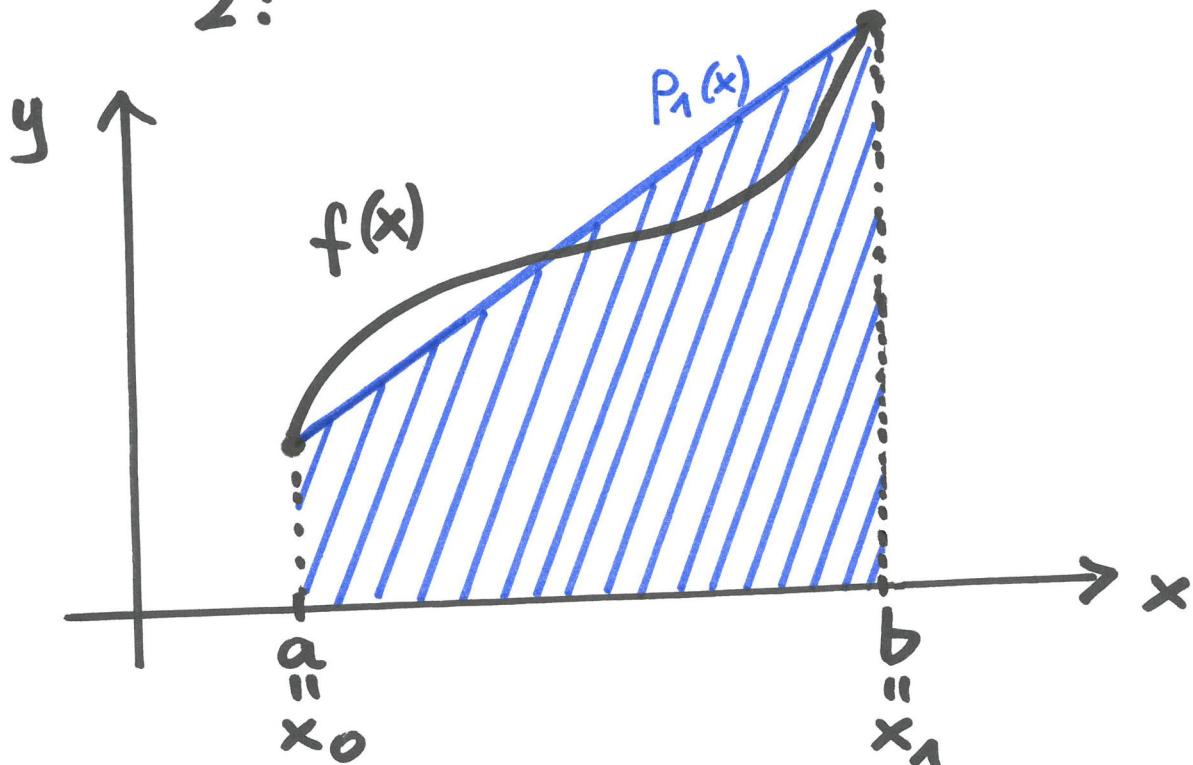
④ Metoda trapezów (zamknięta Newtona-Cotesa)

wielomian interpolacyjny rzędu 1, 2 węzłami

$$x_0 = a, x_1 = b$$

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

$$r_1(x) = \frac{f''(\xi_x)}{2!} (x - x_0)(x - x_1), \quad \xi_x \in (a, b)$$



$$Q(f) = \int_a^b \left[f(x_0) + \frac{f(x_1) - f(x_0)}{x_1 - x_0} (x - x_0) \right] dx = \frac{f(a) + f(b)}{2} (b - a)$$

$$R(f) = -\frac{1}{12} f''(\xi) (b - a)^3, \quad \xi \in (a, b)$$

Bląd większy niż dla metody prostokątów w wariancie C !

5 Metoda parabol (Simpsona) (zamknięta Newtona-Cotesa)

wielomian interpolacyjny rzędu 2, z węzłami

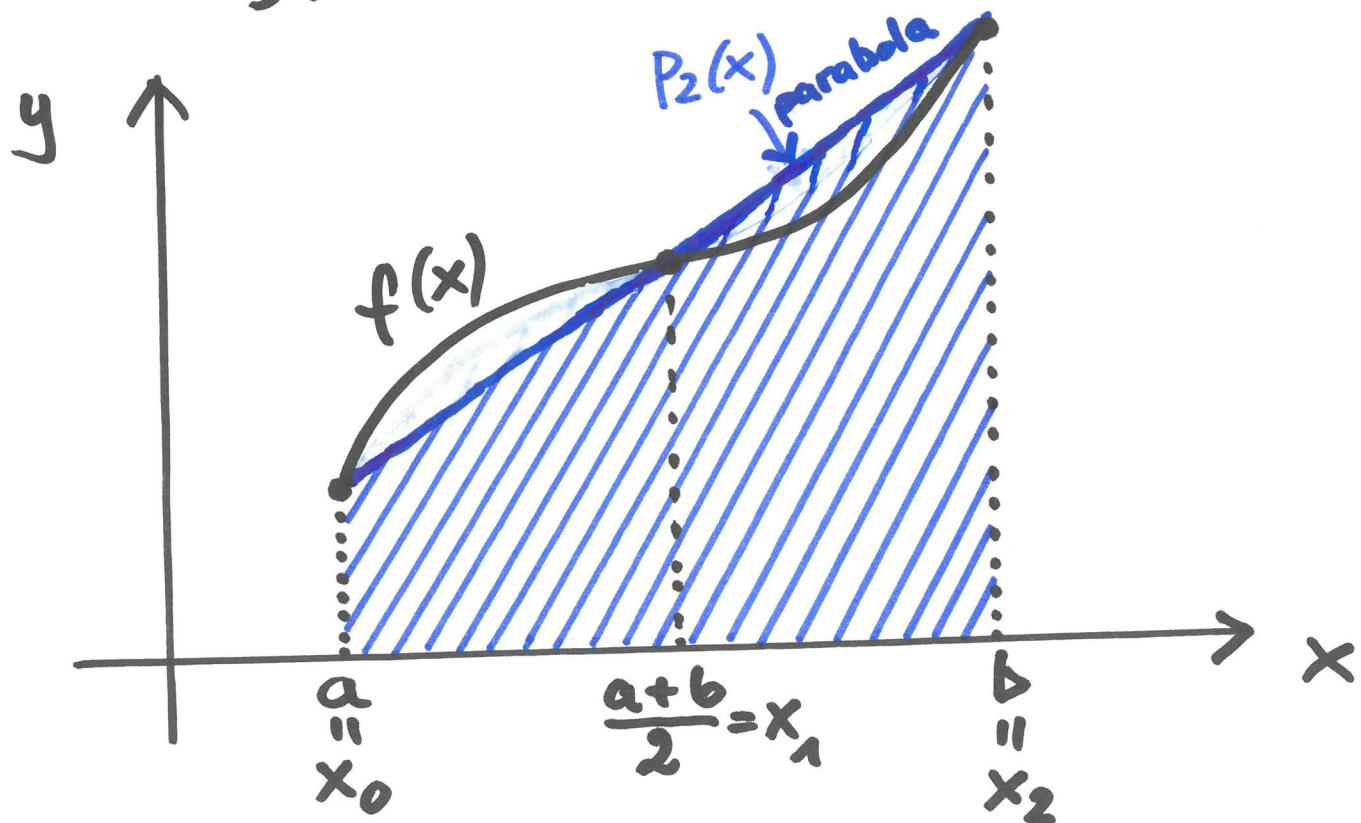
$$x_0 = a, \quad x_1 = \frac{a+b}{2}, \quad x_2 = b$$

$$P_2(x) = f(x_0) \frac{(x-x_1)(x-x_2)}{(x_0-x_1)(x_0-x_2)}$$

$$+ f(x_1) \frac{(x-x_0)(x-x_2)}{(x_1-x_0)(x_1-x_2)}$$

$$+ f(x_2) \frac{(x-x_0)(x-x_1)}{(x_2-x_0)(x_2-x_1)}$$

$$\pi_2(x) = \frac{f'''(\xi_x)}{3!} (x-x_0)(x-x_1)(x-x_2)$$



$$Q(f) = \int_a^b P_2(x) dx =$$

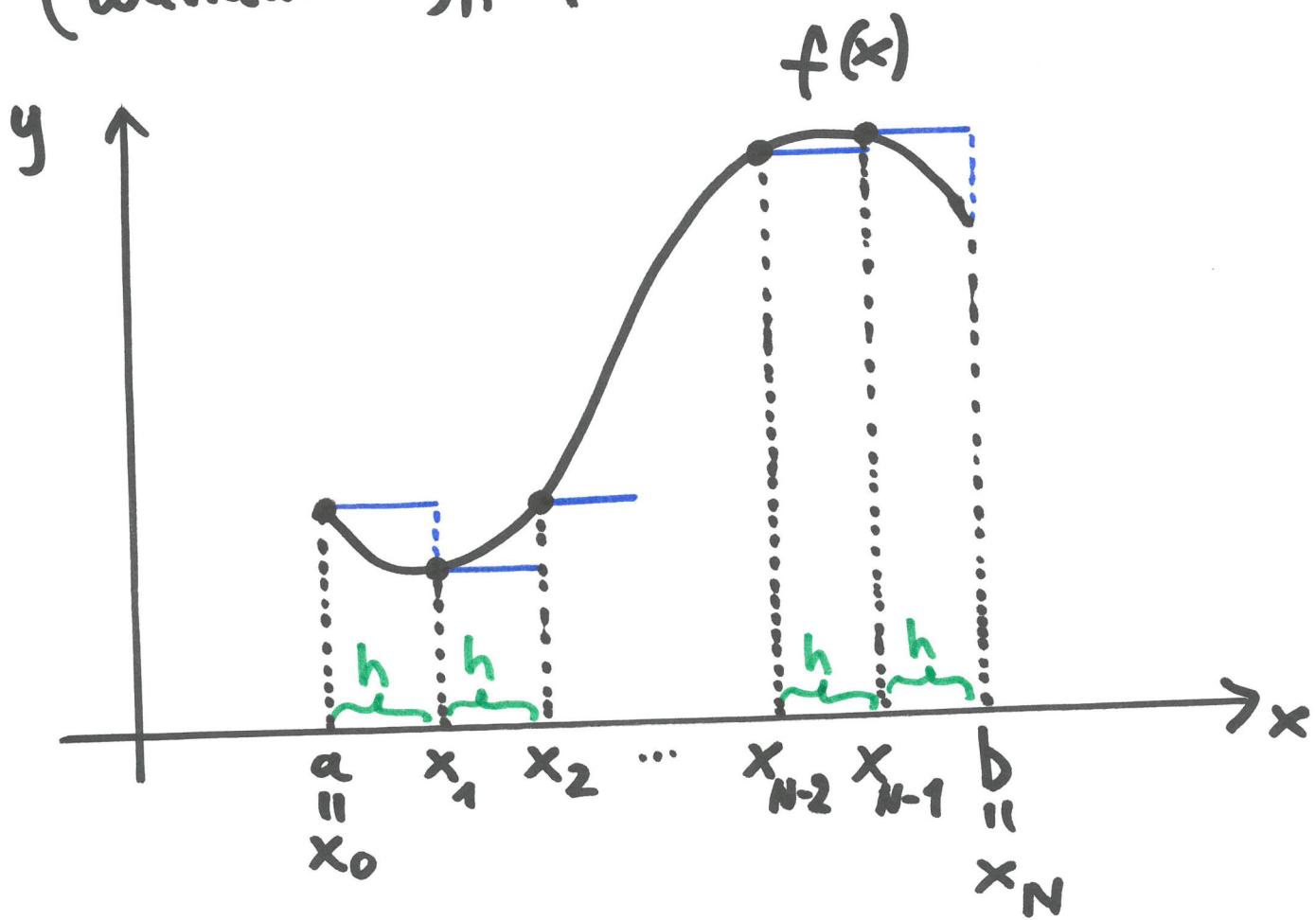
$$\left[\frac{1}{6}f(x_0) + \frac{4}{6}f(x_1) + \frac{1}{6}f(x_2) \right] (b-a)$$

$$R(f) = -\frac{1}{2880} f^{(4)}(\xi) (b-a)^5,$$

$$\xi \in (a, b)$$

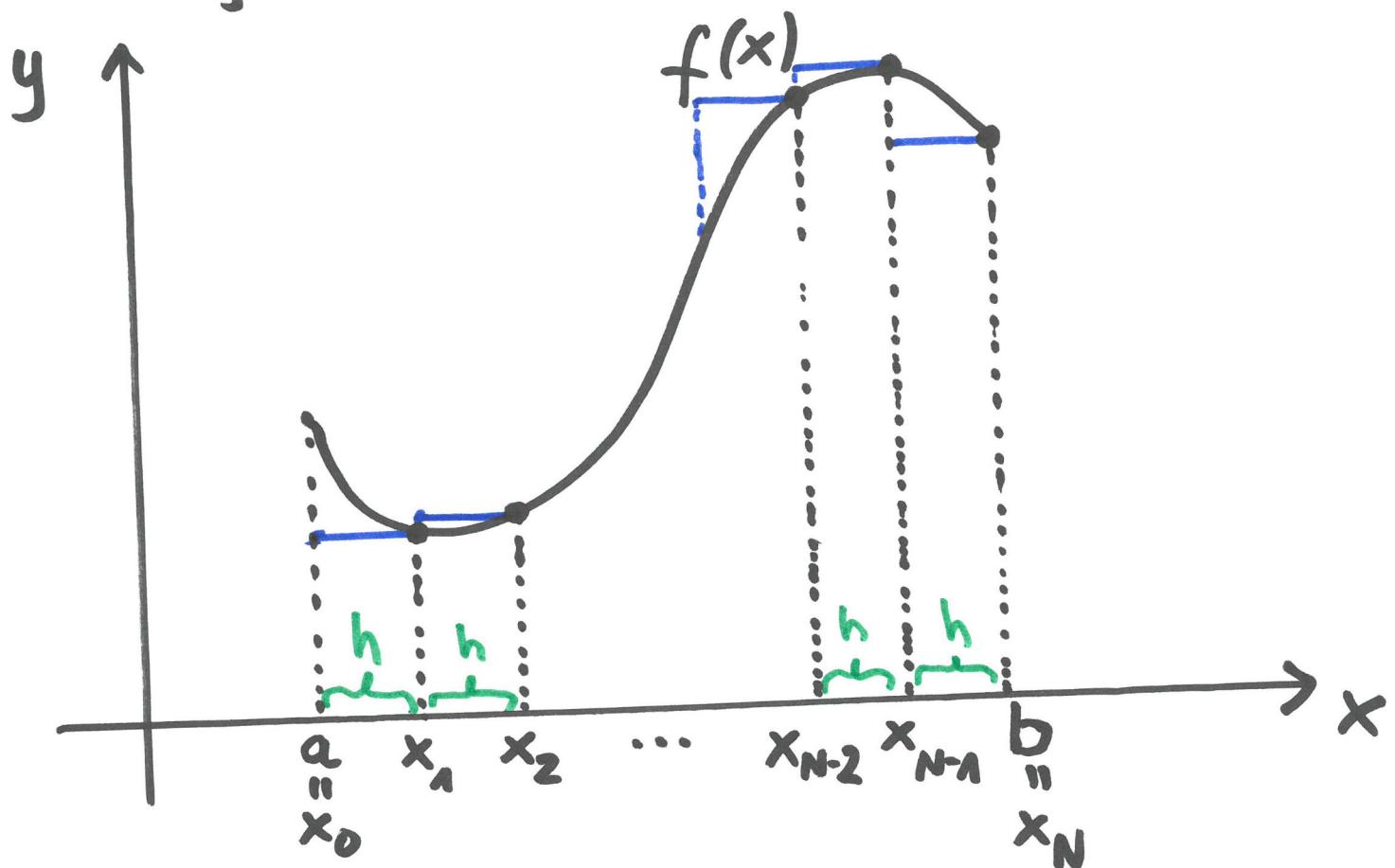
Kwadratury złożone - przykłady -

1 Złożona kwadratura prostokątów (wariant A), podzielony równie długąci h



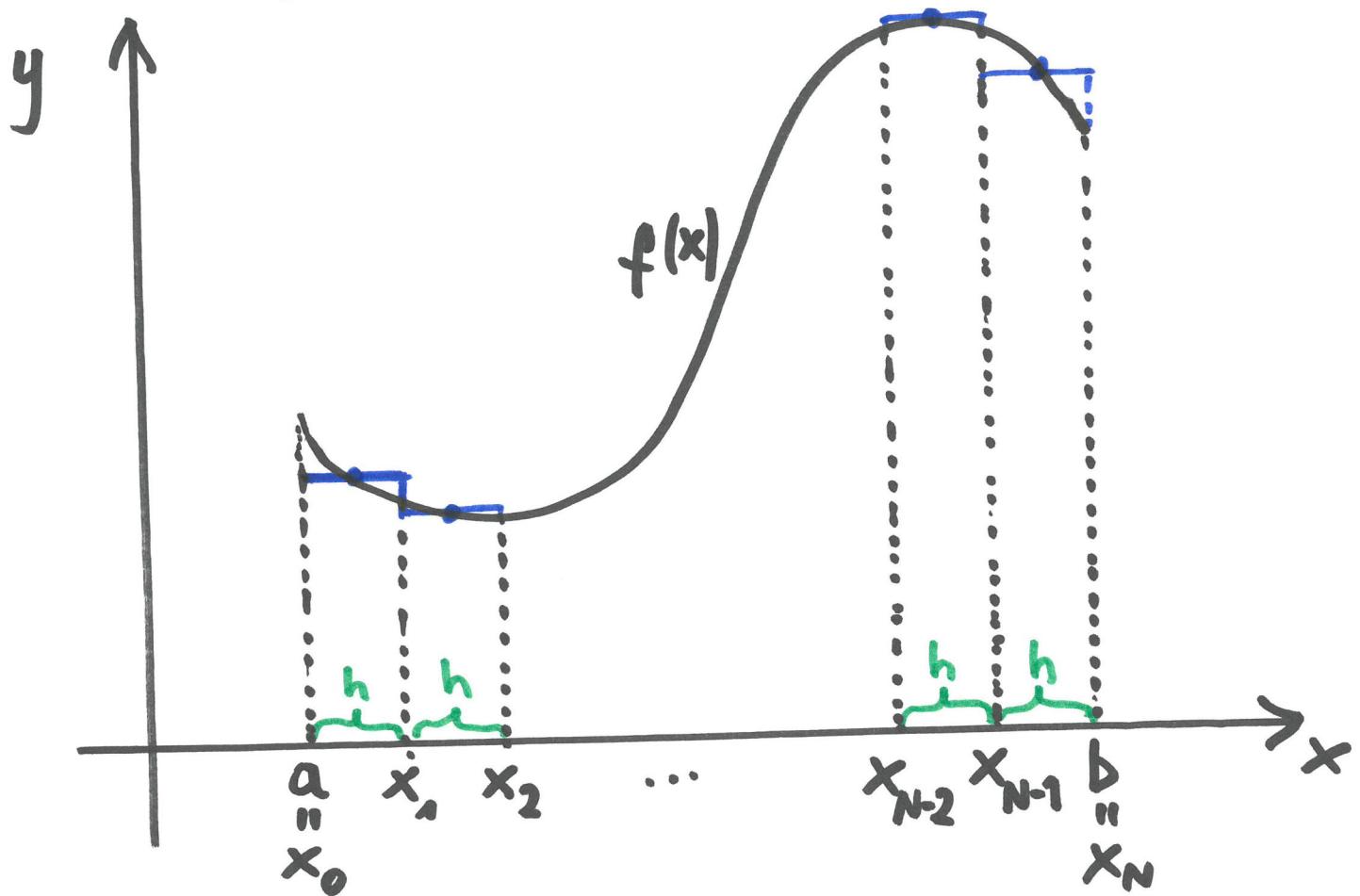
$$Q(f) = \sum_{i=0}^{N-1} f(x_i) h$$

② 2-żona kwadratura prostokątów
 (wariant B), podzielona na równie
 długosci h



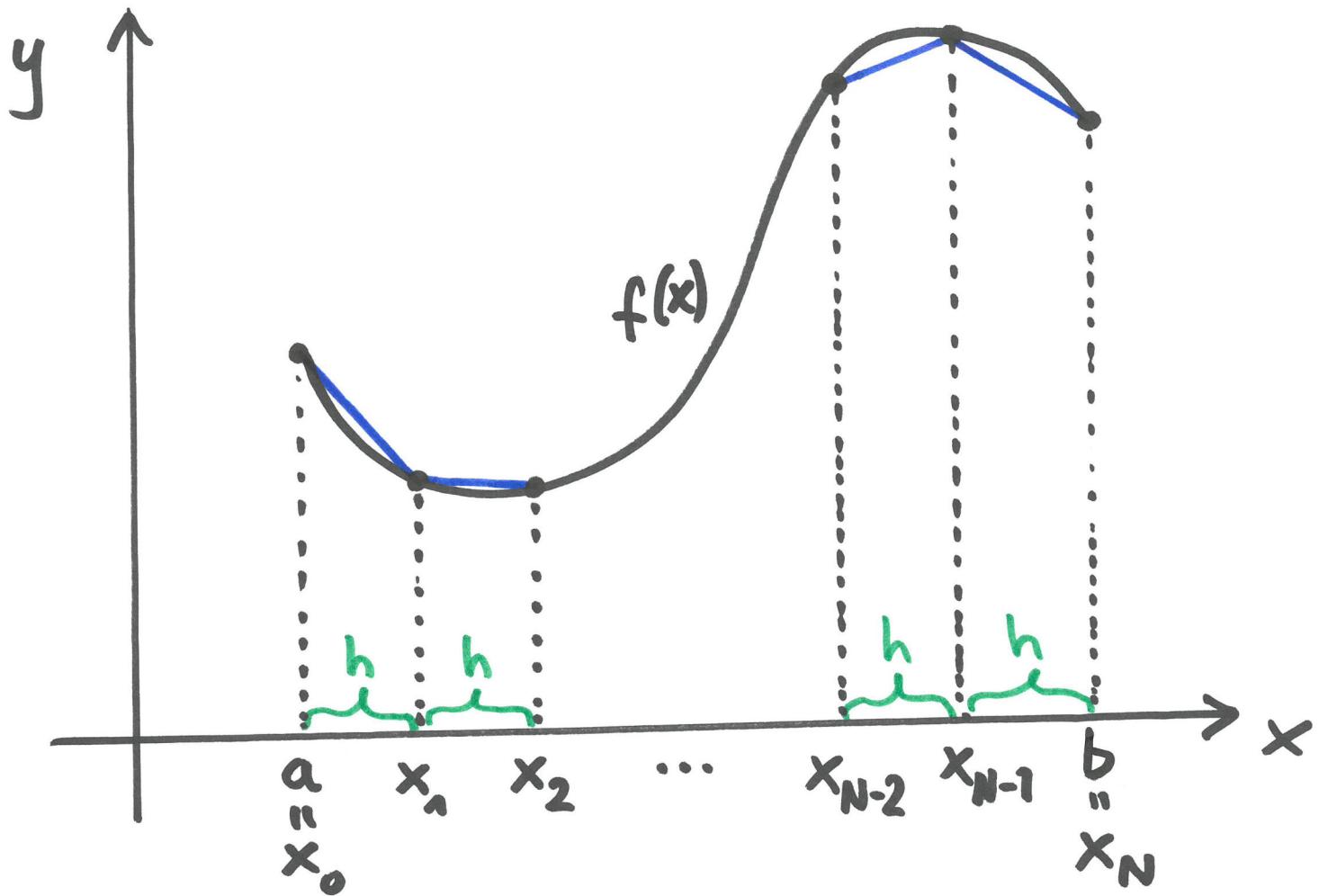
$$Q(f) = \sum_{i=0}^{N-1} f(x_{i+1}) h$$

③ Złożona kwadratura prostokątów (wariant C), podprzedziały o równej długości h



$$Q(f) = \sum_{i=0}^{N-1} f\left(x_i + \frac{h}{2}\right) h$$

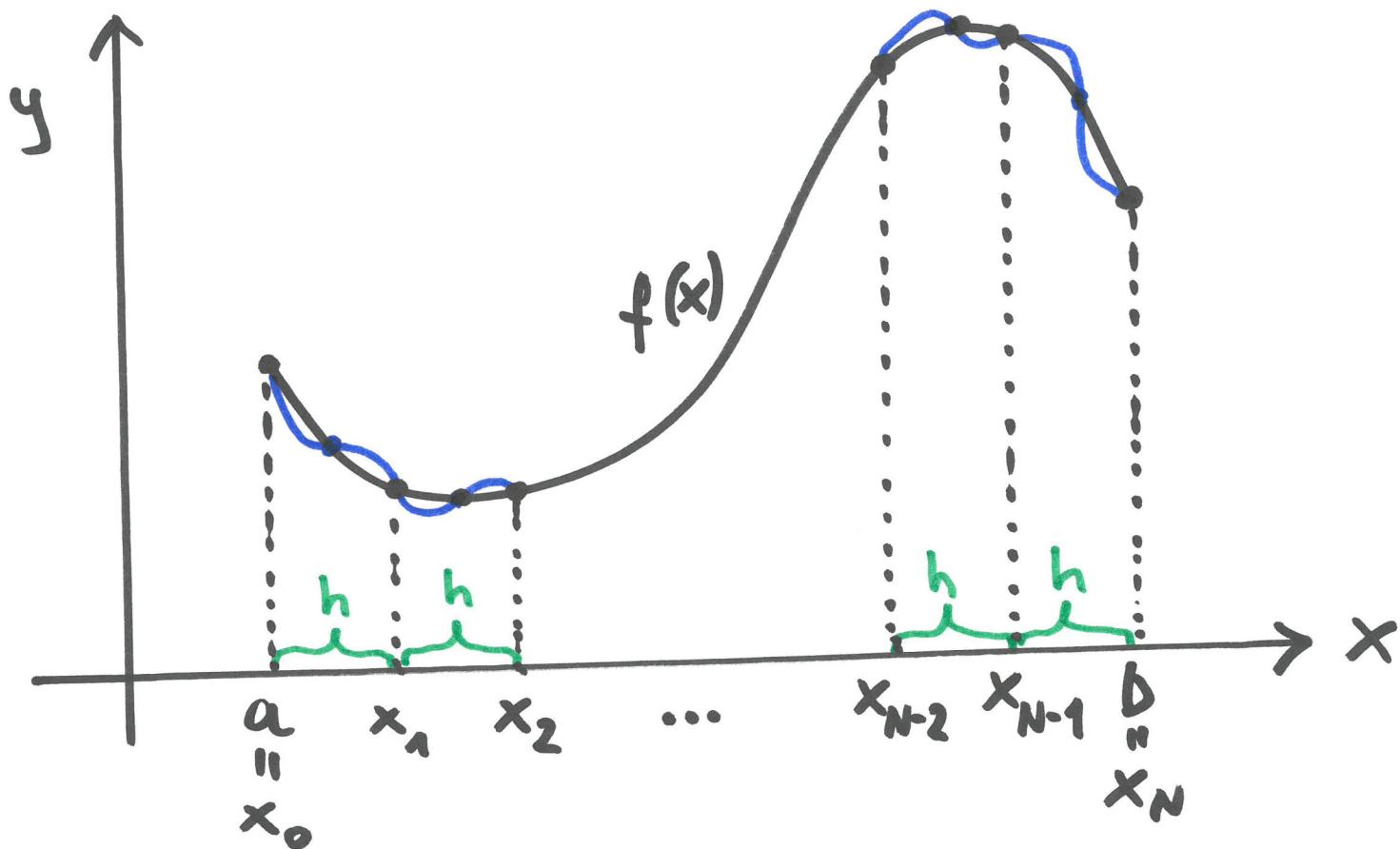
④ Złożona kwadratura trapezów,
podprzedziały o równej długości h



$$Q(f) = \sum_{i=0}^{N-1} \frac{f(x_i) + f(x_{i+1})}{2} h$$

5

Ztożona kwadratura parabol,
podprzedziały o równej dtugosći h



$$Q(f) = \sum_{i=0}^{N-1} \left[\frac{1}{6} f(x_i) + \frac{4}{6} f(x_i + \frac{h}{2}) + \frac{1}{6} f(x_{i+1}) \right] h$$

Kwadratury Gaussa (proste)

Dla kwadratur interpolacyjnych mamy:

$$Q(f) = \int_a^b p_n(x) dx \quad \text{gdzie}$$

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

wielomian
interpolacyjny
Lagrange'a
w bazie
Lagrange'a

Zatem

$$Q(f) = \sum_{i=0}^n A_i f(x_i) \quad \text{gdzie}$$

$$A_i = \int_a^b \prod_{\substack{j=0 \\ j \neq i}}^n \frac{x-x_j}{x_i-x_j} dx$$

Natomiast bieg kwadratury wynosi

$$\int_a^b r_n(x) dx = \frac{1}{(n+1)!} \int_a^b f^{(n+1)}(\xi(x)) \prod_{i=0}^n (x-x_i) dx$$

Wzory te można uogólnić na przypadek
całkowania z wagą, tzn. gdy

$$I = \int_a^b w(x) f(x) dx$$

waga

to

$$Q(f) = \int_a^b w(x) p_n(x) dx = \sum_{i=0}^n A_i f(x_i) , \text{ gdzie}$$

$$A_i = \int_a^b w(x) \prod_{\substack{j=0 \\ j \neq i}}^{n-1} \frac{x-x_j}{x_i-x_j} dx$$

Rzg. kwadratury wynosi

$$\int_a^b w(x) r_n(x) dx = \frac{1}{(n+1)!} \int_a^b w(x) f^{(n+1)}(\xi(x)) \prod_{i=0}^n (x-x_i) dx$$

Definicja: kwadratura jest rzędu m , jeśli jest dokładna dla $f(x) = \text{wielomianowi stopnia} < m$

Wniosek: rzg. powyższych kwadratur wynosi co najmniej $n+1$

(bo $f^{(n+1)}(\xi(x)) = 0$ gdy $f(x)$ jest wielomianem stopnia $< n+1$)

Pytanie: Czy rzg. kwadratury może być $> n+1$?

Odpowiedź: Tak, jeśli odpowiednio rozmiślimy węzły.

Kwadratury Gaussa

to Kwadratury z węzłami dobranymi w taki sposób, aby rzęd kwadratury był jak największy przy zadanej wadze $w(x)$.

Tw. Gaussa:

Jeśli węzły x_0, x_1, \dots, x_n są zerami $(n+1)$ -go wielomianu ortogonalnego $P_{n+1}(x)$ w przedziale (a, b) , z wagą $w(x)$, to kwadratura powyższa jest dokładna dla każdej funkcji $f(x)$ która jest wielomianem stopnia co najwyżej $\underline{2n+1}$!

Uwaga:

Kwadratura Gaussa jest rzędu $2(n+1)$.

Uzupełnienie matematyczne:

iloczyn skalarny wektorów: $x \cdot y = \sum_i x_i y_i$

dwa wektory są ortogonalne jeśli $\boxed{x \cdot y = 0}$

iloczyn skalarny funkcji: $f \cdot g = \int_a^b f(x)g(x) dx$

dwie funkcje są ortogonalne jeśli $\boxed{f \cdot g = 0}$

iloczyn skalarny funkcji z wagą: $f \cdot g = \int_a^b w(x) f(x)g(x) dx$

dwie funkcje są ortogonalne z wagą jeśli $\boxed{f \cdot g = 0}$

Przykład: Dwupunktowy wzór Gaussa
dla całki $\int_{-1}^1 f(x) dx$

Mamy tutaj $w(x) = 1$. W tym przypadku wielomiany ortogonalne w przedziale $(-1, 1)$ to tzw. wielomiany Legendre'a:

$$P_n(x) = \frac{1}{2^n n!} \frac{d^{(n)}}{dx^{(n)}} (x^2 - 1)^n \Rightarrow$$

$$P_0(x) = \frac{1}{2^0 0!} (x^2 - 1)^0 = 1$$

$$P_1(x) = \frac{1}{2 \cdot 1!} \frac{d}{dx} (x^2 - 1) = \frac{1}{2} \cdot 2x = x$$

$$P_2(x) = \frac{1}{2^2 \cdot 2!} \frac{d^2}{dx^2} (x^2 - 1)^2 = \frac{1}{2} (3x^2 - 1)$$

$P_2(x)$ ma pierwiastki $x_1 = -\frac{1}{\sqrt{3}}, x_2 = \frac{1}{\sqrt{3}}$

Stąd

$$A_1 = \int_{-1}^1 \frac{x - x_2}{x_1 - x_2} dx = - \frac{2x_2}{x_1 - x_2} = \frac{-2 \frac{1}{\sqrt{3}}}{-\frac{1}{\sqrt{3}} - \frac{1}{\sqrt{3}}} = 1$$

$$A_2 = \int_{-1}^1 \frac{x - x_1}{x_2 - x_1} dx = \frac{-2x_1}{x_2 - x_1} = \frac{-2 \left(-\frac{1}{\sqrt{3}}\right)}{\frac{1}{\sqrt{3}} + \frac{1}{\sqrt{3}}} = 1$$

A zatem

$$Q(f) = 1 \cdot f\left(-\frac{1}{\sqrt{3}}\right) + 1 \cdot f\left(\frac{1}{\sqrt{3}}\right)$$

Często używane kwadratury Gaussa

Gaussa-Legendre'a: $w(x) = 1, x \in (-1, 1)$

Gaussa-Chebysheva: $w(x) = \frac{1}{\sqrt{1-x^2}}, x \in [-1, 1]$

Gaussa-Laguerre'a: $w(x) = x^\alpha \exp(-x), x \in (0, \infty)$

Gaussa-Hermitela: $w(x) = \exp(-x^2), x \in (-\infty, \infty)$

Gaussa-Jacobiego: $w(x) = (1-x)^\alpha (1+x)^\beta, x \in (-1, 1)$

Węzły i współczynniki kwadratur Gaussa

x_i A_i

Znaleźć można w tabelach.