# Metody numeryczne

## Wykład 1 - Wprowadzenie

Janusz Szwabiński

# Plan wykładu

1. Sprawy administracyjne

2. Warsztat pracy

3. Arytmetyka komputerowa

4. Błędy w obliczeniach numerycznych

# Sprawy administracyjne

- Kontakt:
  https://prac.im.pwr.edu.pl/~szwabin/
- Materiały do kursu:
  https://eportal.pwr.edu.pl/course/view.php?id=1820
- Zasady zaliczenia:
  https://eportal.pwr.edu.pl/course/view.php?id=1820

# Plan kursu

- Układy równań liniowych
- Równania nieliniowe
- Interpolacja i aproksymacja
- Całkowanie numeryczne
- Różniczkowanie numeryczne
- Równania różniczkowe zwyczajne

# Bibliografia


1. Jaan Kiusalaas, *Numerical Methods in Engineering with Python*
2. G. Dahlquist, A. Björk, *Metody numeryczne*
3. pozostałe pozycje jak w karcie przedmiotu

# Warsztat pracy

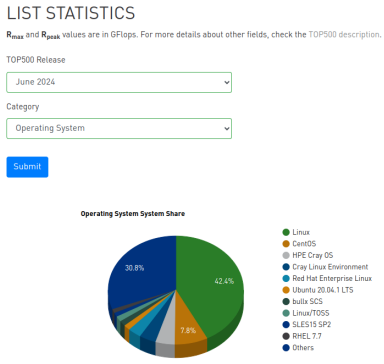
Podstawowe narzędzia:

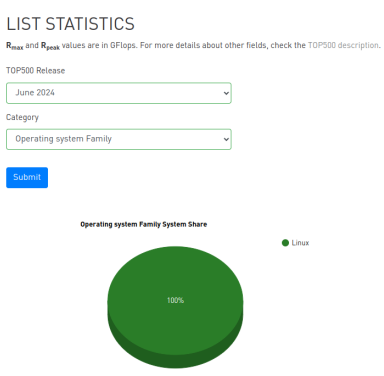
- Python 3.X
- biblioteki numeryczne: numpy, scipy
- inne biblioteki: matplotlib, sympy

Warto znać:

- pygsl
- Maxima, Yacas
- GNU Octave

# Warsztat pracy




https://top500.org/statistics/list/

# Arytmetyka komputerowa (i jej ograniczenia)

Demo

# Reprezentacja stałopozycyjna

Rozważmy liczby w formacie 5-cyfrowym, z dwoma cyframi w części ułamkowej

| | | | , | | |
|---|---|---|---|---|---|

Liczba 256,78 ma w tym formacie naturalną reprezentację

| 2 | 5 | 6 | , | 7 | 8 |
|---|---|---|---|---|---|

# Reprezentacja stałopozycyjna

Najmniejsza liczba

| | | | | | |
|---|---|---|---|---|---|
| 0 | 0 | 0 | , | 0 | 0 |

Największa liczba

| | | | | | |
|---|---|---|---|---|---|
| 9 | 9 | 9 | , | 9 | 9 |

# Reprezentacja stałopozycyjna

Liczba 256,786 będzie miała tylko reprezentację przybliżoną

| 2 | 5 | 6 | , | 7 | 8 |
|---|---|---|---|---|---|

Obcięcie

| 2 | 5 | 6 | , | 7 | 9 |
|---|---|---|---|---|---|

Zaokrąglenie

- w oby przypadkach błąd mniejszy od 0.01
- ogólnie przy zaokrągleniu błąd średnio dwukrotnie mniejszy niż przy obcięciu

# Podstawowe błędy

Błąd bezwzględny:

$$|X - X_o|$$

gdzie $X_o$ to wartość dokładna.

Błąd względny:

$$\frac{|X - X_o|}{X_o}$$

# Podstawowe błędy

## Przykład

Liczby 256,786 i 3,546 mają takie same błędy bezwzględne w naszej reprezentacji z zaokrąglaniem:

$$\begin{aligned} |X^{(1)} - X_0^{(1)}| &= |256.786 - 256.79| = 0,004 \\ |X^{(2)} - X_0^{(2)}| &= |3,546 - 3,55| = 0,004 \end{aligned}$$

# Podstawowe błędy

## Przykład

Błędy względne są większe dla małych liczb:

$$\epsilon_1 = \frac{256.786 - 256.79}{256.786} * 100\% = 0,001558\%$$

$$\epsilon_2 = \frac{3,546 - 3,55}{3,546} * 100\% = 0,11280\%$$

# Arytmetyka zmiennopozycyjna

$$ZNAK \times MANTYSA \times 10^{WYKLADNIK}$$

Co zyskujemy?

- zwiększa się zakres liczb możliwych do przedstawienia
- błędy względne małych i dużych liczb są porównywalne

# Arytmetyka zmiennopozycyjna

## Przykład

Liczba 576329,78 zapisana na 5 miejscach

| 5 | 7 | 6 | 3 | 2 |
|---|---|---|---|---|

$$|X - X_0| = 29,78$$

$$\epsilon_t = 0,0051672\%$$

Liczba 256,78 zapisana na 5 miejscach

| 2 | 5 | 6 | 8 | -1 |
|---|---|---|---|---|

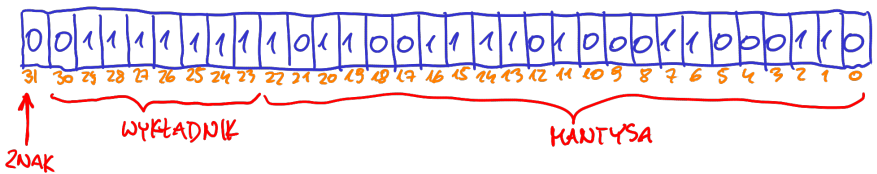
$$|X - X_0| = 0,02$$

$$\epsilon_t = 0,0077888\%$$

# Standard IEEE 754

Dla liczb 32-bitowych mamy:

$$(-1)^{b_{31}} \times 2^{(b_{30}b_{29}\ldots b_{23})_2 - 127} \times (1.b_{22}\ldots b_0)_2$$



# Standard IEEE 754

$$(-1)^{ZNAK} \times 2^{E-127} \times \left(1 + \sum_{i=1}^{23} b_{23-i} 2^i\right)$$

# Jednostka maszynowa

## Definicja

Jednostka maszynowa to najmniejsza liczba $\epsilon$ taka, że spełniony jest dla niej na komputerze warunek

$$1 + \epsilon \neq 1$$

Demo

# Liczby rzeczywiste i maszynowe

Niech $X > 0$ oraz

$$X = q \times 2^m, \quad 1 \leqslant q < 2$$

Wówczas

$$X = (1.a_1 a_2 \ldots)_2 \times 2^m, \quad a_i \in \{0, 1\}$$

# Liczby rzeczywiste i maszynowe

Jeśli mantysa liczb maszynowych ma $t$ bitów po kropce:

$\Rightarrow$ liczba maszynowa bliska $X$ powstanie przez odrzucenie zbędnych bitów $a_{t+1}, a_{t+1}, \ldots$

- obcięcie:
$$X_- = (1.a_1a_2 \ldots a_t)_2 \times 2^m, \quad X_- \leqslant X$$
- zaokrąglenie: odrzucamy zbędne bity i jednoczesnie dodajemy jedynkę na ostatniej pozycji

$$X_+ = \left[(1.a_1a_2 \ldots a_t)_2 + 2^{-t}\right] \times 2^m, \quad X_+ - X_- = 2^{m-t}$$

# Liczby rzeczywiste i maszynowe

## Definicja

Najbliższa względem $X$ liczba maszynowa $fl(X)$ to bliższa z liczb $X_+$ i $X_-$.

Jeśli $fl(X) = X_-$, wówczas

$$|X - fl(X)| \leqslant \frac{1}{2}|X_+ - X_-| = 2^{m-t-1}$$

Podobnie, jeśli $fl(X) = X_+$, to

$$|X - fl(X)| \leqslant \frac{1}{2}|X_+ - X_-| = 2^{m-t-1}$$

# Liczby rzeczywiste i maszynowe

Błąd względny:

$$\frac{|X - fl(X)|}{|X|} \leqslant \frac{2^{m-t-1}}{q2^m} = \frac{1}{q}2^{-t-1} \leqslant 2^{-t-1}$$

Niech

$$\delta = \frac{fl(X) - X}{X}$$

Wówczas

$$fl(X) = X(1 + \delta), \quad |\delta| \leqslant \epsilon$$

# Liczby rzeczywiste i maszynowe

## Przykład

Niech $X = \frac{2}{3}$:

- Jaka jest postać dwójkowa $X$ dla $t = 23$?
- Ile wynoszą liczby $X_-$ i $X_+$?
- Która z tych liczb będzie $fl(X)$?
- Jaki będzie błąd przybliżenia?

# Liczby rzeczywiste i maszynowe

$$\frac{2}{3} = (0.a_1a_2\dots)_2 \quad / \cdot 2$$
$$\frac{4}{3} = (a_1.a_2a_3\dots)_2$$

Część całkowita obu stron jest równa $1 = a_1$. Po jej odjęciu mamy

$$\frac{1}{3} = (0.a_2a_3\dots)_2$$

Powtarzając obustronne mnożenie przez 2 i ewentualne odejmowanie części całkowitej otrzymamy

$$X = \frac{2}{3} = (0.101010\dots)_2 = (1.010101\dots)_2 \times 2^{-1}$$

# Liczby rzeczywiste i maszynowe

Dwie bliskie liczby maszynowe:

$$X_- = (1.\underbrace{010101\ldots010}_{t=23 \text{ bitów}})_2 \times 2^{-1}$$
$$X_+ = (1.\underbrace{010101\ldots011}_{t \text{ bitów}})_2 \times 2^{-1}$$

Obliczmy różnice

$$X - X_- = (0.101010\ldots)_2 \times 2^{-24} = \frac{2}{3} \times 2^{-24}$$
$$X_+ - X = X_+ - X_- - (X - X_-) = 2^{-24} - \frac{2}{3} \times 2^{-24} = \frac{1}{3} \times 2^{-24}$$

# Liczby rzeczywiste i maszynowe

Czyli

$$fl(X) = X_+$$

Błąd bezwzględny

$$|fl(X) - X| = \frac{1}{3} \times 2^{-24}$$

Błąd względny

$$\frac{|fl(X) - X|}{|X|} = \frac{\frac{1}{3} \times 2^{-24}}{\frac{2}{3}} = 2^{-23}$$

# Działania arytmetyczne

Chcemy obliczyć

$$X \odot Y,$$

gdzie $\odot$ oznacza jedno z działań arytmetycznych

Niech:

- $X$ i $Y$ będą liczbami maszynowymi
- mantysa wyniku jest normalizowana i zaokrąglana, tzn. wynikiem działania jest $fl(X \odot Y)$

# Działania arytmetyczne

## Przykład

Komputer działający na liczbach z mantysą pięciocyfrową z przedziału $[0.1, 1)$. Niech:

$$\begin{aligned} X &= 0.31426_{10}\ 3 \\ Y &= 0.92577_{10}\ 5 \end{aligned}$$

# Działania arytmetyczne

Załóżmy, że surowe wyniki umieszczane są w akumulatorze podwójnej długości (typowe w nowoczesnych komputerach):

$$
\begin{aligned}
X + Y &= 0.9289126000_{10}\ 5 \\
X - Y &= -0.9226274000_{10}\ 5 \\
X \times Y &= 0.2909324802_{10}\ 8 \\
X/Y &= 0.3394579647_{10}\ -2
\end{aligned}
$$

# Działania arytmetyczne

Po zaokrągleniu mantys:

$$\left.\begin{array}{rl} X + Y = 0.92891_{10}\ 5 & \delta = 2.8_{10}\ -6 \\ X - Y = -0.92263_{10}\ 5 & \delta = 2.8_{10}\ -6 \\ X \times Y = 0.29093_{10}\ 8 & \delta = 8.5_{10}\ -6 \\ X/Y = 0.33946_{10}\ -2 & \delta = 6.0_{10}\ -6 \end{array}\right\} \delta_{\odot} \text{ mniejsze od } 10^{-5}$$

W dobrze zaprojektowanym komputerze:

$$fl(X \odot Y) = (X \odot Y)(1 + \delta), \quad |\delta| \leqslant \epsilon \quad \forall \odot$$

# Działania arytmetyczne

Jeśli $X$ i $Y$ nie są liczbami maszynowymi:

$$fl\left(fl(X) \odot fl(Y)\right) = \left[X(1+\delta_1) \odot Y(1+\delta_2)\right](1+\delta_3), \quad |\delta_i| \leqslant \epsilon$$

# Błąd względny wyrażeń arytmetycznych

Niech $X$, $Y$ i $Z$ będą liczbami maszynowymi:

$$\begin{aligned} fl(X(Y+Z)) &= [X * fl(Y+Z)]\,(1+\delta_1) = [X * (Y+Z)(1+\delta_2)]\,(1+\delta_1) \\ &= X(Y+Z)(1+\delta_1+\delta_2+\delta_1\delta_2) \simeq X(Y+Z)(1+\delta_1+\delta_2) \\ &= X(Y+Z)(1+\delta_3) \end{aligned}$$

$$|\delta_1| \leqslant \epsilon, \quad |\delta_2| \leqslant \epsilon, \quad |\delta_3| \leqslant 2\epsilon$$

# Błąd względny wyrażeń arytmetycznych

## Twierdzenie

*Jeśli $X_0, X_1, \ldots, X_n$ są dodatnimi liczbami maszynowymi, to błąd względny sumy $\sum_{i=0}^{n} X_i$ jest równy co najwyżej*

$$(1+\epsilon)^n - 1 \simeq n\epsilon.$$

# Utrata cyfr znaczących

## Przykład

Niech

$$X = 0.3721478693, \quad Y = 0.3720230772$$

$$X - Y = 0.0001248121$$

# Utrata cyfr znaczących

## Przykład

W obliczeniach z 5-cyfrowymi mantysami:

$$fl(X) = 0.37215, \quad fl(Y) = 0.37202$$

$$fl(X) - fl(Y) = 0.00013$$

Różnica ma mniej cyfr znaczących w porównaniu z odjemną i odjemnikiem $\Rightarrow$ duży błąd względny!

$$\frac{|X - Y - [fl(X) - fl(Y)]|}{x - y} \simeq 0.04$$

# Utrata cyfr znaczących

## Twierdzenie

*Jeśli liczby maszynowe $X$ i $Y$ są takie, że $X > Y > 0$ oraz*

$$2^{-q} \leqslant 1 - \frac{Y}{X} \leqslant 2^{-p}$$

*($p$ i $q$ są całkowite), to liczba bitów znaczących straconych przy odejmowaniu $X - Y$ jest równa co najmniej $p$ i co najwyżej $q$.*

- utraty cyfr znaczących można uniknąć odpowiednio planując obliczenia!
- w szczególności należy unikać odejmowania bliskich sobie liczb

# Utrata cyfr znaczących

Demo

# Niestabilność numeryczna

## Definicja

Algorytm numeryczny jest niestabilny, jeżeli małe błędy popełnione na jakimś etapie obliczeń rosną w następnych etapach.

# Niestabilność numeryczna

## Przykład

Rozważmy ciąg

$$X_0 = 1, \quad X_1 = \frac{1}{3}, \quad X_{n+1} = \frac{13}{3}X_n - \frac{4}{3}X_{n-1}, \quad n > 1$$

# Niestabilność numeryczna

Jeśli będziemy liczyć wyrazy ciągu w arytmetyce z 24-bitowymi mantysami:

$$
\begin{aligned}
X_0 &= 1.0000000 & \\
X_1 &= 0.3333333 & \text{(7 cyfr znaczących)} \\
X_2 &= 0.1111112 & \text{(6 cyfr znaczących)} \\
X_3 &= 0.0370373 & \text{(5 cyfr znaczących)} \\
X_4 &= 0.0123466 & \text{(4 cyfr znaczących)} \\
X_5 &= 0.0041187 & \text{(3 cyfr znaczących)} \\
X_6 &= 0.0013857 & \text{(2 cyfr znaczących)} \\
X_7 &= 0.0005131 & \text{(1 cyfra znacząca)} \\
X_8 &= 0.0003757 & \text{(brak cyfr znaczących)} \\
&\vdots & \\
X_{15} &= 3.657493 & \text{(błąd względny } 10^8\text{)}
\end{aligned}
$$

# Niestabilność numeryczna

Można pokazać, że rozważany ciąg równoważny jest ciągowi o wyrazach

$$X_n = \left(\frac{1}{3}\right)^n, \quad n \geqslant 0$$

# Niestabilność numeryczna

Demo

# Uwarunkowanie

Uwarunkowanie to wrażliwość rozwiązania zadania na małe zmiany danych początkowych:

$$\begin{aligned} a &\rightarrow a + \delta a \\ W(a) &\rightarrow W(a + \delta a) \end{aligned}$$

Niech $w$ będzie wektorem wyników oraz

$$\delta w = \underbrace{WN(a, \epsilon)}_{\text{wynik numeryczny}} - W(a)$$

# Uwarunkowanie

Wskaźnik uwarunkowania $B(a)$:

$$\frac{\|\delta w\|}{\|w\|} \leqslant B(a) \frac{\|\delta a\|}{\|a\|}$$

# Uwarunkowanie

Jeśli naszym zadaniem jest policzenie funkcji:

$$f(\underbrace{x+h}_{\text{zaburzenie}}) = \underbrace{f'(\xi)h}_{\text{tw. o wart. śr.}} \simeq f'(x)h$$

Wówczas

$$\frac{f(x+h)-f(x)}{f(x)} \simeq \frac{hf'(x)}{f(x)} = \underbrace{\frac{xf'(x)}{f(x)}}_{B(x)} \left(\frac{h}{x}\right)$$

# Źródła błędów w obliczeniach numerycznych

- błędy wejściowe
- błędy obcięcia
- błędy zaokrągleń

# Błędy wejściowe

- dane wejściowe są wynikiem pomiarów
- skończona długość słów binarnych
- wstępne zaokrąglanie liczb niewymiernych

$$\pi = 3.14\ldots$$

Warto wiedzieć

$$\pi = 4 * arctg(1.0)$$

# Błędy obcięcia

Spowodowane użyciem przybliżonej formuły zamiast pełnej operacji matematycznej

- przejścia graniczne (np. pochodne i całki oznaczone)
- sumy nieskończone szeregów

# Błędy zaokrągleń

- skończona długość słów binarnych

# Metody numeryczne

## Wykład 2/3 - Układy równań liniowych

Janusz Szwabiński

# Plan wykładu

1. Układy równań liniowych
2. Pojęcia podstawowe
3. Metody dokładne
4. Metody iteracyjne
5. Układy niedookreślone
6. Układy nadokreślone

# Układy równań liniowych

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

- układ może mieć nieskończenie wiele rozwiązań, jedno rozwiązanie lub nie mieć ich wcale
- warunki istnienia rozwiązań układu są znane
- istnieją też gotowe wzory na wyliczenie $\vec{x}$ w wielu przypadkach
- numeryczne rozwiązanie może się okazać dość trudnym zadaniem

# Układy równań liniowych

- jedno z ważniejszych zagadnień w ramach tego kursu
- wiele problemów fizycznych sprowadza się do rozwiązania układów równań liniowych
- w analizie numerycznej wiele algorytmów opartych jest o takie układy

# Numeryczne metody rozwiązań

**metody dokładne** przy braku błędów zaokrągleń dają dokładne rozwiązanie po skończonej liczbie przekształceń układu wyjściowego

**metody iteracyjne** pozwalają na wyznaczenie zbieżnego ciągu rozwiązań przybliżonych

# Normy

## Definicja

Normą w przestrzeni $\mathbf{R}^n$ nazywamy funkcję

$$\|\cdot\| : \mathbf{R}^n \to \langle 0, +\infty)$$

o następujących własnościach:

1. $\|\vec{x}\| \geqslant 0$ dla każdego $x \in \mathbf{R}^n$,
2. $\|\alpha\vec{x}\| = |\alpha|\|\vec{x}\|$ dla każdego $\alpha \in \mathbf{R}$ i każdego $\vec{x} \in \mathbf{R}^n$,
3. $\|\vec{x}_1 - \vec{x}_2\| \leqslant \|\vec{x}_1\| + \|\vec{x}_2\|$ dla każdej pary $\vec{x}_1, \vec{x}_2 \in \mathbf{R}^n$ (nierówność trójkąta),
4. $\|\vec{x}\| = 0 \Leftrightarrow \vec{x} = 0$.

# Normy wektorowe w $\mathbf{R}^n$

- dla $\vec{x} = [x_1, x_2, \ldots, x_n]^{\mathrm{T}} \in \mathbf{R}^n$ można wprowadzić wiele norm
- najczęściej stosowane w obliczeniach numerycznych:

$$\begin{aligned} \|\vec{x}\|_1 &= |x_1| + |x_2| + \cdots + |x_n| \\ \|\vec{x}\|_2 &= \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \\ \|\vec{x}\|_\infty &= \max\{|x_1|, |x_2|, \ldots, |x_n|\} \end{aligned}$$

- równoważne w tym sensie, że jeśli ciąg wektorów $\vec{x}_1, \vec{x}_2, \vec{x}_3, \ldots$ dąży do wektora zerowego w jednej normie, to zbieżność zachodzi również w dowolnej innej

# Normy macierzowe

## Definicja

Normą macierzy $\mathbf{A}$ nazywamy

$$\|\mathbf{A}\|_{pq} = \max_{\vec{x}\in\mathbf{R}^n, \vec{x}\neq \mathrm{o}} \frac{\|\mathbf{A}\vec{x}\|_q}{\|\vec{x}\|_p}.$$

Przy tym, jeżeli $p = q$, będziemy pisać $\|\mathbf{A}\|_p$.

# Normy macierzowe

$$
\begin{aligned}
\|\mathbf{A}\|_1 &= \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}| \\
\|\mathbf{A}\|_2 &= \sqrt{\lambda_{max}} \\
\|\mathbf{A}\|_\infty &= \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{ij}|
\end{aligned}
$$

$\lambda_{max}$ - największa wartość własna macierzy $\mathbf{A}^{\mathrm{T}}\mathbf{A}$

# Normy macierzowe

## Definicja

Euklidesową normą macierzy (normą Schura, normą Frobeniusza) nazywamy

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

Norma Euklidesowa spełnia warunek zgodności z $\|\cdot\|_2$, tzn.:

$$\|\mathbf{A}\vec{x}\|_2 \leqslant \|\mathbf{A}\|_E \|\vec{x}\|_2.$$

# Wyznaczniki

## Definicja

Wyznacznikiem macierzy kwadratowej **A**,

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{21} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

nazywamy liczbę

$$\det \mathbf{A} = \sum_f (-1)^{l_f} a_{1\alpha_1} a_{2\alpha_2} \ldots a_{n\alpha_n},$$

gdzie $\sum_f$ oznacza sumowanie po wszystkich permutacjach liczb naturalnych $1, 2, \ldots, n$, a $l_f$ to liczba inwersji w permutacji $f$.

# Wyznaczniki

- definicja ma niewielkie znaczenie praktyczne
- możemy próbować policzyć wyznacznik z rozwinięcia Laplace'a wzdłuż $i$-tego wiersza lub $j$-tej kolumny,

$$\det \mathbf{A} = \sum_{j=1}^{n} a_{ij} A_{ij}$$

$$\det \mathbf{A} = \sum_{j=1}^{n} a_{jk} A_{jk}$$

$A_{ij}$ - dopełnienie algebraiczne elementu $a_{ij}$ macierzy $\mathbf{A}$

- rozwinięcie Laplace'a wymaga $n!$ mnożeń
- można je stosować tylko dla bardzo małych $n$

# Macierze trójkątne

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & \cdots & l_{nn} \end{pmatrix}, \quad \mathbf{R} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ 0 & r_{22} & \cdots & r_{21} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & r_{nn} \end{pmatrix}.$$

- sumy, iloczyny i odwrotności macierzy trójkątnych tego samego rodzaju są znowu macierzami trójkątnymi
- łatwo wyliczyć ich wyznacznik

$$\det \mathbf{L} = l_{11} l_{22} \dots l_{nn}, \quad \det \mathbf{R} = r_{11} r_{22} \dots r_{nn}$$

# Układy równań liniowych

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
a_{31}x_1 + a_{32}x_2 + \cdots + a_{3n}x_n &= b_3 \\
a_{41}x_1 + a_{42}x_2 + \cdots + a_{4n}x_n &= b_4 \\
&\vdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}
$$

# Układy równań liniowych - twierdzenie Capellego

Macierz rozszerzona układu

$$\mathbf{D} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} & b_1 \\ a_{21} & a_{22} & \cdots & a_{21} & b_2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} & b_m \end{pmatrix}$$

# Układy równań liniowych - twierdzenie Capellego

## Twierdzenie

*Warunkiem koniecznym i wystarczającym rozwiązywalności dowolnego układu równań liniowych jest, aby rząd r macierzy* **A** *układu był równy rzędowi macierzy rozszerzonej* **D**

- jeśli warunek jest spełniony, układ ma rozwiązanie zależne od $n - r$ parametrów
- dla $n = r$ istnieje jednoznaczne rozwiązanie

# Wzory Cramera

Macierz układu jest nieosobliwa i kwadratowa:

$$x_k = \frac{\det \mathbf{A_k}}{\det \mathbf{A}}, \quad k = 1, 2, \ldots, n$$

- $\mathbf{A}_k$ powstaje z macierzy $\mathbf{A}$ przez zastąpienie $k$–tej kolumny przez wektor $b$
- metoda wymaga bardzo dużego nakładu obliczeń (wyznaczniki)
- może prowadzić do dużych błędów w rozwiązaniu
- nieprzydatna w obliczeniach numerycznych

# Analiza zaburzeń

- obliczenia na komputerach nie są dokładne
- rozwiązanie układu równań obarczone pewnym błędem
- wynik niedokładnego działania w arytmetyce zmiennopozycyjnej możemy przedstawić jako wynik działania nieobarczonego błędami wykonanego na zaburzonych argumentach (interpretacja Wilkinsona)

# Analiza zaburzeń

- zastępujemy macierz **A** macierzą zaburzoną $\mathbf{A} + \delta\mathbf{A}$
- podobnie, $\vec{b} \to \vec{b} + \delta\vec{b}$
- zamiast rozwiązania $\vec{x}$ układu $\mathbf{A}\vec{x} = \vec{b}$ szukamy rozwiązania $\vec{x} + \delta\vec{x}$ układu

$$(\mathbf{A} + \delta\mathbf{A})(\vec{x} + \delta\vec{x}) = \vec{b} + \delta\vec{b}$$

- błąd $\delta\vec{x}$ zależeć będzie od zaburzeń danych wejściowych $\delta\mathbf{A}$ i $\delta\vec{b}$ oraz od uwarunkowania układu

## $\delta\mathbf{A} = \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

Z równania

$$(\mathbf{A} + \delta\mathbf{A})(\vec{x} + \delta\vec{x}) = \vec{b} + \delta\vec{b}$$

otrzymamy

$$\begin{aligned} \mathbf{A}(\vec{x} + \delta\vec{x}) &= \vec{b} + \delta\vec{b} \\ \mathbf{A}\vec{x} + \mathbf{A}\delta\vec{x} &= \vec{b} + \delta\vec{b} \\ \mathbf{A}\delta\vec{x} &= \delta\vec{b} \\ \delta\vec{x} &= \mathbf{A}^{-1}\delta\vec{b} \end{aligned}$$

## $\delta\mathbf{A} = \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

Dla dowolnych norm norm wektorów $\delta\vec{b}$ i $\delta\vec{x}$ oraz indukowanej przez nie normy macierzy $\mathbf{A}^{-1}$ mamy

$$\|\delta\vec{x}\|_p \leqslant \|\mathbf{A}^{-1}\|_{qp}\|\delta\vec{b}\|_q.$$

Jeśli $\vec{x} \neq \mathrm{o}$, to

$$\begin{aligned}\frac{\|\delta\vec{x}\|_p}{\|\vec{x}\|_p} &\leqslant \frac{\|\mathbf{A}^{-1}\|_{qp}}{\|\vec{x}\|_p}\|\delta\vec{b}\|_q = \frac{\|\mathbf{A}^{-1}\|_{qp}\|\vec{b}\|_q}{\|\vec{x}\|_p}\frac{\|\delta\vec{b}\|_q}{\|\vec{b}\|_q} \\ &= \frac{\|\mathbf{A}^{-1}\|_{qp}\|\mathbf{A}\vec{x}\|_q}{\|\vec{x}\|_p}\frac{\|\delta\vec{b}\|_q}{\|\vec{b}\|_q} = \underbrace{\|\mathbf{A}^{-1}\|_{qp}\|\mathbf{A}\|_{pq}}_{\text{wsk. uwarunkowania}}\frac{\|\delta\vec{b}\|_q}{\|\vec{b}\|_q} = K_{pq}\frac{\|\delta\vec{b}\|_q}{\|\vec{b}\|_q}\end{aligned}$$

$$\delta \mathbf{A} = \mathrm{o} \text{ i } \delta \vec{b} \neq \mathrm{o}$$

- wartość wskaźnika zależy od wyboru norm
- wskaźnik bliski jedności $\rightarrow$ zadanie dobrze uwarunkowane
- duży wskaźnik $\rightarrow$ zadanie źle uwarunkowane
  - nawet niewielkie zaburzenie w wektorze wyrazów wolnych jest wzmacniane i powoduje duży błąd w wyniku

$$\delta \mathbf{A} = 0 \text{ i } \delta \vec{b} \neq 0$$

## Przykład

Rozważmy układ

$$\mathbf{A} = \begin{pmatrix} 1,2969 & 0,8648 \\ 0,2161 & 0,1441 \end{pmatrix}, \quad \mathbf{A}^{-1} = 10^8 \begin{pmatrix} 0,1441 & -0,8648 \\ -0,2161 & 1,2969 \end{pmatrix}$$

## $\delta\mathbf{A} = 0$ i $\delta\vec{b} \neq 0$

Mamy

$$\|\mathbf{A}\|_\infty = 2,1617, \quad \|\mathbf{A}^{-1}\|_\infty = 1,513 * 10^8$$

oraz

$$K = \|\mathbf{A}^{-1}\|_\infty \|\mathbf{A}\|_\infty \approx 3,3 * 10^8$$

- wskaźnik uwarunkowania $\gg 1$
- przy rozwiązaniu układu w najgorszym wypadku możemy utracić 8 miejsc istotnych dokładności
- bardzo złe uwarunkowanie

# Wskaźnik uwarunkowania w praktyce

- w przypadku dużych macierzy wyliczenie wskaźnika może być czasochłonne
- w praktyce często jako kryterium uwarunkowania stosuje się porównanie wartości wyznacznika macierzy **A** z jej elementami
- jeżeli jest on dużo mniejszy niż najmniejszy element macierzy, wówczas zadanie jest na ogół źle uwarunkowane

## $\delta\mathbf{A} \neq \mathrm{o}$ i $\delta\vec{b} = \mathrm{o}$

Z równania macierzowego wynika

$$\delta\vec{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\vec{x} + \delta\vec{x})$$

Wówczas

$$\|\delta\vec{x}\| \leqslant \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|\|\vec{x} + \delta\vec{x}\|$$

czyli

$$\frac{\|\delta\vec{x}\|}{\|\vec{x} + \delta\vec{x}\|} \leqslant \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\| = K\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}$$

# $\delta\mathbf{A} \neq \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

- nawet, jeżeli $\mathbf{A}$ i $\vec{b}$ są znane dokładnie, zwykle nie będą miały dokładnej reprezentacji maszynowej
- najczęściej będziemy mieli do czynienia z sytuacją $\delta\mathbf{A} \neq \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

# $\delta\mathbf{A} \neq \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

Załóżmy, że zaburzenie $\delta\mathbf{A}$ jest na tyle małe, że macierz $\mathbf{A} + \delta\mathbf{A}$ pozostaje nieosobliwa. Wówczas otrzymamy

$$\delta\vec{x} = -\mathbf{A}^{-1}\left(\delta\vec{b} - \delta\mathbf{A}\vec{x} - \delta\mathbf{A}\delta\vec{x}\right)$$

$$\|\delta\vec{x}\| \leqslant \|\mathbf{A}^{-1}\|\left(\|\delta\vec{b}\| + \|\delta\mathbf{A}\|\|\vec{x}\| + \|\delta\mathbf{A}\|\|\delta\vec{x}\|\right)$$

czyli

$$\|\delta\vec{x}\| \leqslant \frac{1}{1 - \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|}\|\mathbf{A}^{-1}\|\left(\|\delta\vec{b}\| + \|\delta\mathbf{A}\|\|\vec{x}\|\right)$$

## $\delta\mathbf{A} \neq \mathrm{o}$ i $\delta\vec{b} \neq \mathrm{o}$

Z równości $\mathbf{A}\vec{x} = \vec{b}$ wynika

$$\frac{\|\vec{b}\|}{\|\vec{x}\|\|\mathbf{A}\|} \leqslant 1$$

Ostatecznie

$$\begin{aligned}
\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} &\leqslant \frac{1}{1 - \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|}\|\mathbf{A}^{-1}\|\|\mathbf{A}\|\left(\frac{\|\delta\vec{b}\|}{\|\vec{b}\|}\frac{\|\vec{b}\|}{\|\vec{x}\|\|\mathbf{A}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right) \\
&\leqslant \frac{K}{1 - \|\mathbf{A}^{-1}\|\|\delta\mathbf{A}\|}\left(\frac{\|\delta\vec{b}\|}{\|\vec{b}\|} + \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}\right)
\end{aligned}$$

# Układy z macierzami trójkątnymi

- szczególnie łatwe do rozwiązania
- aby istniało jednoznaczne rozwiązanie, macierz musi być nieosobliwa…
- …czyli wszystkie elementy na głównej przekątnej muszą być różne od zera

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ &\vdots \\ a_{nn}x_n &= b_n \end{aligned}$$

# Podstawianie w tył

- wstawiając $x_n$ do przedostatniego równania obliczymy $x_{n-1}$
- procedurę kontynuujemy aż do wyliczenia $x_1$

$$\begin{aligned} x_n &= \frac{b_n}{a_{nn}}, \\ x_i &= \frac{b_i - \sum_{k=i+1}^{n} a_{ik} x_k}{a_{ii}}, \ i = n-1, n-2, \ldots, 1 \end{aligned}$$

- koszt obliczeń:

$$M = \frac{1}{2}n^2 + \frac{1}{2}n \ \text{ mnożeń i dzieleń}, \quad D = \frac{1}{2}n^2 - \frac{1}{2}n \ \text{ dodawań}$$

- niewiele większy od kosztu mnożenia wektora przez macierz trójkątną

# Podstawianie w przód

$$\begin{aligned} a_{11}x_1 &= b_1 \\ a_{21}x_1 + a_{22}x_2 &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned}$$

- wstawiając $x_1$ do drugiego równania obliczymy $x_2$ itd.

$$x_1 = \frac{b_1}{a_{11}}, \quad x_i = \frac{b_i - \sum_{k=1}^{i-1} a_{ik}x_k}{a_{ii}}, \ i = 2, 3, \ldots, n$$

- koszt obliczeń ten sam, co poprzednio

# Jak rozwiązać dowolny układ?

1. Sprowadź układ wyjściowy do postaci trójkątnej
2. Zastosuj wzory na podstawianie w tył lub w przód

# Eliminacja Gaussa

- jedna z metod sprowadzenia układu równań do postaci trójkątnej
- nazwana na cześć Carla Friedricha Gaussa
- po raz pierwszy zaprezentowana została dużo wcześniej, bo już około 150 roku p.n.e w słynnym chińskim podręczniku matematyki „Dziewięć rozdziałów sztuki matematycznej”

# Eliminacja Gaussa - algorytm

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned}$$

Odejmujemy od drugiego wiersza pierwszy pomnożony przez $a_{21}/a_{11}$, a od trzeciego pierwszy pomnożony przez $a_{31}/a_{11}$

$$\begin{aligned} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + a_{13}^{(0)}x_3 &= b_1^{(0)} \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{32}^{(1)}x_2 + a_{33}^{(1)}x_3 &= b_3^{(1)} \end{aligned}$$

# Eliminacja Gaussa - algorytm

Przy tym

$$a_{ij}^{(0)} = a_{ij}, \quad b_i^{(0)} = b_i, \quad i, j = 1, 2, 3$$

oraz

$$a_{ij}^{(1)} = a_{ij}^{(0)} - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} a_{1j}^{(0)}, \quad b_i^{(1)} = b_i^{(0)} - \frac{a_{i1}^{(0)}}{a_{11}^{(0)}} b_1^{(0)}, \quad i, j = 2, 3$$

# Eliminacja Gaussa - algorytm

Odejmujemy od trzeciego równania drugie pomnożone przez $a_{32}^{(1)}/a_{22}^{(1)}$

$$\begin{aligned} a_{11}^{(0)}x_1 + a_{12}^{(0)}x_2 + a_{13}^{(0)}x_3 &= b_1^{(0)} \\ a_{22}^{(1)}x_2 + a_{23}^{(1)}x_3 &= b_2^{(1)} \\ a_{33}^{(2)}x_3 &= b_3^{(2)} \end{aligned}$$

$$a_{ij}^{(2)} = a_{ij}^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}a_{2j}^{(1)}, \quad b_i^{(2)} = b_i^{(1)} - \frac{a_{i2}^{(1)}}{a_{22}^{(1)}}b_2^{(1)}, \quad i,j = 3$$

# Eliminacja Gaussa - przypadek ogólny

$$a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} a_{kj}^{(k-1)}, \quad i,j = k+1, k+2, \ldots, n,$$

$$b_{i}^{(k)} = b_{i}^{(k-1)} - \frac{a_{ik}^{(k-1)}}{a_{kk}^{(k-1)}} b_{k}^{(k-1)}, \quad i,j = k+1, k+2, \ldots, n.$$

- otrzymaliśmy układ trójkątny
- jego rozwiązanie ma postać

$$x_i = \frac{b_i^{(i-1)} - \sum_{j=i+1}^{n} a_{ij}^{(i-1)} x_j}{a_{ii}^{(i-1)}}, \quad i = n, n-1, \ldots, 1$$

# Eliminacja Gaussa - przypadek ogólny

- nakład obliczeń to

$$M = \frac{1}{3}n^3 + n^2 - \frac{1}{3} \text{ mnożeń i dzieleń}$$

$$D = \frac{1}{3}n^3 + \frac{1}{2}n^2 - \frac{5}{6}n \text{ dodawań}$$

- większa część przypada na sprowadzenie układu do postaci trójkątnej
- liczba operacji bez porównania mniejsza niż w przypadku wzorów Cramera

# Niezawodność eliminacji Gaussa

## Przykład

$$\begin{pmatrix} 0 & 2 & 2 \\ 3 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

- macierz jest nieosobliwa, a zatem istnieje jednoznaczne rozwiązanie
- mimo to eliminacja Gaussa zawodzi już w pierwszym kroku
- algorytm wymaga dzielenia przez $a_{11}$, które tutaj jest równe 0
  $\Rightarrow$ eliminacja Gaussa w formie przedstawionej powyżej nie jest niezawodna

# Częściowy wybór elementu podstawowego

## Definicja

Elementem podstawowym nazywamy ten element macierzy **A**, za pomocą którego dokonujemy eliminacji zmiennej z dalszych równań.

- rozwiązanie równania nie zmieni się, jeżeli zamienimy kolejność wierszy w układzie równań
- możemy to wykorzystać, aby uniknąć problemów związanych z dzieleniem przez zero

# Częściowy wybór elementu podstawowego

$$\begin{pmatrix} 0 & 2 & 2 \\ 3 & 3 & 0 \\ 1 & 0 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 1 \\ 3 \\ 2 \end{pmatrix}$$

Rozważmy macierz rozszerzoną (z położeniem wierszy):

$$\left(\begin{array}{ccc|c} 0 & 2 & 2 & 1 \\ 3 & 3 & 0 & 3 \\ 1 & 0 & 1 & 2 \end{array}\right) \begin{array}{c} : \\ : \\ : \end{array} \begin{array}{c} w1 \\ w2 \\ w3 \end{array}$$

# Częściowy wybór elementu podstawowego

Zamieniamy wiersze w macierzy układu, tak aby nowy element diagonalny w jej pierwszym wierszu był różny od zera:

$$\left(\begin{array}{ccc|c} 3 & 3 & 0 & 3 \\ 0 & 2 & 2 & 1 \\ 1 & 0 & 1 & 2 \end{array}\right) \begin{array}{c} : \\ : \\ : \end{array} \begin{array}{c} w1^{(1)} \\ w2^{(1)} \\ w3^{(1)} \end{array}$$

# Częściowy wybór elementu podstawowego

Po zamianie wierszy możemy wykonać pierwszy krok eliminacji Gaussa

$$
\begin{array}{lcl}
w1^{(1)} & \rightarrow & \left(\begin{array}{ccc|c} 3 & 3 & 0 & 3 \end{array}\right) : w1^{(2)} \\
w2^{(1)} - (a_{21}^{(1)}/a_{11}^{(1)}) \times w1^{(1)} & \rightarrow & \left(\begin{array}{ccc|c} 0 & 2 & 2 & 1 \end{array}\right) : w2^{(2)} \\
w3^{(1)} - (a_{31}^{(1)}/a_{11}^{(1)}) \times w1^{(1)} & \rightarrow & \left(\begin{array}{ccc|c} 0 & -1 & 1 & 1 \end{array}\right) : w3^{(2)}
\end{array}
$$

# Częściowy wybór elementu podstawowego

W kolejnym kroku nie musimy zamieniać wierszy ze sobą:

$$\begin{array}{lcl} w1^{(2)} & \rightarrow & \left(\begin{array}{cccc} 3 & 3 & 0 & 3 \\ 0 & 2 & 2 & 1 \\ 0 & 0 & 2 & 3/2 \end{array}\right) \begin{array}{c} : \\ : \\ : \end{array} \begin{array}{l} w1^{(3)} \\ w2^{(3)} \\ w3^{(3)} \end{array} \\ w2^{(2)} & \rightarrow & \\ w3^{(2)} - (a_{32}^{(2)}/a_{22}^{(2)}) \times w2^{(2)} & \rightarrow & \end{array}$$

Końcowe rozwiązanie znajdziemy podstawiając w tył

$$x_3 = \frac{b_3^{(3)}}{a_{33}^{(3)}} = \frac{3}{4}, \quad x_2 = \frac{b_2^{(3)} - a_{23}^{(3)} x_3}{a_{22}^{(3)}} = -\frac{1}{4}, \quad x_1 = \frac{b_1^{(3)} - a_{12}^{(3)} x_2 - a_{13}^{(3)} x_3}{a_{11}^{(3)}} = \frac{5}{4}$$

# Częściowy wybór elementu podstawowego

- teoretycznie możemy dowolnie dobierać wiersze do zamiany
- ze względu na błędy zaokrągleń w *i*-tym kroku eliminacji powinniśmy wybierać wiersz, który ma największy element w *i*-tej kolumnie
- częściowy wybór elementu podstawowego zalecany jest również dla układów, których macierze nie mają zerowych elementów na głównej przekątnej, ponieważ w większości przypadków prowadzi do redukcji błędów zaokrągleń

# Częściowy wybór elementu podstawowego - ograniczenia

- nie zawsze prowadzi do poprawy dokładności obliczeń
- odpowiedni wybór jest sprawą delikatną
- czasami warto przeprowadzić równoważenie układu
- ostatecznie można zmienić strategię wyboru z częściowego na całkowity
  - bierzemy pod uwagę wartości elementów w *i*-tej kolumnie i w *i*-tym wierszu
  - duży nakład obliczeń

# Ograniczenia - przykład

$$\begin{pmatrix} 10^{-15} & 1 \\ 1 & 10^{11} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + 10^{-15} \\ 10^{11} + 1 \end{pmatrix}$$

Przy dokładnych obliczeniach eliminacja Gaussa bez wyboru elementu podstawowego da poprawne rozwiązanie

$$\left(\begin{array}{cc|c} 10^{-15} & 1 & 1 + 10^{-15} \\ 1 & 10^{11} & 10^{11} + 1 \end{array}\right) \xrightarrow{el.} \left(\begin{array}{cc|c} 1 & 10^{15} & 1 + 10^{15} \\ 0 & 10^{11} - 10^{15} & 10^{11} - 10^{15} \end{array}\right)$$

$$\xrightarrow{podstawianie} \vec{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

# Ograniczenia - przykład

Błędy zaokrągleń spowodują, że wynik będzie znacznie odbiegał od idealnego:

$$\xrightarrow{el.} \left(\begin{array}{cc|c} 1 & 9.999999999999999e+14 & 1.0000000000000001e+015 \\ 0 & -9.9989999999999999e+014 & -9.9990000000000000e+014 \end{array}\right)$$

$$\xrightarrow{podstawianie} \vec{x} = \begin{pmatrix} 8.7500000000000000e-001 \\ 1.0000000000000000e+000 \end{pmatrix}$$

# Ograniczenia - przykład

Lepszy wynik uzyskamy, dokonując częściowego wyboru elementu podstawowego:

$$\left(\begin{array}{cc|c} 10^{-15} & 1 & 1+10^{-15} \\ 1 & 10^{11} & 10^{11}+1 \end{array}\right) \xrightarrow{\textit{zamiana wierszy}} \left(\begin{array}{cc|c} 1 & 10^{11} & 10^{11}+1 \\ 10^{-15} & 1 & 1+10^{-15} \end{array}\right)$$

$$\xrightarrow{\textit{eliminacja}} \left(\begin{array}{cc|c} 1 & 1.000e+011 & 1.0000000000010000e+011 \\ 0 & 9.999e-001 & 9.9990000000000001e-001 \end{array}\right)$$

$$\xrightarrow{\textit{podstawianie}} \vec{x} = \begin{pmatrix} 9.9998474121093750e-001 \\ 1.0000000000000000e+000 \end{pmatrix}$$

# Ograniczenia - przykład 2

Rozważmy układ

$$\begin{pmatrix} 10^{-14.6} & 1 \\ 1 & 10^{15} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 + 10^{-14.6} \\ 10^{15} + 1 \end{pmatrix}$$

Jego dokładne rozwiązanie wynosi $\vec{x} = (1, 1)^{\mathrm{T}}$. Eliminacja Gaussa da poprawny wynik

$$\xrightarrow{\textit{eliminacja}} \left(\begin{array}{cc|c} 1 & 3.9810717055349 69e+014 & 3.9810717055349 79e+014 \\ 0 & 6.0189282944650 30e+014 & 6.0189282944650 30e+014 \end{array}\right)$$

$$\xrightarrow{\textit{podstawianie}} \vec{x} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

# Ograniczenia - przykład 2

Częściowy wybór elementu podstawowego „zepsuje” wynik

$$\left(\begin{array}{cc|c} 10^{-14.6} & 1 & 1+10^{-14.6} \\ 1 & 10^{15} & 10^{15}+1 \end{array}\right)$$

$$\xrightarrow{\textit{zamiana wierszy}} \left(\begin{array}{cc|c} 1 & 10^{15} & 10^{15}+1 \\ 10^{-14.6} & 1 & 1+10^{-14.6} \end{array}\right)$$

$$\xrightarrow{\textit{eliminacja}} \left(\begin{array}{cc|c} 1 & 1.000e+015 & 1.0000000000000001e+015 \\ 0 & -1.5118864315095819 & -1.5118864315095821 \end{array}\right)$$

$$\xrightarrow{\textit{podstawianie}} \vec{x} = \begin{pmatrix} 0.750000000000000 \\ 1.000000000000002 \end{pmatrix}$$

# Równoważenie układu - przykład

Rozważmy układ

$$\begin{pmatrix} 1 & 10000 \\ 1 & 0,0001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 10000 \\ 1 \end{pmatrix}$$

Układ ten ma rozwiązanie $x_1 = x_2 = 0,9999$, poprawnie zaokrąglone do czterech cyfr dziesiętnych.

Przyjmijmy $a_{11}$ jako element podstawowy i poszukajmy rozwiązań układu w trzycyfrowej arytmetyce zmiennopozycyjnej. Otrzymamy następujące, złe rozwiązanie

$$x_1 = 0.00, \quad x_2 = 1.00.$$

# Równoważenie układu - przykład

Pomnóżmy teraz pierwsze równanie przez $10^{-4}$

$$\begin{pmatrix} 0,0001 & 1 \\ 1 & 0,0001 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

Wybierając $a_{21}$ jako element podstawowy, otrzymamy

$$x_1 = 1.00, \quad x_2 = 1.00,$$

co w trzycyfrowej arytmetyce jest wynikiem dobrym

# Eliminacja Gaussa i macierze osobliwe - przykład

$$\begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 1 \\ 1 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

Po kilku krokach dojdziemy do sytuacji (sprawdzić!):

$$\left(\begin{array}{ccc|c} 1 & 0 & 1 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{array}\right)$$

# Eliminacja Gaussa i macierze osobliwe - przykład

- same zera w ostatnim wierszu sygnalizują, że wyjściowa macierz była osobliwa
- nie istnieje rozwiązanie jednoznaczne
- ponieważ ostatni element wektora wyrazów wolnych jest również równy zero, rozwiązań jest nieskończenie wiele

# Macierze odwrotne

- wiele układów różniących się tylko wyrazem wolnym

$$\mathbf{A}\vec{x}_1 = \vec{b}_1, \quad \mathbf{A}\vec{x}_2 = \vec{b}_2, \quad \ldots, \mathbf{A}\vec{x}_N = \vec{b}_N$$

$$\mathbf{A}\left(\vec{x}_1\vec{x}_2\ldots\vec{x}_N\right) = \left(\vec{b}_1\vec{b}_2\ldots\vec{b}_N\right)$$

$$\mathbf{AX} = \mathbf{B}$$

- formalne rozwiązanie ostatniego równania macierzowego ma postać

$$\mathbf{X} = \mathbf{A}^{-1}\mathbf{B}$$

- jeżeli **B** będzie macierzą jednostkową, znajdziemy w ten sposób macierz odwrotną do macierzy **A**

# Eliminacja Jordana

$$
\begin{aligned}
a_{11}^{(1)}x_1 &+ a_{12}^{(1)}x_2 + \ldots + a_{1n}^{(1)}x_n &&= b_1^{(1)} \\
a_{21}^{(1)}x_1 &+ a_{22}^{(1)}x_2 + \ldots + a_{2n}^{(1)}x_n &&= b_2^{(1)} \\
&&&\vdots \\
a_{n1}^{(1)}x_1 &+ a_{n2}^{(1)}x_2 + \ldots + a_{nn}^{(1)}x_n &&= b_n^{(1)}
\end{aligned}
$$

# Eliminacja Jordana

Dzielimy pierwsze równanie obustronnie przez $a_{11}^{(1)}$,a następnie od $i$–tego wiersza ($i = 2, 3, \ldots, n$) odejmujemy pierwszy pomnożony przez $a_{i1}^{(1)}$,

$$\begin{aligned} x_1 \quad + \quad a_{12}^{(2)} x_2 \quad + \quad \ldots \quad + \quad a_{1n}^{(2)} x_n \quad &= \quad b_1^{(2)} \\ a_{22}^{(2)} x_2 \quad + \quad \ldots \quad + \quad a_{2n}^{(2)} x_n \quad &= \quad b_2^{(2)} \\ &\vdots \\ a_{n2}^{(2)} x_2 \quad + \quad \ldots \quad + \quad a_{nn}^{(2)} x_n \quad &= \quad b_n^{(2)} \end{aligned}$$

# Eliminacja Jordana

W kolejnym kroku dzielimy drugie równanie obustronnie przez $a_{22}^{(2)}$ i odejmujemy od $i$–tego wiersza ($i = 1, 3, 4, \ldots, n$) wiersz drugi pomnożony przez $a_{i2}^{(2)}$,

$$\begin{array}{rcccccl} x_1 & & + & \ldots & + & a_{1n}^{(3)} x_n & = & b_1^{(3)} \\ & x_2 & + & \ldots & + & a_{2n}^{(3)} x_n & = & b_2^{(3)} \\ & & & & & \vdots & & \\ & & & \ldots & + & a_{nn}^{(3)} x_n & = & b_n^{(3)} \end{array}$$

# Eliminacja Jordana

Po $(n-1)$ eliminacjach otrzymujemy układ

$$\begin{aligned} x_1 &= b_1^{(n)} \\ x_2 &= b_2^{(n)} \\ \ddots &\ \vdots \\ x_n &= b_n^{(n)} \end{aligned}$$

# Eliminacja Jordana

- koszt obliczeń
$$M = \frac{1}{2}n^3 + \frac{1}{2}n^2, \quad D = \frac{1}{2}n^3 - \frac{1}{2}$$
- potrzebujemy wyboru elementu podstawowego w celu zagwarantowania niezawodności
- zalety:
  - oszczędne gospodarowanie pamięcią
  - możliwość określenia rozwiązania „obciętego” układu równań
- wady:
  - duży nakład obliczeń (około 1,5 raza większy niż w eliminacji Gaussa)
  - brak odpowiednika rozkładu **LU** (o tym zaraz)

# Rozkład LU

- przypuśćmy, że macierz **A** układu da się przedstawić w postaci iloczynu macierzy trójkątnej dolnej **L** i trójkątnej górnej **U**

$$\mathbf{A} = \mathbf{LU}$$

- jeżeli macierz **A** jest nieosobliwa, zachodzi

$$\mathbf{A}^{-1} = (\mathbf{LU})^{-1} = \mathbf{U}^{-1}\mathbf{L}^{-1}$$

- rozwiązanie układu da się przedstawić w postaci

$$\vec{x} = \mathbf{A}^{-1}\vec{b} = \mathbf{U}^{-1}\left(\mathbf{L}^{-1}\vec{b}\right)$$

# Rozkład LU

$\Rightarrow$ aby znaleźć rozwiązanie $\vec{x}$ układu dysponując rozkładem **LU** jego macierzy, wystarczy rozwiązać dwa układy trójkątne

$$\begin{aligned} \mathbf{L}\vec{y} &= \vec{b} \\ \mathbf{U}\vec{x} &= \vec{y} \end{aligned}$$

# Eliminacja Gaussa a rozkład LU

Przekształcenie

$$\mathbf{A}^{(1)}x = b^{(1)} \quad \rightarrow \quad \mathbf{A}^{(2)}x = b^{(2)}$$

jest równoważne pomnożeniu obu stron układu $\mathbf{A}^{(1)}x = b^{(1)}$ przez macierz

$$\mathbf{L}^{(1)} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ -l_{21} & 1 & 0 & \dots & 0 \\ -l_{31} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -l_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}, \quad l_{i1} = \frac{a_{i1}^{(1)}}{a_{11}^{(1)}}, \quad i = 2, 3, \dots, n$$

# Eliminacja Gaussa a rozkład LU

W ten sposób otrzymujemy dwa równania:

$$\mathbf{L}^{(1)}\mathbf{A}^{(1)} = \mathbf{A}^{(2)}, \quad \mathbf{L}^{(1)}b^{(1)} = b^{(2)}$$

Podobnie

$$\mathbf{L}^{(2)}\mathbf{A}^{(2)} = \mathbf{A}^{(3)}, \quad \mathbf{L}^{(2)}b^{(2)} = b^{(3)}$$

$$\mathbf{L}^{(2)} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & -l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & -l_{n2} & 0 & \dots & 1 \end{pmatrix}, \quad l_{i2} = \frac{a_{i2}^{(2)}}{a_{22}^{(2)}}, \quad i = 3, \dots, n$$

# Eliminacja Gaussa a rozkład LU

Ostatecznie

$$\mathbf{L}^{(n-1)}\mathbf{L}^{(n-2)}\dots\mathbf{L}^{(1)}\mathbf{A}^{(1)} = \mathbf{A}^{(n)}$$

oraz

$$\mathbf{L}^{(n-1)}\mathbf{L}^{(n-2)}\dots\mathbf{L}^{(1)}b^{(1)} = b^{(n)}$$

Macierze $\mathbf{L}^{(i)}$, $i = 1, \dots, n-1$ są nieosobliwe, więc

$$\mathbf{A}^{(1)} = (\mathbf{L}^{(1)})^{-1}(\mathbf{L}^{(2)})^{-1}\dots(\mathbf{L}^{(n)})^{-1}\mathbf{A}^{(n)}$$

# Eliminacja Gaussa a rozkład LU

Ponadto

$$(\mathbf{L}^{(1)})^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & 0 & 0 & \dots & 1 \end{pmatrix}, \quad (\mathbf{L}^{(2)})^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & l_{n2} & 0 & \dots & 1 \end{pmatrix} \dots$$

# Eliminacja Gaussa a rozkład LU

Stąd

$$\mathbf{L} \equiv (\mathbf{L}^{(1)})^{-1}(\mathbf{L}^{(2)})^{-1}\dots(\mathbf{L}^{(n)})^{-1} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ l_{21} & 1 & 0 & \dots & 0 \\ l_{31} & l_{32} & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & 1 \end{pmatrix}.$$

Z drugiej strony wiemy, że $\mathbf{A}^{(n)} = \mathbf{U}$ jest macierzą trójkątną górną.

# Eliminacja Gaussa a rozkład LU

- zapamiętując macierze **L** i **U**, możemy szybko rozwiązać wiele układów różniących się tylko kolumnami wyrazów wolnych
- w ramach oszczędności pamięci możemy zapisywać elementy tych macierzy w miejsce elementów macierzy **A**,

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{pmatrix} \rightarrow \begin{pmatrix} u_{11} & u_{12} & u_{13} & \dots & u_{1n} \\ l_{21} & u_{22} & u_{23} & \dots & u_{2n} \\ l_{31} & l_{32} & u_{33} & \dots & u_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ l_{n1} & l_{n2} & l_{n3} & \dots & u_{nn} \end{pmatrix}$$

# Eliminacja Gaussa a rozkład LU

- nie każdą macierz nieosobliwą można przedstawić w postaci
- aby rozkład istniał, wszystkie minory główne macierzy muszą być różne od zera
- jeżeli eliminację Gaussa można przeprowadzić do końca, rozkład LU na pewno istnieje

# Rozkład LU a wybór elementu podstawowego

Jeżeli eliminacja Gaussa wymaga zamiany wierszy, wówczas zamiast rozkładu LU macierzy **A** znajdziemy rozkład permutacji jej wierszy

$$\mathbf{PA} = \mathbf{LU}$$

Znaczenie macierzy permutacji **P** ilustruje następujący przykład:

$$\mathbf{PA} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} a_{31} & a_{32} & a_{33} \\ a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}$$

# Rozkład LU a wybór elementu podstawowego

Macierz permutacji ma następującą własność:

$$\mathbf{P}^{\mathrm{T}}\mathbf{P} = \mathbf{I} \;\Rightarrow\; \mathbf{P}^{\mathrm{T}} = \mathbf{P}^{-1}$$

Stąd wynika

$$\mathbf{A} = \mathbf{P}^{\mathrm{T}}\mathbf{L}\mathbf{U}$$

# Rozkład LU i metoda Doolittle'a

Potraktujmy równość

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ l_{21} & 1 & 0 \\ l_{31} & l_{32} & 1 \end{pmatrix} \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{pmatrix}$$

jako układ $n^2$ równań dla $n^2$ niewiadomych $l_{ij}$ i $u_{ij}$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} & u_{13} \\ l_{21}u_{11} & l_{21}u_{12} + u_{22} & l_{21}u_{13} + u_{23} \\ l_{31}u_{11} & l_{31}u_{12} + l_{32}u_{22} & l_{31}u_{13} + l_{32}u_{23} + u_{33} \end{pmatrix}$$

# Rozkład LU i metoda Doolittle'a

Stąd

$$
\begin{array}{lll}
u_{11} = a_{11}, & u_{12} = a_{12}, & u_{13} = a_{13} \\
l_{21} = \frac{a_{21}}{u_{11}}, & u_{22} = a_{21} - l_{21}u_{12}, & u_{23} = a_{23} - l_{21}u_{13} \\
l_{31} = \frac{a_{31}}{u_{11}}, & l_{32} = \frac{a_{32} - l_{31}u_{12}}{u_{22}} & u_{33} = a_{33} - l_{31}u_{13} - l_{32}u_{23}
\end{array}
$$

# Rozkład LU i metoda Doolittle'a

- w przypadku ogólnym elementy macierzy **L** i **U** obliczamy dla $i = 1, 2, \ldots, n$ ze wzorów

$$
\begin{aligned}
u_{ij} &= a_{ij} - \sum_{k=1}^{i-1} l_{ik} u_{kj}, \quad j = i, i+1, \ldots, n \\
l_{ji} &= \frac{a_{ji} - \sum_{k=1}^{i-1} l_{jk} u_{ki}}{u_{ii}}, \quad j = i+1, i+2, \ldots, n
\end{aligned}
$$

# Rozkład LU i metoda Doolittle'a

- koszt obliczeń (łącznie z rozw. układów trójkątnych)

$$M = \frac{1}{3}n^3 + n^2 - \frac{1}{3}n, \quad D = \frac{1}{3}n^3 + \frac{1}{3}n^2 - \frac{5}{6}n$$

- koszt taki sam, jak w eliminacji Gaussa
- niezawodna w połączeniu z wyborem elementu podstawowego
- wiersze zamieniamy ze sobą miejscami tak, aby element $u_{ii}$ był jak największy

# Metoda Doolitle'a - przykład

Chcemy wyznaczyć rozkład LU macierzy

$$\begin{pmatrix} 20 & 31 & 23 \\ 30 & 24 & 18 \\ 15 & 32 & 21 \end{pmatrix}$$

metodą Doolittle'a z częściowym wyborem elementu podstawowego. W tym celu wprowadzamy dodatkową kolumnę indeksującą wiersze

$$\begin{pmatrix} 20 & 31 & 23 \\ 30 & 24 & 18 \\ 15 & 32 & 21 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$$

# Metoda Doolitle'a - przykład

Element podstawowy wybieramy tak, aby element $u_{ii}$ występujący we wzorach ogólnych miał jak największą wartość.

Dla $i = 1$ w zależności od tego, czy na pierwszym miejscu ustawimy wiersz pierwszy, drugi czy trzeci, uzyskamy odpowiednio $u_{11} = 20$, $u_{11} = 30$ oraz $u_{11} = 15$.

Zamieniamy miejscami wiersz pierwszy z drugim

$$\begin{pmatrix} 30 & 24 & 18 \\ 20 & 31 & 23 \\ 15 & 32 & 21 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$

# Metoda Doolitle'a - przykład

Otrzymamy

$$u_{11} = 30, \quad u_{12} = a_{21} = 24, \quad u_{13} = a_{13} = 18$$
$$l_{21} = \frac{2}{3}, \qquad l_{31} = \frac{1}{2}$$

Wartości te wpisujemy do macierzy **A**

$$\begin{pmatrix} 30 & 24 & 18 \\ \frac{2}{3} & 31 & 23 \\ \frac{1}{2} & 32 & 21 \end{pmatrix} \begin{pmatrix} 2 \\ 1 \\ 3 \end{pmatrix}$$

# Metoda Doolitle'a - przykład

Dla $i = 2$ otrzymamy

$$u_{22} = a_{22} - a_{21}a_{12} = 31 - \frac{2}{3} * 24 = 15$$

lub

$$u_{22} = a_{32} - a_{31}a_{12} = 32 - \frac{1}{2} * 24 = 20$$

w zależności od tego, czy na drugim miejscu ustawimy wiersz drugi czy trzeci.

# Metoda Doolitle'a - przykład

Zamieniamy wiersze miejscami

$$\begin{pmatrix} 30 & 24 & 18 \\ \frac{1}{2} & 32 & 21 \\ \frac{2}{3} & 31 & 23 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

Znajdujemy

$$u_{22} = 20, \quad u_{23} = a_{23} - a_{21}a_{13} = 12, \quad u_{32} = \frac{15}{20}$$

# Metoda Doolitle'a - przykład

Uzyskane wartości wpisujemy do macierzy

$$\begin{pmatrix} 30 & 24 & 18 \\ \frac{1}{2} & 20 & 12 \\ \frac{2}{3} & \frac{3}{4} & 23 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

Dla $i = 3$ obliczamy

$$u_{33} = 2.$$

# Metoda Doolitle'a - przykład

Stąd

$$\begin{pmatrix} 30 & 24 & 18 \\ \frac{1}{2} & 20 & 12 \\ \frac{2}{3} & \frac{3}{4} & 2 \end{pmatrix} \begin{pmatrix} 2 \\ 3 \\ 1 \end{pmatrix}$$

W ten sposób w miejsce macierzy **A** otrzymaliśmy rozkład **LU** macierzy, która składa się z wierszy 2, 3 i 1 macierzy wyjściowej **A**.

# Rozkład LU i metoda Crouta

Przyjmujemy dla odmiany, że **U** ma na głównej przekątnej same jedynki

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{pmatrix} = \begin{pmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{pmatrix} \begin{pmatrix} 1 & u_{12} & u_{13} \\ 0 & 1 & u_{23} \\ 0 & 0 & 1 \end{pmatrix}$$

i ponownie potraktujemy powyższe wyrażenie jak równanie na niewiadome elementy macierzy trójkątnych.

# Rozkład LU i wyznaczniki

$$\det \mathbf{A} = \det(\mathbf{LU}) = \det \mathbf{L} \det \mathbf{U} = \begin{cases} u_{11}u_{22}\ldots u_{nn}, & l_{ii} = 1 \\ l_{11}l_{22}\ldots l_{nn}, & u_{ii} = 1 \end{cases}$$

# Macierze dominujące diagonalnie

## Definicja

Macierz kwadratową **A** nazywamy diagonalnie dominującą, jeżeli

$$|a_{ii}| \geqslant \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ik}|, \quad i = 1, 2, \ldots, n$$

Jeżeli nierówności są ostre, mówimy o macierzy silnie diagonalnie dominującej.

# Macierze dominujące diagonalnie

## Definicja

Macierz **A** jest diagonalnie dominująca kolumnowo, jeżeli $\mathbf{A}^{\mathrm{T}}$ jest diagonalnie dominująca, tzn.

$$|a_{ii}| \geqslant \sum_{\substack{k=1 \\ k \neq i}}^{n} |a_{ki}|, \quad i = 1, 2, \ldots, n$$

# Macierze dominujące diagonalnie

## Twierdzenie

*Jeżeli macierz* **A** *jest nieosobliwa i diagonalnie dominująca kolumnowo, to przy eliminacji metodą Gaussa nie ma potrzeby przestawiania wierszy.*

# Macierze trójdiagonalne

$$\mathbf{T} = \begin{pmatrix} b_1 & c_1 & & & & \\ a_2 & b_2 & c_2 & & \mathrm{o} & \\ & a_3 & b_3 & c_3 & & \\ & & a_4 & b_4 & \ddots & \\ & & & \ddots & \ddots & \ddots \\ & \mathrm{o} & & & \ddots & b_{n-1} & c_{n-1} \\ & & & & & a_n & b_n \end{pmatrix}$$

# Rozkład LU macierzy trójdiagonalnej

$$\mathbf{L} = \begin{pmatrix} 1 & & & 0 \\ l_2 & \ddots & & \\ & \ddots & \ddots & \\ 0 & & l_n & 1 \end{pmatrix}, \quad \mathbf{U} = \begin{pmatrix} u_1 & c_1 & & 0 \\ & \ddots & \ddots & \\ & & \ddots & c_{n-1} \\ 0 & & & u_n \end{pmatrix}$$

$$u_1 = b_1, \quad l_i = \frac{a_i}{u_{i-1}}, \quad u_i = b_i - l_i c_{i-1}, \quad i = 2, 3, \ldots, n$$

- rozkład wymaga $O(n)$ operacji
- metoda niezawodna, jeśli $\mathbf{T}$ jest diagonalnie dominująca kolumnowo

# Błędy zaokrągleń

- macierze **L** i **U** spełniają warunek

$$\mathbf{LU} = \mathbf{A} + \mathbf{E}$$

$\mathbf{E}$ - błąd rozkładu

- $\vec{y}$ i $\vec{x}$ możemy potraktować jako dokładne rozwiązania układów

$$\begin{aligned}(\mathbf{L} + \delta\mathbf{L})\vec{y} &= \vec{b} \\ (\mathbf{U} + \delta\mathbf{U})\vec{x} &= \vec{y}\end{aligned}$$

# Błędy zaokrągleń

Stąd

$$(\mathbf{A} + \mathbf{E} + \mathbf{L}\delta\mathbf{U} + \delta\mathbf{L}\mathbf{U} + \delta\mathbf{L}\delta\mathbf{U})\vec{x} = \vec{b}$$

Można pokazać, że zaburzenie

$$\delta\mathbf{A} = \mathbf{E} + \mathbf{L}\delta\mathbf{U} + \delta\mathbf{L}\mathbf{U} + \delta\mathbf{L}\delta\mathbf{U}$$

ma oszacowanie

$$\frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|} \leqslant \epsilon\left(\frac{9}{2}n^3 + \frac{61}{2}n^2 - 18n - 16\right) + O(\epsilon)$$

gdzie $\epsilon$ to dokładność maszynowa. Stąd wynika

$$\frac{\|\delta\vec{x}\|}{\|\vec{x}\|} \leqslant \frac{\alpha}{1-\alpha}, \quad \alpha = \epsilon KO(\frac{9}{2}n^3)$$

# Inne rozkłady macierzy

- rozkład LU nie jest jedynym przydatnym rozkładem macierzy
- do innych często stosowanych rozkładów należą
  - rozkład Cholesky'ego (Banachiewicza)
  - rozkład SVD
  - rozkład QR

# Rozkład Cholesky'ego (Banachiewicza)

Jeżeli macierz układu jest macierzą symetryczną, tzn.

$$a_{ij} = a_{ji}, \quad i, j = 1, \ldots, n$$

i dodatnio określoną

$$\vec{x}^{\mathrm{T}} \mathbf{A} \vec{x} > 0 \text{ dla każdego } \vec{x}$$

to istnieje dla niej bardziej wydajny od LU rozkład na macierze trójkątne

$$\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$$

gdzie **L** to macierz trójkątna dolna

# Rozkład Cholesky'ego (Banachiewicza)

Traktując ostatnie równanie jako układ równań ze względu na elementy macierzy **L**, znajdziemy:

$$
\begin{aligned}
l_{ii} &= \left( a_{ii} - \sum_{k=1}^{i-1} l_{ik}^2 \right)^{1/2} \\
l_{ji} &= \frac{1}{l_{ii}} \left( a_{ij} - \sum_{k=1}^{i-1} l_{ik} l_{jk} \right), \quad j = i+1, i+2, \ldots, n
\end{aligned}
$$

- liczba operacji o połowę mniejsza od LU
- niezawodność (metoda nie wymaga wyboru elementu podstawowego)
- stabilność numeryczna

# Rozkład SVD (ang. *Singular Value Decomposition*)

## Twierdzenie

*Każdą macierz* $\mathbf{A} \in \mathbf{R}^{m \times n}$ *rzędu r możemy przedstawić jako*

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}, \quad \Sigma = \begin{pmatrix} \Sigma_1 & \mathbf{O} \\ \mathbf{O} & \mathbf{O} \end{pmatrix} \in \mathbf{R}^{m \times n}, \quad \Sigma_1 = \operatorname{diag}(\sigma_1, \ldots, \sigma_r),$$

*gdzie* $\mathbf{U} \in \mathbf{R}^{m \times m}$ *i* $\mathbf{V} \in \mathbf{R}^{n \times n}$ *są macierzami ortogonalnymi, a* $\sigma_1 \geqslant \sigma_2 \geqslant \cdots \geqslant \sigma_r > 0$. *Elementy* $\sigma_i$ *macierzy* $\Sigma$ *nazywane są wartościami osobliwymi macierzy* $\mathbf{A}$.

# Szkic algorytmu rozkładu SVD

Krok 1 przekształcamy **A** do postaci

$$\mathbf{A} = \mathbf{QCH}^{\mathrm{T}}$$

gdzie **C** to macierz dwudiagonalna, a **Q** i **H** są iloczynami macierzy odpowiadających transformacji Householdera

Krok 2 nadajemy macierzy **C** postać diagonalną,

$$\mathbf{C} = \mathbf{U}'\Sigma'\mathbf{V}'^{\mathrm{T}}$$

gdzie $\mathbf{U}'$ i $\mathbf{V}'$ opisują transformację Givensa

Krok 3 porządkujemy elementy diagonalne macierzami ortogonalnymi $\mathbf{U}''$ i $\mathbf{V}''$, wyrażającymi się poprzez iloczyny macierzy permutacji

$$\Sigma = \mathbf{U}''^{\mathrm{T}}\Sigma'\mathbf{V}''$$

# Szkic algorytmu rozkładu SVD

Macierze **U** i **V** rozkładu SVD to po prostu

$$\mathbf{U} = \mathbf{Q}\mathbf{U}'\mathbf{U}'', \quad \mathbf{V} = \mathbf{H}\mathbf{V}'\mathbf{V}''$$

Zastosowania

- do przybliżonych rozwiązań układów z macierzami osobliwymi albo prawie osobliwymi
- do układów niedookreślonych i nadokreślonych
- numeryczny rząd macierzy
- wskaźnik uwarunkowania macierzy

# Rozkład QR

$$\mathbf{A} = \mathbf{QR}$$

$$\mathbf{Q}^{\mathrm{T}}\mathbf{Q} = \mathbf{1}, \quad \mathbf{R} \text{ - macierz trójkątna górna}$$

Do wyznaczenia tego rozkładu stosuje się zmodyfikowaną metodę Grama-Schmidta. Polega ona na obliczeniu ciągu macierzy

$$\mathbf{A} = \mathbf{A}^{(1)}, \mathbf{A}^{(2)}, \ldots, \mathbf{A}^{(n+1)} = \mathbf{Q},$$

gdzie $\mathbf{A}^{(k)}$ ma postać

$$\mathbf{A}^{(k)} = \left(\vec{q}_1, \ldots, \vec{q}_{k-1}, \vec{a}_k^{(k)}, \ldots, \vec{a}_n^{(k)}\right).$$

Kolumny $\vec{q}_1,\ldots,\vec{q}_{k-1}$ są $k-1$ początkowymi kolumnami macierzy $\mathbf{Q}$, a kolumny $\vec{a}_k^{(k)}, \ldots, \vec{a}_n^{(k)}$ powinny być ortogonalne do $\vec{q}_1,\ldots,\vec{q}_{k-1}$.

# Rozkład QR

Ortogonalność w $k$-tym kroku kolumn od $k+1$ do $n$ względem $\vec{q}_k$ zapewnia się w następujący sposób:

$$\vec{q}_k = \vec{a}_k^{(k)}, \quad d_k = \vec{q}_k^{\mathrm{T}} \vec{q}_k, \quad r_{kk} = 1, \quad \vec{a}_j^{k+1} = \vec{a}_j^{(k)} - r_{jk}\vec{q}_k$$

$$r_{jk} = \frac{\vec{q}_k^{\mathrm{T}} \vec{a}_j^{(k)}}{d_k}, \quad j = k+1, \ldots, n$$

Po $n$ krokach ($k = 1, \ldots, n$) otrzymamy macierze $\mathbf{Q} = \left(\vec{q}_1, \ldots, \vec{q}_n\right)$ i $\mathbf{R} = \left(r_{kj}\right)$ o pożądanych własnościach.

# Iteracyjne poprawianie rozwiązań

- rozwiązanie układu równań $\mathbf{A}\vec{x} = \vec{b}$ dowolną metodą bezpośrednią będzie zwykle obarczone pewnym błędem
- błąd ten możemy wykryć, sprawdzając, jak bardzo tzw. wektor reszt

$$\vec{r} = \vec{b} - \mathbf{A}\vec{x}$$

  różni się od zera
- powinniśmy przy tym liczyć $\vec{r}$ z dokładnością większą niż dokładność uzyskanego rozwiązania

# Przykład

Układ

$$\begin{pmatrix} 0,99 & 0,70 \\ 0,70 & 0,50 \end{pmatrix} \vec{x} = \begin{pmatrix} 0,54 \\ 0,36 \end{pmatrix}$$

ma rozwiązanie dokładne

$$\vec{x}_{dok} = \begin{pmatrix} 0,80 \\ -0,36 \end{pmatrix}$$

# Przykład

Obliczmy najpierw $\vec{r}$ w arytmetyce zmiennopozycyjnej o dwóch miejscach dziesiętnych w mantysie, dokonując zaokrągleń

$$\begin{aligned}\vec{r}(\vec{x}_{dok}) &= \begin{pmatrix} 0,54 \\ 0,36 \end{pmatrix} - \begin{pmatrix} 0,99 & 0,70 \\ 0,70 & 0,50 \end{pmatrix} \begin{pmatrix} 0,80 \\ -0,36 \end{pmatrix} \\ &= \begin{pmatrix} 0,54 - 0,79 + 0,25 \\ 0,38 - 0,56 + 0,18 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\end{aligned}$$

Nie możemy jednak wnioskować stąd, że $\vec{x}_{dok}$ jest dokładnym rozwiązaniem równania.

# Przykład

Dla

$$\vec{x}_1 = \begin{pmatrix} 0,02 \\ 0,74 \end{pmatrix}$$

mamy również

$$\begin{aligned} \vec{r}(\vec{x}_1) &= \begin{pmatrix} 0,54 \\ 0,36 \end{pmatrix} - \begin{pmatrix} 0,99 & 0,70 \\ 0,70 & 0,50 \end{pmatrix} \begin{pmatrix} 0,02 \\ 0,74 \end{pmatrix} \\ &= \begin{pmatrix} 0,54 - 0,02 - 0,52 \\ 0,38 - 0,01 - 0,37 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \end{aligned}$$

# Przykład

$\vec{x}_1$ rozwiązaniem równania nie jest i różni się dość sporo od rozwiązania dokładnego,

$$\|\vec{x}_{dok} - \vec{x}_1\|_\infty = 1,1$$

# Przykład

Policzmy teraz wektory reszt z większą liczbą miejsc dziesiętnych w mantysie. Otrzymamy

$$\begin{aligned}\vec{r}(\vec{x}_{dok}) &= \begin{pmatrix} 0,54 \\ 0,36 \end{pmatrix} - \begin{pmatrix} 0,99 & 0,70 \\ 0,70 & 0,50 \end{pmatrix} \begin{pmatrix} 0,80 \\ -0,36 \end{pmatrix} \\ &= \begin{pmatrix} 0,54 - 0,792 + 0,252 \\ 0,38 - 0,56 + 0,18 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}\end{aligned}$$

# Przykład

oraz

$$\begin{aligned}\vec{r}(\vec{x}_1) &= \begin{pmatrix} 0,54 \\ 0,36 \end{pmatrix} - \begin{pmatrix} 0,99 & 0,70 \\ 0,70 & 0,50 \end{pmatrix} \begin{pmatrix} 0,02 \\ 0,74 \end{pmatrix} \\ &= \begin{pmatrix} 0,54 - 0,0198 - 0,518 \\ 0,38 - 0,014 - 0,37 \end{pmatrix} = \begin{pmatrix} 0,0022 \\ -0,004 \end{pmatrix}\end{aligned}$$

Dopiero teraz widać, że $\vec{x}_{dok}$ jest rozwiązaniem naszego układu równań, natomiast $\vec{x}_1$ nim nie jest.

# Pierwsza poprawka rozwiązania

Szukamy poprawki $\delta\vec{x}$ takiej, że

$$\vec{x} + \delta\vec{x} = \vec{x}_{dok}$$

Ponieważ zachodzi

$$\vec{r} = \vec{b} - \mathbf{A}\vec{x} = \mathbf{A}\vec{x}_{dok} - \mathbf{A}\vec{x} = \mathbf{A}(\vec{x}_{dok} - \vec{x}) = \mathbf{A}\delta\vec{x}$$

wystarczy, że rozwiążemy układ

$$\vec{r} = \mathbf{A}\delta\vec{x}$$

- łatwe, jeżeli dysponujemy już rozkładem LU macierzy **A**
- wymaga $n^2$ mnożeń i $n^2 - n$ dodawań

# Dalsze poprawki

- w rzeczywistych obliczeniach numerycznych nie potrafimy liczyć dokładnie
- zamiast poprawki $\delta\vec{x}$ znajdziemy tylko poprawkę przybliżoną

$$\delta\vec{x} + \delta(\delta\vec{x})$$

- do ulepszonego rozwiązania $\vec{x} + \delta\vec{x}$ możemy znaleźć kolejną poprawkę

# Przepis praktyczny

1. rozwiąż układ równań $\mathbf{A}\vec{x}^{(1)} = \vec{b}$ stosując rozkład LU macierzy $\mathbf{A}$
2. oblicz wektor reszt $\vec{r}^{(1)} = \vec{b} - \mathbf{A}\vec{x}^{(1)}$ (w podwójnej precyzji)
3. jeśli $\|\vec{r}^{(1)}\|_\infty \leqslant \|\mathbf{A}\vec{x}^{(1)}\|_\infty u$ (lub $\|\vec{r}^{(1)}\|_\infty \leqslant \|\vec{b}\|_\infty u$), gdzie $u$ to jednostka maszynowa, przerwij obliczenia. Jeżeli nie , to …
4. oblicz $\delta\vec{x}^{(1)}$ i wyznacz $\vec{x}^{(2)} = \vec{x}^{(1)} + \delta\vec{x}^{(1)}$,
5. oblicz $\vec{r}^{(2)} = \vec{b} - \mathbf{A}\vec{x}^{(2)}$ i przejdź ponownie do punktu 3

# Przepis praktyczny

- jeżeli macierz układu jest źle uwarunkowana, może się zdarzyć, że metoda ta nie doprowadzi do rozwiązania bliższego dokładnemu
- wtedy należy jest spróbować liczyć wszystkie wielkości w podwójnej precyzji
- w pozostałych przypadkach metoda pozwala na wyznaczenie rozwiązania, którego wektor reszt jest rzędu $u\|\vec{b}\|_\infty$

# Metody iteracyjne

- przybliżone metody rozwiązywania układów równań
- startują z pewnego przybliżenia początkowego, które jest stopniowo ulepszane aż do uzyskania dostatecznie dokładnego rozwiązania
- najczęściej stosowane do dużych układów rzadkich, tzn. takich, których macierze zawierają w większości zera

# Pojęcia podstawowe

## Definicja

Promieniem spektralnym $\rho(\mathbf{A})$ macierzy $\mathbf{A}$ nazywamy liczbę

$$\rho(\mathbf{A}) = \max_{i=1,\ldots,n} |\lambda_i|,$$

przy czym $\lambda_i$ są wartościami własnymi macierzy $\mathbf{A}$.
Dla dowolnej normy macierzowej zgodnej z normą wektorów obowiązuje

$$|\lambda_i| \leqslant \|\mathbf{A}\|, \quad \text{dla każdego } i = 1, \ldots, n$$

Zatem

$$\rho(\mathbf{A}) \leqslant \|\mathbf{A}\|_p, \quad p = 1, 2, \infty, E$$

# Pojęcia podstawowe

Rozważmy ciąg wektorów $\vec{x}^{(0)}, \vec{x}^{(1)}, \ldots, \vec{x}^{(i)}$, określony dla dowolnego wektora $\vec{x}^{(0)}$ w następujący sposób:

$$\vec{x}^{(i+1)} = \mathbf{M}\vec{x}^{(i)} + \vec{w}, \quad i = 0, 1, \ldots,$$

gdzie $\mathbf{M}$ jest pewną macierzą kwadratową, a $\vec{w}$ wektorem

## Twierdzenie

Ciąg określony powyższym wzorem przy dowolnym wektorze $x^{(0)}$ jest zbieżny do jedynego punktu granicznego wtedy i tylko wtedy, gdy

$$\rho(\mathbf{M}) < 1$$

# Jak konstruować metody iteracyjne?

Należy tak dobrać macierz **M**, aby

- ciąg

$$\vec{x}^{(i+1)} = \mathbf{M}\vec{x}^{(i)} + \vec{w}, \quad i = 0, 1, \ldots$$

  był zbieżny, tzn. $\rho(\mathbf{M}) < 1$

- spełniony był warunek zgodności

$$\vec{x}_{dok} = \mathbf{M}\vec{x}_{dok} + \vec{w}$$

# Jak konstruować metody iteracyjne?

Teoretycznie wystarczy wziąć dowolną macierz $\mathbf{M}$ o promieniu spektralnym mniejszym od 1, a następnie wyliczyć $\vec{w}$ ze warunku zgodności

$$\vec{w} = (\mathbf{I} - \mathbf{M})\mathbf{A}^{-1}\vec{b}$$

jednak wymagałoby to wyliczenia macierzy $\mathbf{A}^{-1}$

# Jak konstruować metody iteracyjne? - inny sposób

Załóżmy, że

$$\vec{w} = \mathbf{N}\vec{b}, \quad \mathbf{N} \text{ - macierz kwadratowa}$$

Z warunku zgodności mamy

$$\vec{x}_{dok} = \mathbf{M}\vec{x}_{dok} + \vec{w} \;\Rightarrow (\mathbf{A}^{-1} - \mathbf{N} - \mathbf{M}\mathbf{A}^{-1})\vec{b} = \mathrm{o} \;\Rightarrow \mathbf{M} = \mathbf{I} - \mathbf{N}\mathbf{A},$$

co prowadzi do

$$\vec{x}^{(i+1)} = (\mathbf{I} - \mathbf{N}\mathbf{A})\vec{x}^{(i)} + \mathbf{N}\vec{b}$$

# Jak konstruować metody iteracyjne?

- rodzina iteracyjna zbieżna dla

$$\rho(\mathbf{I} - \mathbf{NA}) < 1$$

- przy pewnych szczególnych własnościach macierzy układu **A** stosunkowo proste metody wyboru macierzy **N**

# Kryteria przydatności metody iteracyjnej

- liczba działań niezbędnych do wykonania
- potrzebna pamięć
- wielkość błędów zaokrągleń
- szybkość zmian wektora błędu

$$\vec{e}^{(i)} = x^{(i)} - \vec{x}_{dok}$$

Może się okazać, że mimo spełnionego warunku zbieżności zagadnienie jest na tyle źle uwarunkowane, że osiągnięcie zadowalającej dokładności w rozsądnym czasie jest niemożliwe.

# Przykład

$$\mathbf{M} = \begin{pmatrix} \frac{1}{2} & 1 & & & \\ & \frac{1}{2} & 1 & & \\ & & \frac{1}{2} & \ddots & \\ & & & \ddots & 1 \\ & & & & \frac{1}{2} \end{pmatrix}, \vec{w} = \begin{pmatrix} -\frac{1}{2} \\ -\frac{1}{2} \\ \vdots \\ -\frac{1}{2} \\ \frac{1}{2} \end{pmatrix}$$

Mamy tutaj $\rho(\mathbf{M}) = \frac{1}{2}$, a więc dla dowolnego $\vec{x}^{(0)}$ rodzina iteracyjna dąży do $\vec{x}_{dok} = (1, \ldots, 1)^{\mathrm{T}}$.

# Przykład

Przyjmijmy $\vec{x}^{(0)} = 0$:

$$\|\vec{e}^{(0)}\|_\infty = 1, \quad \|\vec{e}^{(1)}\|_\infty = \frac{3}{2}, \quad \|\vec{e}^{(2)}\|_\infty = \frac{9}{4}, \ldots$$

Wzrost błędu w początkowych krokach iteracji może uniemożliwić numeryczne wyznaczenie rozwiązania.

# Rola błędów zaokrągleń

- w skrajnym przypadku mogą doprowadzić do uzyskania

$$\vec{x}^{(i+1)} = \vec{x}^{(0)}$$

- powstanie ciąg wektorów, który nie jest zbieżny do rozwiązania
- przed taką sytuacją trudno się ustrzec

# Metoda Jacobiego

Zapiszmy macierz układu w postaci

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U},$$

gdzie macierze **L**, **D** i **U** to odpowiednio macierz poddiagonalna, diagonalna i naddiagonalna, np.

$$\begin{pmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 \\ 4 & 0 & 0 \\ 7 & 8 & 0 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 5 & 0 \\ 0 & 0 & 9 \end{pmatrix} + \begin{pmatrix} 0 & 2 & 3 \\ 0 & 0 & 6 \\ 0 & 0 & 0 \end{pmatrix}$$

Jako macierz **N** wybierzemy

$$\mathbf{N} = \mathbf{D}^{-1}$$

# Metoda Jacobiego

Wówczas

$$\begin{aligned} \mathbf{M_J} &= \mathbf{I} - \mathbf{NA} = \mathbf{I} - \mathbf{D}^{-1}\mathbf{A} \\ &= \mathbf{I} - \mathbf{D}^{-1}(\mathbf{L} + \mathbf{D} + \mathbf{U}) = -\mathbf{D}^{-1}(\mathbf{L} + \mathbf{U}) \end{aligned}$$

Wzór Jacobiego na rodzinę iteracyjną wektorów będzie miał postać

$$\mathbf{D}\vec{x}^{(i+1)} = -(\mathbf{L} + \mathbf{U})\vec{x}^{(i)} + \vec{b}, \quad i = 0, 1, 2, \ldots$$

# Metoda Jacobiego

Aby wzór Jacobiego był niezawodny, należy wcześniej (w razie konieczności) tak pozmieniać kolejność równań w układzie $\mathbf{A}\vec{x} = \vec{b}$, aby na diagonali macierzy układu były tylko elementy niezerowe:

1. spośród kolumn z elementem zerowym na diagonali wybieramy tę, w której jest największa liczba zer
2. w kolumnie tej wybieramy element o największym module i tak przestawiamy wiersze, aby znalazł się on na głównej przekątnej; wiersz ustalamy i pomijamy go w dalszych rozważaniach
3. spośród pozostałych kolumn z elementem zerowym na diagonali wybieramy tę o największej liczbie zer i wracamy do punktu 2 aż do usunięcia wszystkich zer z głównej przekątnej

# Przykład

Rozważmy macierz

$$\begin{pmatrix} 0 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 \\ 7 & 3 & 0 & 1 \\ 0 & 5 & 0 & 0 \end{pmatrix}$$

Najwięcej zer znajduje się w kolumnie trzeciej, a element o największym module w tej kolumnie to element $a_{13}$.

# Przykład

Zamieniamy miejscami wiersze 1 i 3, tak, aby element ten znalazł się na diagonali,

$$\begin{pmatrix} 7 & 3 & 0 & 1 \\ 2 & 1 & 0 & 2 \\ 0 & 0 & 1 & 2 \\ 0 & 5 & 0 & 0 \end{pmatrix}$$

# Przykład

Zero na diagonali znajduje się jeszcze w kolumnie czwartej, a element o największym module w niej to $a_{24}$. Zamieniamy więc miejscami wiersze 2 i 4,

$$\begin{pmatrix} 7 & 3 & 0 & 1 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 1 & 2 \\ 2 & 1 & 0 & 2 \end{pmatrix}$$

W ten sposób otrzymaliśmy macierz, dla której można zastosować metodę Jacobiego.

# Metoda Jacobiego - niezawodności ciąg dalszy

- zamiana wierszy w macierzy gwarantuje jedynie, że będzie istniała macierz odwrotna do macierzy **D**
- spełnienie warunku zbieżności metody Jacobiego, tzn.

$$\rho(-\mathbf{D}^{-1}(\mathbf{L}+\mathbf{U})) < 1$$

  nie jest gwarantowane w każdym przypadku
- można jedynie pokazać, że jest tak zawsze, jeżeli macierz **A** jest silnie diagonalnie dominująca lub silnie diagonalnie dominująca kolumnowo

# Metoda Gaussa–Seidla

Rozkładamy macierz układu na sumę macierzy poddiagonalnej, diagonalnej i naddiagonalnej (w razie konieczności odpowiednio przestawiając wiersze)

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U}$$

Przyjmujemy

$$\mathbf{N} = (\mathbf{D} + \mathbf{L})^{-1}$$

co prowadzi do

$$\mathbf{M}_{GS} = -(\mathbf{D} + \mathbf{L})^{-1}\mathbf{U}$$

# Metoda Gaussa–Seidla

Stąd

$$\mathbf{D}\vec{x}^{(i+1)} = -\mathbf{L}\vec{x}^{(i+1)} - \mathbf{U}\vec{x}^{(i)} + b, \quad i = 0, 1, 2, \ldots$$

- na pierwszy rzut oka powyższe równanie wygląda tak, jakby niewiadome występowały po obu stronach jednocześnie
- jednak przy obliczaniu pierwszej współrzędnej szukanego wektora po prawej stronie równania nie wystąpi żadna współrzędna wektora $\vec{x}^{(i+1)}$
- przy obliczaniu $x_2^{(i+1)}$ prawa strona równości będzie zależała tylko od $x_1^{(i+1)}$
- ogólnie, przy obliczaniu kolejnej składowej szukanego wektora będziemy korzystali z wyznaczonych już poprzednio składowych

# Niezawodność metody Gaussa–Seidla

- odpowiednie przestawienie wierszy nie gwarantuje w ogólnym przypadku spełnienia warunku zbieżności

$$\rho(\mathbf{M}_{GS}) = \rho(-(\mathbf{D}+\mathbf{L})^{-1}\mathbf{U}) < 1$$

- jeżeli potrafimy uzasadnić, że dla danej macierzy $\mathbf{A}$ metoda Jacobiego jest zbieżna oraz macierz $\mathbf{M}_J$ ma nieujemne elementy, to zbieżna jest również metoda Gaussa–Seidla
- zachodzi przy tym

$$\rho(\mathbf{M}_{GS}) < \rho(\mathbf{M}_J) < 1$$

# Niezawodność metody Gaussa–Seidla

- zbieżność metody jest gwarantowana, jeśli macierz **A** układu równań jest:
  - symetryczna, dodatnio określona
  - silnie diagonalnie dominująca
  - silnie diagonalnie dominująca kolumnowo

# Analiza błędów zaokrągleń

Jeżeli w każdej iteracji zamiast wartości $\mathbf{M}\vec{x}^{(i)} + \vec{w}$ obliczamy

$$\mathbf{M}\vec{x}^{(i)} + \vec{w} + \vec{\delta}^{(i)}, \quad \delta^{(i)} \text{ - błąd zaokrągleń}$$

to

$$\vec{x}^{(i+1)} = \mathbf{M}^{i+1}\vec{x}^{(0)} + \mathbf{M}^{i}\vec{w} + \ldots + \vec{w} + \vec{\delta}^{(i)} + \mathbf{M}\vec{\delta}(i-1) + \ldots + \mathbf{M}^{i}\vec{\delta}^{(0)}$$

Łączny błąd zaokrągleń wynosi

$$\vec{x}_{dok}^{(i+1)} - \vec{x}^{(i+1)} = \vec{\delta}^{(i)} + \mathbf{M}\vec{\delta}^{(i-1)} + \ldots + \mathbf{M}^{i}\vec{\delta}^{(0)}.$$

Jeżeli algorytm iteracyjny jest zbieżny i indeks iteracji jest dostatecznie duży, możemy przyjąć

$$\frac{1}{2}\|\vec{x}_{dok}\| < \|\vec{x}^{(j)}\| < 2\|\vec{x}_{dok}\|$$

# Analiza błędów zaokrągleń

Stąd wynika, że jeżeli $\vec{x}_{dok} \neq 0$, to

$$\frac{\|\vec{x}_{dok}^{(i+1)} - \vec{x}^{(i+1)}\|}{\|\vec{x}^{(i+1)}\|} \leqslant \frac{2}{\|\vec{x}_{dok}\|}(\|\vec{\delta}^{(i)}\| + \|\mathbf{M}\| \cdot \|\vec{\delta}^{(i-1)}\| + \ldots + \|\mathbf{M}\|^i \cdot \|\vec{\delta}^{(0)}\|)$$

czyli

$$\frac{\|\vec{x}_{dok}^{(i+1)} - \vec{x}^{(i+1)}\|}{\|\vec{x}^{(i+1)}\|} \leqslant \frac{2\kappa}{\|\vec{x}_{dok}\|}(1 + \|\mathbf{M}\| + \ldots + \|\mathbf{M}\|^i)$$

gdzie $\kappa$ to wspólne oszacowanie błędów $\vec{\delta}^{(j)}$, tzn.

$$\|\vec{\delta}^{(j)}\| < \kappa, \quad j = 0, 1, 2, \ldots, i$$

# Analiza błędów zaokrągleń

Jeśli $\|\mathbf{M}\| < 1$, to

$$\frac{\|\vec{x}_{dok}^{(i+1)} - \vec{x}^{(i+1)}\|}{\|\vec{x}^{(i+1)}\|} < \frac{1}{1 - \|\mathbf{M}\|} \frac{2\kappa}{\|\vec{x}_{dok}\|}$$

# Analiza błędów zaokrągleń

Gdy macierz układu jest macierzą silnie diagonalnie dominującą, można pokazać, że

$$\frac{\|\vec{x}_{dok}^{(i+1)} - \vec{x}^{(i+1)}\|_\infty}{\|\vec{x}^{(i+1)}\|_\infty} \leqslant \frac{1}{1 - \|\mathbf{M}_{GS}\|_\infty} \frac{12\alpha}{1 - \alpha}$$

gdzie

$$\alpha = \epsilon O(2n^2) \|\mathbf{D}\|_\infty \|\mathbf{D}^{-1}\|_\infty$$

$\Rightarrow$ z porównania powyższego oszacowania z błędem eliminacji Gaussa wynika, że stosując metodę Gaussa–Seidla można zyskać na dokładności, jeżeli tylko wskaźnik $\|\mathbf{D}\|_\infty \|\mathbf{D}^{-1}\|_\infty$ jest mały w porównaniu ze wskaźnikiem uwarunkowania $K_\infty$

# Nakłady obliczeń

- w każdej iteracji wykonujemy około $n^2$ mnożeń (jeżeli macierz układu nie jest rzadka)
- dla porównania, metody dokładne wymagają około $\frac{1}{3}n^3$ mnożeń do uzyskania rozwiązania
- aby metody dokładne i iteracyjne były porównywalne pod względem nakładu obliczeń, powinniśmy wykonać tylko około $n$ iteracji
- proste przykłady pokazują, że liczba iteracji musi być dużo większa niż $n$, aby dokładność była zadowalająca

Metody iteracyjne w przypadku ogólnym są nieefektywne!

# Przykład

Układ

$$\begin{pmatrix} 1 & \frac{3}{4} \\ \frac{3}{4} & 1 \end{pmatrix} \vec{x} = \begin{pmatrix} 448 \\ 448 \end{pmatrix}$$

ma rozwiązanie $x = (256, 256)^{\mathrm{T}}$. Stosując np. eliminację Gaussa, musimy wykonać 6 mnożeń, aby otrzymać wynik dokładny. Jeżeli zastosujemy metodę Gaussa–Seidla, po ośmiu iteracjach (32 mnożenia) mamy

$$\|x^{(8)} - \vec{x}_{dok}\| > 0,1$$

# Warunki przerwania obliczeń

- niezbędną do uzyskania zaplanowanej dokładności liczbę iteracji trudno jest przewidzieć
- w praktyce nie zakłada się konkretnej liczby iteracji z góry
- zamiast tego stosuje się testy na przerwanie obliczeń (tzw. testy stopu):

$$\|\vec{x}^{(i+1)} - \vec{x}^{(i)}\| < \Delta$$

$$\frac{1}{\|\vec{b}\|}\|\mathbf{A}\vec{x}^{(i+1)} - \vec{b}\| < \Delta$$

gdzie $\Delta$ to żądana dokładność

# „Niedoskonałości” testów stopu

- jeżeli norma macierzy **A** jest mała, to wartość reszty

$$\|\mathbf{A}\vec{x}^{(i+1)} - \vec{b}\| = \|\mathbf{A}(\vec{x}^{(i+1)} - \vec{x}_{dok})\| \leqslant \|\mathbf{A}\| \cdot \|\vec{x}^{(i+1)} - \vec{x}_{dok}\|$$

  może być mała, mimo dużego odchylenia wektora $\vec{x}^{(i+1)}$ od rozwiązania dokładnego $\vec{x}_{dok}$
- ponieważ

$$\|\vec{x}^{(i+1)} - \vec{x}^{(i)}\| = \|\vec{e}^{(i+1)} - \vec{e}^{(i)}\| = \|\mathbf{M}\vec{e}^{(i)} - \vec{e}^{(i)}\| = \|(\mathbf{M} - \mathbf{I})\vec{e}^{(i)}\|$$

  gdy norma macierzy $\mathbf{M} - \mathbf{I}$ jest mała, wektory $\vec{x}^{(i+1)}$ i $\vec{x}^{(i)}$ mogą się mało różnić, mimo że błąd $\vec{e}^{(i)}$ jest duży
- testy mogą się okazać mało przydatne z powodu błędów zaokrągleń (wektory reszt należy zawsze liczyć z dużą dokładnością)

# Niedookreślone układy równań ($m < n$)

- liczba równań $m$ jest mniejsza od liczby niewiadomych $n$
- dość często spotykane w praktyce (np. w zagadnieniach optymalizacji)
- nie są one często dyskutowane w literaturze poświęconej metodom numerycznym
- nigdy nie są rozwiązywalne jednoznacznie
  - jeżeli wektor wyrazów wolnych $\vec{b}$ należy do przestrzeni rozpinanej przez kolumny macierzy **A**, wówczas układ

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

    będzie miał nieskończenie wiele rozwiązań
  - w przeciwnym wypadku rozwiązań nie będzie wcale

# Niedookreślone układy równań ($m < n$)

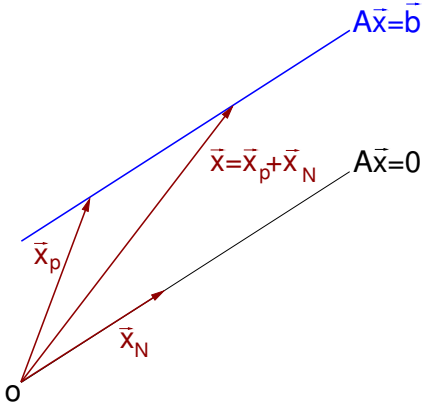
- jeżeli rząd macierzy **A** jest równy liczbie równań, wówczas $\vec{b}$ zawsze będzie należał do przestrzeni rozpinanej przez **A** (układ będzie rozwiązywalny)
- ogólne rozwiązanie takiego układu zapisze się w postaci:

$$\vec{x} = \vec{x}_p + \vec{x}_N$$

gdzie $\vec{x}_p$ jest specjalnym rozwiązaniem równania, a $\vec{x}_N$ należy do jądra przekształcenia liniowego **A**

$$\mathbf{A}\vec{x}_N = \mathrm{o}$$

# Graficzna interpretacja rozwiązania dla $m = 1$ i $n = 2$



# Rozwiązanie szczególne $\vec{x}_p$

## Twierdzenie

Jeżeli macierz $\mathbf{A} \in \mathbf{R}^{m \times n}$ ma rząd $m$, układ $\mathbf{A}\vec{x} = \vec{b}$ jest zawsze rozwiązywalny. Dla każdego $\vec{b}$ istnieje wówczas nieskończenie wiele rozwiązań, z których

$$\vec{x}_p = \mathbf{A}^{\mathrm{T}} \left( \mathbf{A}\mathbf{A}^{\mathrm{T}} \right)^{-1} \vec{b}$$

jest tym o najmniejszej normie. Macierz $\mathbf{A}^{\mathrm{T}} \left( \mathbf{A}\mathbf{A}^{\mathrm{T}} \right)^{-1}$ nazywana jest przy tym macierzą pseudoodwrotną macierzy $\mathbf{A}$.

# Rozwiązanie szczególne $\vec{x}_p$

**Dowód.**
Dla każdego $\vec{x}$ zachodzi

$$\vec{x}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}}\vec{x} = \left(\mathbf{A}^{\mathrm{T}}\vec{x}\right)^{\mathrm{T}}\left(\mathbf{A}^{\mathrm{T}}\vec{x}\right) = \|\mathbf{A}^{\mathrm{T}}\vec{x}\|^2 \geqslant 0.$$

Ponadto, jeśli $\mathrm{rank}\mathbf{A} = m$, to $\|\mathbf{A}^{\mathrm{T}}\vec{x}\| = 0$ wtedy i tylko wtedy, gdy $\vec{x} = 0$. Czyli macierz $\mathbf{A}\mathbf{A}^{\mathrm{T}}$ jest dodatnio określona i nieosobliwa. Ponieważ

$$\mathbf{A}\vec{x}_p = \mathbf{A}\mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1}\vec{b} = \vec{b},$$

więc $x_p$ rzeczywiście jest rozwiązaniem równania $\mathbf{A}\vec{x} = \vec{b}$. Pozostaje nam pokazać, że każde inne rozwiązanie ma normę większą od $\|\vec{x}_p\|$. □

# Rozwiązanie szczególne $\vec{x}_p$

Niech $\vec{x}$ będzie innym rozwiązaniem naszego układu. Wówczas

$$\|\vec{x}\|^2 = \|\vec{x}_p + (\vec{x} - \vec{x}_p)\|^2 = \|\vec{x}_p\|^2 + \|\vec{x} - \vec{x}_p\|^2 + 2\vec{x}_p^{\mathrm{T}}(\vec{x} - \vec{x}_p).$$

Ponieważ z założenia $\mathbf{A}\vec{x}_p = \mathbf{A}\vec{x}$, trzeci wyraz w powyższym równaniu jest równy zero:

$$\vec{x}_p^{\mathrm{T}}(\vec{x} - \vec{x}_p) = \left[\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\vec{b}\right]^{\mathrm{T}}(\vec{x} - \vec{x}_p) = \vec{b}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}(\vec{x} - \vec{x}_p) = 0.$$

Stąd

$$\|\vec{x}\|^2 = \|\vec{x}_p\|^2 + \|\vec{x} - \vec{x}_p\|^2 \geqslant \|\vec{x}_p\|^2,$$

przy czym równość zachodzi tylko dla $\vec{x} = \vec{x}_p$. □

# Układy niedookreślone - przykład

Rozważmy układ

$$\begin{pmatrix} 1 & 2 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 3$$

czyli

$$x_1 + 2x_2 = 3, \quad x_2 = -\frac{1}{2}x_1 + \frac{3}{2}$$

Dowolne z tych dwóch wyrażeń jest rozwiązaniem układu. Rozwiązaniem o najmniejszej normie będzie

$$\vec{x}_p = \mathbf{A}^{\mathrm{T}} \left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1} \vec{b} = \begin{pmatrix} 0,6 \\ 1,2 \end{pmatrix}$$

# Układy niedookreślone - przykład

Wektory należące do jądra przekształcenia **A** będą miały postać

$$\mathbf{A}\vec{x}_N = 0 \quad \rightarrow \quad x_{N2} = -\frac{1}{2}x_{N1}$$

więc ogólne rozwiązanie jest następujące:

$$\vec{x} = \begin{pmatrix} 0,6 \\ 1,2 \end{pmatrix} + \alpha \begin{pmatrix} 1 \\ -0,5 \end{pmatrix}$$

gdzie $\alpha$ jest dowolną liczbą rzeczywistą.

# Macierz pseudoodwrotna i rozkład Cholesky'ego

Ponieważ $\mathbf{A}^{\mathrm{T}}\mathbf{A}$ jest macierzą symetryczną i dodatnio określoną, możemy rozłożyć ją na iloczyn dwóch macierzy trójkątnych

$$\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathrm{T}}$$

Teraz wystarczy rozwiązać układy równań

$$\begin{aligned} \mathbf{L}\vec{w} &= \vec{b} \\ \mathbf{L}^{\mathrm{T}}\vec{z} &= \vec{w} \end{aligned}$$

i na tej podstawie wyliczyć $\vec{x}_p$

$$\vec{x}_p = \mathbf{A}^{\mathrm{T}}\vec{z}$$

# Macierz pseudoodwrotna i rozkład SVD

Z równości

$$\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}$$

wynika

$$\begin{aligned}
\vec{x}_p &= \mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1}\vec{b} = \mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\left(\mathbf{U}\Sigma\mathbf{V}^{\mathrm{T}}\mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\right)^{-1}\vec{b} \\
&= \mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\left(\mathbf{U}\Sigma\Sigma\mathbf{U}^{\mathrm{T}}\right)^{-1}\vec{b} = \mathbf{V}\Sigma\mathbf{U}^{\mathrm{T}}\left(\mathbf{U}^{\mathrm{T}}\right)^{-1}\Sigma^{-1}\Sigma^{-1}\mathbf{U}^{-1}\vec{b} \\
&= \mathbf{V}\Sigma^{-1}\mathbf{U}^{\mathrm{T}}\vec{b}
\end{aligned}$$

# Macierz pseudoodwrotna i rozkład SVD

Zapisując rozkład SVD w postaci

$$\mathbf{A} = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^{\mathrm{T}}, \quad r = \mathrm{rank}\mathbf{A},$$

gdzie $\vec{u}_i$ i $\vec{v}_i$ to kolumny macierzy **U** i **V**, otrzymamy

$$\vec{x}_p = \sum_{i=1}^{r} \frac{\vec{u}_i^{\mathrm{T}} \vec{b}}{\sigma_i} \vec{v}_i.$$

# Macierz pseudoodwrotna i rozkład QR

Jeżeli dysponujemy rozkładem QR macierzy $\mathbf{A}^{\mathrm{T}}$

$$\mathbf{A}^{\mathrm{T}} = \mathbf{QR},$$

wówczas

$$\begin{aligned} \vec{x}_p &= \mathbf{A}^{\mathrm{T}} \left(\mathbf{A}\mathbf{A}^{\mathrm{T}}\right)^{-1} \vec{b} = \mathbf{QR} \left(\mathbf{R}^{\mathrm{T}}\mathbf{Q}^{\mathrm{T}}\mathbf{QR}\right)^{-1} \vec{b} \\ &= \mathbf{QR} \left(\mathbf{R}^{\mathrm{T}}\mathbf{R}\right)^{-1} \vec{b} = \mathbf{QRR}^{-1} \left(\mathbf{R}^{\mathrm{T}}\right)^{-1} \vec{b} \\ &= \mathbf{Q} \left(\mathbf{R}^{\mathrm{T}}\right)^{-1} \vec{b} \end{aligned}$$

Aby wyliczyć $\vec{x}_p$, musimy wyznaczyć $\left(\mathbf{R}^{\mathrm{T}}\right)^{-1} \vec{b}$. Ale to nic innego, jak rozwiązanie równania trójkątnego

$$\mathbf{R}^{\mathrm{T}}\vec{z} = \vec{b}$$

# Nadokreślone układy równań ($m > n$)

- równań ($m$) jest więcej niż niewiadomych ($n$)
- w zależności od wektora wyrazów wolnych nie ma rozwiązań, jest ich nieskończona liczba lub tylko jedno rozwiązanie jednoznaczne
- w praktyce najczęściej dokładne rozwiązanie układu nie istnieje, ale możliwe jest na ogół znalezienie rozwiązania przybliżonego (np. regresja liniowa)

# Nadokreślone układy równań ($m > n$)

- równania $\mathbf{A}\vec{x} = \vec{b}$ dla macierzy $\mathbf{A} \in \mathbf{R}^{m \times n}$ przy $m > n$ nie można rozwiązać uniwersalnie, ponieważ rząd tej macierzy jest mniejszy od $m$
- rozwiązanie dokładne nie istnieje w ogóle, gdy wektor $\vec{b}$ nie należy do przestrzeni rozpinanej przez kolumny macierzy układu
- w tym przypadku zadany układ można potraktować jak zadanie aproksymacyjne i poszukać takiego $\vec{x}$, który zminimalizuje kwadrat normy wektora błędu

$$\vec{e} = \mathbf{A}\vec{x} - \vec{b}.$$

- takie przybliżone rozwiązanie może okazać się bardzo użyteczne w wielu praktycznych zagadnieniach
- to nic innego jak metoda najmniejszych kwadratów

# Rozwiązanie układu nadokreślonego

Szukamy minimum wyrażenia

$$J = \frac{1}{2}\|\vec{e}\|_2^2 = \frac{1}{2}\|\mathbf{A}\vec{x} - \vec{b}\|_2^2 = \frac{1}{2}\left(\mathbf{A}\vec{x} - \vec{b}\right)^{\mathrm{T}}\left(\mathbf{A}\vec{x} - \vec{b}\right)$$

Z warunku na istnienie minimum,

$$\frac{\partial}{\partial \vec{x}} J = \mathbf{A}^{\mathrm{T}}\left(\mathbf{A}\vec{x} - \vec{b}\right) = \mathrm{o}$$

znajdziemy

$$\vec{x}_p = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\vec{b}$$

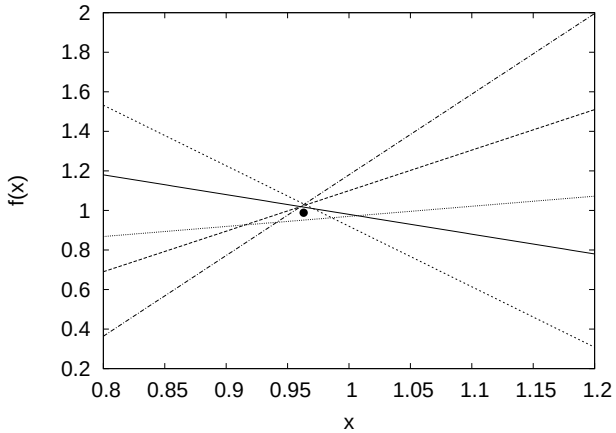
- do obliczenia macierzy pseudoodwrotnej do macierzy **A** możemy znowu wykorzystać rozkłady SVD lub QR macierzy **A**

# Układ nadokreślony - przykład

Rozważmy układ

$$\begin{aligned} x + y &= 1,98 \\ 2,05 * x - y &= 0,95 \\ 3,06 * x + y &= 3,98 \\ -1,02 * x + 2 * y &= 0,92 \\ 4,08 * x - y &= 2,90 \end{aligned}$$

# Układ nadokreślony - przykład



# Układ nadokreślony - przykład

- rozwiązanie ma prostą interpretację geometryczną - to punkt przecięcia prostych zdefiniowanych poszczególnymi równaniami
- dokładne rozwiązanie układu nie istnieje
- rozwiązanie przybliżone wynosi

$$\vec{x}_p = \begin{pmatrix} 0,963101 \\ 0,988543 \end{pmatrix}$$

- błąd przybliżenia

$$\|\mathbf{A}\vec{x}_p - \vec{b}\|_2 = 0,10636$$

# Metody numeryczne

## Wykład 4 - Równania nieliniowe

Janusz Szwabiński

# Plan wykładu

1. Równania z jedną niewiadomą

2. Równania algebraiczne

3. Układy równań nieliniowych

# Równania nieliniowe

- szukamy $x$, dla którego

$$f(x) = 0$$

- trudniejsze niż rozwiązanie układu równań liniowych
- rozwiązanie analityczne albo nie istnieje (równania przestępne, równania algebraiczne rzędu wyższego niż 4), albo jest tak skomplikowane, że zupełnie nie nadaje się do użycia w praktycznych obliczeniach
- iteracyjne poprawianie początkowego przybliżenia szukanego pierwiastka
- przybliżone rozwiązania wystarczają w większości przypadków

# Twierdzenie o punkcie stałym

## Twierdzenie

*Niech $g(x)$ i jej pochodna $g'(x)$ będą funkcjami ciągłymi na pewnym przedziale $I = [\tilde{x} - r, \tilde{x} + r]$ wokół punktu $\tilde{x}$ takiego, że*

$$g(\tilde{x}) = \tilde{x}.$$

*Wówczas, jeżeli*

$$|g'(x)| \leqslant \alpha < 1,$$

*gdzie $\alpha$ to pewna liczba dodatnia, to iteracja*

$$x_{k+1} = g(x_k)$$

*startująca z dowolnego $x_0 \in I$ dąży do punktu stałego $\tilde{x}$ przekształcenia $g$.*

# Twierdzenie o punkcie stałym

- praktyczny przepis na znalezienie przybliżonego rozwiązania równania, o ile tylko da się ono zapisać w postaci

$$x = g(x)$$

- trudność polega na tym, że istnieje zwykle kilka różnych możliwości przekształcenia równania
- w myśl twierdzenia należy wybrać postać, dla której

$$|g'(x)| < 1, \quad x \in I$$

- bez znajomości zgrubnego oszacowania rozwiązania określenie przedziału $I$ może okazać się niemożliwe

# Twierdzenie o punkcie stałym - przykład

Rozważmy równanie

$$f(x) = x^2 - 2 = 0$$

„Zgadujemy", że rozwiązanie powinno leżeć w przedziale $I = (1; 1,5)$ i przekształcamy równanie do postaci

$$x = \frac{2}{x} \quad \Rightarrow \quad g(x) = 2/x$$

Po wyliczeniu pochodnej funkcji $g(x)$ okaże się, że warunek

$$|g'(x)| = \frac{2}{x^2} < 1$$

nie jest spełniony dla wszystkich $x \in I$.

# Twierdzenie o punkcie stałym - przykład

W tej sytuacji procedura iteracyjna

$$x_{k+1} = \frac{2}{x_k}$$

raczej nie zadziała.

# Twierdzenie o punkcie stałym - przykład

I rzeczywiście, już po kilku iteracjach widać, że otrzymaliśmy naprzemienny ciąg wartości

$$x_0 = 1, \quad x_1 = 2, \quad x_2 = 1, \quad x_3 = 2, \ldots$$

który nigdy nie osiągnie poszukiwanego rozwiązania.

# Twierdzenie o punkcie stałym - przykład

Równanie

$$f(x) = x^2 - 2 = 0$$

możemy również zapisać w postaci

$$x = -\frac{1}{2}\{(x-1)^2 - 3\}$$

W tym wypadku funkcja $g(x)$ spełnia warunek zbieżności

$$|g'(x)| = |x - 1| \leqslant 0.5 < 1, \quad \forall x \epsilon I$$

Można więc użyć iteracji

$$x_{k+1} = -\frac{1}{2}\{(x_k - 1)^2 - 3\}$$

do znalezienia rozwiązania.

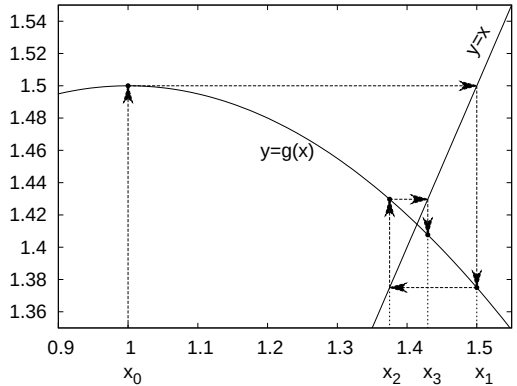
# Twierdzenie o punkcie stałym - przykład

Szereg iteracyjny

$$x_0 = 1; \quad x_1 = 1,5; \quad x_2 = 1,375; \quad x_3 = 1,4297; \quad x_4 = 1,4077, \ldots$$

rzeczywiście dąży do rozwiązania $\sqrt{2} = 1,414\ldots$.

# Twierdzenie o punkcie stałym - przykład



# Twierdzenie o punkcie stałym - przykład

Przekształćmy równanie

$$f(x) = x^2 - 2 = 0$$

do postaci

$$x = \frac{1}{2}\left(x + \frac{2}{x}\right)$$

Pochodna funkcji $g(x)$ spełnia warunek zbieżności

$$|g'(x)| = \frac{1}{2}|1 - \frac{2}{x^2}| \leqslant \frac{1}{2} < 1, \quad \forall x \epsilon I$$

Dodatkowo $g'(x) = 0$ dla $x^2 = 2$ stanowiącego rozwiązanie równania. W tym przypadku szereg iteracyjny zbiega szczególnie szybko do punktu stałego:

$$x_0 = 1; \quad x_1 = 1,5; \quad x_2 = 1,4167; \quad x_3 = 1,4142; \quad x_4 = 1,4142; \ldots$$

```python
import numpy as np

def fixedpoint(g, x0, tol=1e-6, maxit=100):
    xx = np.zeros(maxit)
    xx[0] = x0
    for k in range(1, maxit):
        xx[k] = g(xx[k - 1])
        err = abs(xx[k] - xx[k - 1])
        if err < tol:
            break

    x = xx[k]
    if k == maxit - 1:
        print("No real convergence!")  #

    return x, err, xx[:k+1]
```

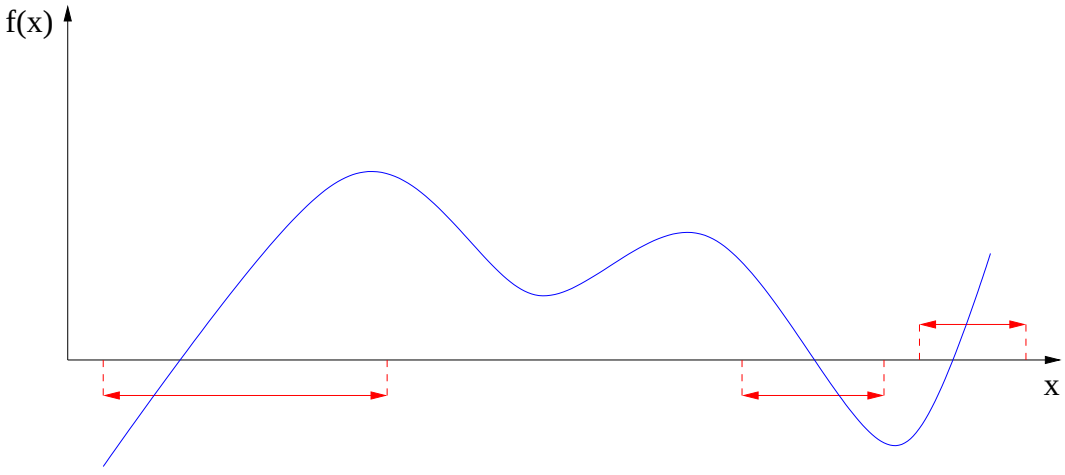
```
def fun(x):
    return 0.5*(x+2/x)

result, error, values = fixedpoint(fun, 0.5)
print("Result:", result)
print("Error:", error)
print("Values:", values)
```

```
Result: 1.414213562373095
Error: 1.5183720947220536e-10
Values: [0.5 2.25 1.56944444 1.42189036 1.41423429 1.41421356  1.41421356]
```

# Lokalizacja miejsc zerowych

- wybór wartości startowej (lub przedziału) odgrywa dużą rolę w rozwiązywaniu równań
- źle wybrany punkt startowy może spowodować, że metoda iteracyjna w ogóle nie będzie zbieżna lub znajdzie „złe” rozwiązanie
- nawet niezbyt dokładny wykres pozwala wybrać rozsądne przybliżenie początkowe
- jeżeli metoda wymaga od nas przedziału, w którym znajduje się rozwiązanie, a nie tylko wartości początkowej, powinniśmy wybrać tzw. przedział izolacji pierwiastka
- wiele metod zawodzi, kiedy podaje im się na starcie przedział zawierający więcej pierwiastków

# Lokalizacja miejsc zerowych



# Lokalizacja miejsc zerowych

- wykres funkcji jako metoda lokalizacji rozwiązań sprawdza się znakomicie przy rozwiązaniu jednego (lub kilku równań), o ile tylko stanowi to cel sam w sobie
- czasochłonne, gdy mamy wiele różnych równań do rozwiązania
- niepraktyczne, gdy rozwiązanie równania nieliniowego stanowi tylko krok pośredni obliczeń komputerowych

# Automatyczne oddzielanie pierwiastków (ang. *incremental search*)

- jeżeli $f(a)f(b) < 0$, to ciągła funkcja $f(x)$ musi mieć w przedziale $(a, b)$ przynajmniej jeden pierwiastek
- jeżeli dodatkowo przedział $(a, b)$ będzie mały, istnieje duże prawdopodobieństwo, że będzie on przedziałem izolacji danego pierwiastka
- wystarczy zbadać zmiany znaku w ciągu wartości funkcji wyliczonych dla dyskretnego zbioru punktów

$$x_i = x_0 + i\Delta x, \quad n = 0, 1, 2, 3, \ldots$$

odległych od siebie o pewien niewielki krok $\Delta$

# Automatyczne oddzielanie pierwiastków - przykład

$$x^2 - 2 = 0$$

| $x$ | $f(x)$ |
|---|---|
| 0.00000 | -2.000000 |
| 0.20000 | -1.960000 |
| 0.40000 | -1.840000 |
| 0.60000 | -1.640000 |
| 0.80000 | -1.360000 |
| 1.00000 | -1.000000 |
| 1.20000 | -0.560000 |
| 1.40000 | -0.040000 |
| 1.60000 | 0.560000 |
| 1.80000 | 1.240000 |
| 2.00000 | 2.000000 |

# Automatyczne oddzielanie pierwiastków - implementacja

```
from numpy import sign
def rootsearch(f,a,b,dx):
    x1 = a; f1 = f(a)
    x2 = a + dx; f2 = f(x2)
    while sign(f1) == sign(f2):
        if x1 >= b: return None,None
        x1 = x2; f1 = f2
        x2 = x1 + dx; f2 = f(x2)
    else:
        return x1,x2
```

# Automatyczne oddzielanie pierwiastków - implementacja

Szukamy pierwiastka funkcji

$$f(x) = x^3 - 10x^2 + 5$$

z dokładnością do 4 cyfr dziesiętnych

- w podejściu naiwnym $dx = 0.0001 \rightarrow 10000$ wyliczeń funkcji
- możemy lokalizować pierwiastek w 4 etapach, poprawiając za każdym razem dokładność $\rightarrow 40$ wyliczeń funkcji

# Automatyczne oddzielanie pierwiastków - implementacja

```
def f(x): return x**3 - 10.0*x**2 + 5.0

x1 = 0.0; x2 = 1.0
for i in range(4):
    dx = (x2 - x1)/10.0
    x1,x2 = rootsearch(f,x1,x2,dx)
x = (x1 + x2)/2.0
print('x =', '{:6.4f}'.format(x))

x = 0.7346
```

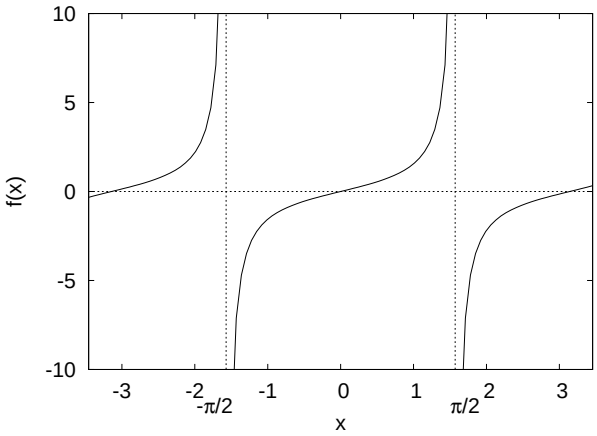
# Automatyczne oddzielanie pierwiastków - ograniczenia

- jeśli krok $\Delta x$ jest większy, niż odległość między dwoma sąsiednimi pierwiastkami, możemy je przeoczyć
- pierwiastek o parzystej krotności nie zostanie znaleziony
- niektóre osobliwości mogą zostać potraktowane jako pierwiastki

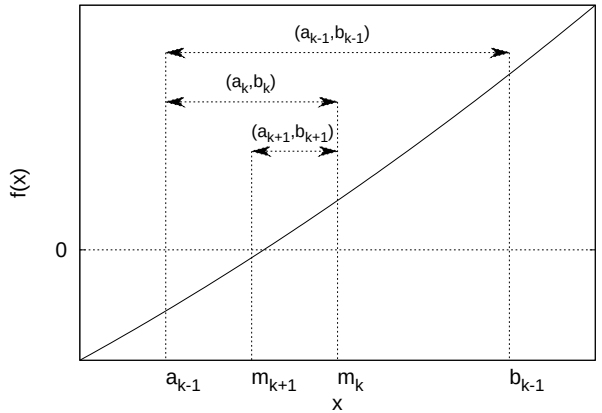
# Automatyczne oddzielanie pierwiastków - ograniczenia

```
import math
def f(x): return math.tan(x)

x1,x2 = rootsearch(f,1,2,0.02)

print((x1+x2)/2)

1.5700000000000005
```

# Automatyczne oddzielanie pierwiastków - ograniczenia



# Metoda połowienia przedziału (bisekcji)



# Metoda połowienia przedziału

- jeżeli w przedziale $(a, b)$ znajduje się miejsce zerowe ciągłej funkcji $f(x)$, to $f(a)f(b) < 0$
- dla pierwiastka $\alpha \in (a_1, b_1)$ generujemy ciąg przedziałów

$$(a_1, b_1) \supset (a_2, b_2) \supset (a_3, b_3) \supset \ldots, \quad \forall i \quad \alpha \in (a_i, b_i)$$

- dla $I_{k-1} = (a_{k-1}, b_{k-1})$ kolejny przedział wyznaczamy według przepisu:
  1. obliczamy środek $m_k$ przedziału $I_{k-1}$

$$m_k = \frac{1}{2}(a_{k-1} + b_{k-1})$$

  2. jeśli $f(m_k) = 0$, znaleźliśmy pierwiastek; w przeciwnym razie

$$(a_k, b_k) = \begin{cases} (m_k, b_{k-1}), & \text{jeśli } f(m_k)f(b_{k-1}) < 0 \\ (a_{k-1}, m_k), & \text{jeśli } f(a_{k-1})f(m_k) < 0 \end{cases}$$

# Metoda połowienia przedziału - błędy zaokrągleń

Używamy arytmetyki z sześcioma liczbami dziesiętnymi. Niech

$$a_{k-1} = 0,742531, \quad b_{k-1} = 0,742533$$

Stąd (po zaokrągleniu)

$$a_{k-1} + b_{k-1} = 1.48506, \quad \frac{1}{2}(a_{k-1} + b_{k-1}) = 0,742530$$

$$\Rightarrow \quad \frac{1}{2}(a_{k-1} + b_{k-1}) < a_{k-1}$$

# Metoda połowienia przedziału - błędy zaokrągleń

- w obliczeniach ze skończoną dokładnością w arytmetyce dziesiętnej nierówności
$$a_{k-1} \leqslant \frac{1}{2}(a_{k-1} + b_{k-1}) \leqslant b_{k-1}$$
mogą nie być spełnione dla wszystkich liczb zmiennoprzecinkowych $a_{k-1}$ i $b_{k-1}$!
- aby zagwarantować spełnienie nierówności w arytmetyce o dowolnej podstawie, wystarczy wzór
$$m_k = a_{k-1} + \frac{1}{2}(b_{k-1} - a_{k-1})$$

# Metoda połowienia przedziału

- po $n$ krokach otrzymamy przedział o długości

$$\frac{1}{2^n}(b-a)$$

- jako wartość przybliżoną pierwiastka przyjmujemy

$$\alpha = m_{n+1} \pm d_n, \quad d_n = \frac{1}{2^{n+1}}(b-a)$$

- wadą jest wolna zbieżność
- w każdym kroku iteracji zyskujemy jedną dokładną cyfrę dwójkową
- ponieważ $10^{-1} \simeq 2^{-3{,}3}$, jedną cyfrę dziesiętną uzyskamy średnio co 3,3 kroków

# Metoda połowienia przedziału - przykład

$$x^2 - 2 = 0,\ I_0 = (1; 1,5)$$

| n | $a_{n-1}$ | $b_{n-1}$ | $m_n$ | $f(m_n)$ |
|---|---|---|---|---|
| 1 | 1,0000 | 1,5000 | 1,2500 | -0,43750000 |
| 2 | 1,2500 | 1,5000 | 1,3750 | -0,10938000 |
| 3 | 1,3750 | 1,5000 | 1,4375 | 0,06640600 |
| 4 | 1,3750 | 1,4375 | 1,4062 | -0,02260200 |
| 5 | 1,4062 | 1,4375 | 1,4219 | 0,02180000 |
| 6 | 1,4062 | 1,4219 | 1,4141 | -0,00032119 |
| 7 | 1,4141 | 1,4219 | 1,4180 | 0,01072400 |
| 8 | 1,4141 | 1,4180 | 1,4160 | 0,00505600 |
| 9 | 1,4141 | 1,4160 | 1,4150 | 0,00222500 |
| 10 | 1,4141 | 1,4150 | 1,4146 | 0,00109320 |
| 11 | 1,1441 | 1,4146 | 1,4143 | 0,00024449 |
| 12 | 1,4141 | 1,4143 | 1,4142 | 0,00003836 |

```
import math, sys
from numpy import sign

def bisection(f,x1,x2,switch=1,tol=1.0e-9):
    f1 = f(x1)
    if f1 == 0.0: return x1
    f2 = f(x2)
    if f2 == 0.0: return x2
    if sign(f1) == sign(f2):
        print('Wrong interval!')
        sys.exit(1)
    n = int(math.ceil(math.log(abs(x2 - x1)/tol)/math.log(2.0)))

    for i in range(n):
        x3 = 0.5*(x1 + x2); f3 = f(x3)
        if (switch == 1) and (abs(f3) > abs(f1)) and (abs(f3) > abs(f2)):
            return None
        if f3 == 0.0: return x3
        if sign(f2)!= sign(f3): x1 = x3; f1 = f3
        else: x2 = x3; f2 = f3

    return (x1 + x2)/2.0
```

```
def f(x): return x**3 - 10.0*x**2 + 5.0

x = bisection(f, 0.0, 1.0, tol = 1.0e-4)
print('x =', '{:6.4f}'.format(x))
```

```
x = 0.7346
```

# Metoda wielopodziału przedziału izolacji

- uogólnienie metody bisekcji
- w jednym kroku dzielimy przedział na $k$ podprzedziałów $I_i = [x_i, x_{i+1}]$

$$x_i = a + i\left(\frac{b-a}{k}\right), \quad i = 0, 1, 2, \ldots, k$$

- wybór nowego przedziału izolacji odbywa się jak poprzednio
- aby znaleźć pierwiastek z dokładnością $\epsilon$ musimy wykonać

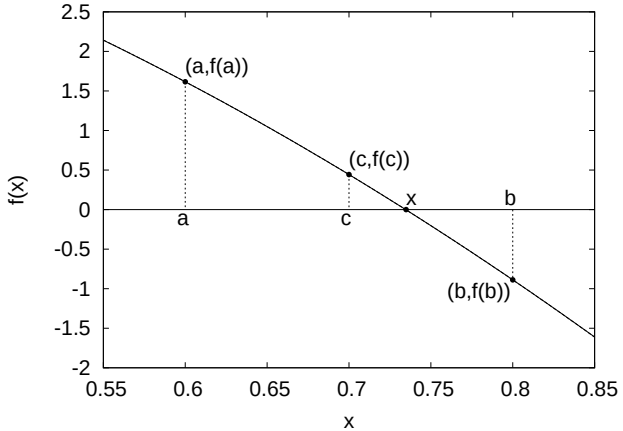
$$n_k = \frac{\log_2\left(\frac{b-a}{2\epsilon}\right)}{\log_2 k} \text{ podziałów}$$

- przydatna, gdy w przedziale początkowym jest kilka pierwiastków i mamy możliwość równoległego operowania na większej liczbie podprzedziałów

# Metoda Brenta

- łączy w sobie niezawodność bisekcji z odwrotną interpolacją kwadratową
- dzielimy wyjściowy przedział $(a, b)$ izolacji pierwiastka na połowę
- określamy, w którym z przedziałów $\left(a, c = \frac{a+b}{2}\right)$ i $(c, b)$ leży poszukiwany pierwiastek
- przez punkty $(a, f(a))$, $(c, f(c))$ i $(b, f(b))$ prowadzimy parabolę i szukamy punktu przecięcia z osią $X$

# Metoda Brenta



# Metoda Brenta

- wzór paraboli przechodzącej przez trzy punkty

$$x = \frac{[y - f(b)][y - f(c)]}{[f(a) - f(b)][f(a) - f(c)]}a + \frac{[y - f(a)][y - f(c)]}{[f(b) - f(a)][f(b) - f(c)}b + \frac{[y - f(a)][y - f(b)]}{[f(c) - f(a)][f(c) - f(b)]}c$$

- kładąc $y = 0$ znajdziemy nowe przybliżenie poszukiwanego pierwiastka

$$x = -\frac{af(b)f(c)[f(b) - f(c)] + bf(c)f(a)[f(c) - f(a)] + cf(a)f(b)[f(a) - f(b)]}{[f(a) - f(b)][f(b) - f(c)][f(c) - f(a)]}$$

- przybliżenie przyjmujemy, jeśli leży ono w nowym przedziale izolacji
- w przeciwnym razie wynik interpolacji porzucamy i przeprowadzamy następny krok bisekcji
- procedurę powtarzamy do uzyskania żądanej dokładności

# Metoda Brenta - przykład

Szukamy pierwiastka funkcji

$$f(x) = x^3 - 10x^2 + 5$$

znajdującego się początkowo w przedziale $(0,6; 0,8)$.
W punktach startowych mamy

$$\begin{aligned} a &= 0,6, \quad f(a) = 1,616 \\ b &= 0,8, \quad f(b) = -0,888 \end{aligned}$$

Połowimy przedział izolacji pierwiastka:

$$c = 0,7, \quad f(c) = 0,443$$

Stąd wynika, że nowym przedziałem izolacji pierwiastka jest $(c, b) = (0,7; 0,8)$.

# Metoda Brenta - przykład

Przez punkty $(a, f(a))$, $(c, f(c))$ i $(b, f(b))$ prowadzimy parabolę. Znajdujemy punkt przecięcia z osią $X$:

$$x = 0,73487$$

Ponieważ leży on w nowym przedziale izolacji pierwiastka, akceptujemy wynik. W ten sposób pierwszy krok metody Brenta został ukończony. W drugim kroku wartości $c$, $x$ i $b$ będziemy traktować jako nowe wartości $a$, $c$ i $b$

$$\begin{aligned} c = 0,7 &\rightarrow a \\ x = 0,73487 &\rightarrow c \\ b = 0,8 &\rightarrow b \end{aligned}$$

Nowym przedziałem izolacji będzie $(a, c) = (0,7; 0,73487)$.

# Metoda Brenta - przykład

Interpolacja kwadratowa prowadzi do

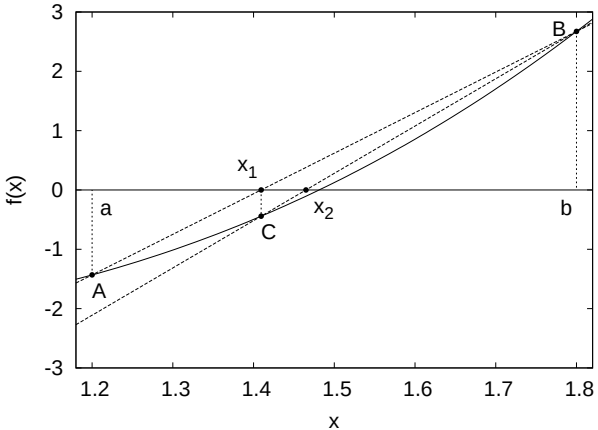
$$x = 0,73460$$

Ponownie akceptujemy wynik, ponieważ leży on w przedziale izolacji. W ten sposób, po dwóch krokach, otrzymaliśmy rozwiązanie z pięcioma poprawnymi cyframi dziesiętnymi.

# Regula falsi

- zakładamy, że równanie $f(x) = 0$ ma w przedziale $(a, b)$ pojedynczy pierwiastek $\alpha$
- funkcja $f(x)$ jest klasy $C^2$ na przedziale $\langle a, b \rangle$
- jej pierwsza i druga pochodna mają stały znak na tym przedziale
- dla ustalenia uwagi rozważymy przypadek $f'(x) > 0$ i $f''(x) > 0$ dla $x \in \langle a, b \rangle$ (wybór stałego punktu iteracji)

# Regula falsi



# Regula falsi

- przez punkty $A = (a, f(a))$ i $B = (b, f(b))$ prowadzimy cięciwę

$$y - f(a) = \frac{f(b) - f(a)}{b - a}(x - a)$$

- punkt przecięcia cięciwy z osią $X$ to pierwsze przybliżenie pierwiastka:

$$-f(a) = \frac{f(b) - f(a)}{b - a}(x_1 - a) \quad \Rightarrow x_1 = a - \frac{f(a)}{f(b) - f(a)}(b - a)$$

- jeśli przybliżenie nie jest wystarczające, przez punkt $C = (x_1, f(x_1))$ oraz jeden z punktów $A$ i $B$ (wybieramy ten, w którym funkcja jest przeciwnego znaku niż w $C$) prowadzimy następną cięciwę itd.

# Regula falsi

- w ten sposób otrzymamy ciąg kolejnych przybliżeń

$$\begin{aligned} x_0 &= a \\ x_{k+1} &= x_k - \frac{f(x_k)}{f(b) - f(x_k)}(b - x_k), \quad k = 1, 2, \ldots \end{aligned}$$

- powyższy ciąg jest rosnący i ograniczony z góry $\Rightarrow$ jest zbieżny
- można przejść w powyższym równaniu do granicy $k \to \infty$

$$g = g - \frac{f(g)}{f(b) - f(g)}(b - g)$$

$$g = \lim_{k \to \infty} x_k, \quad a < g < b$$

$$\Rightarrow \quad f(g) = 0$$

# Regula falsi

Korzystając z twierdzenia Lagrange'a o przyrostach,

$$f(x_n) - f(\alpha) = f'(c)(x_n - \alpha), \;\; x_n < c < \alpha$$

możemy oszacować błąd $n$-tego przybliżenia ($f(\alpha) = 0$)

$$|x_n - \alpha| \leqslant \frac{f(x_n)}{m}, \;\; m = \inf_{x \in \langle a,b \rangle} |f'(x)|$$

# Regula falsi

Błąd możemy ocenić również znając dwa kolejne przybliżenia:

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

Ponieważ $f(\alpha) = 0$, więc

$$f(\alpha) - f(x_k) = \frac{f(x_k) - f(b)}{x_k - b}(x_{k+1} - x_k)$$

# Regula falsi

Z twierdzenia Lagrange'a otrzymujemy

$$(\alpha - x_k)f'(\xi_k) = (x_{k+1} - x_k)f'(\bar{x}_k), \quad \xi_k \in (x_k, \alpha), \quad \bar{x}_k \in (x_k, b)$$

Dodajemy obustronnie $-x_{k+1}f'(\xi_k)$

$$|\alpha - x_{k+1}| = \frac{|x_k - x_{k+1}| \cdot |f'(\xi_k) - f'(\bar{x}_k)|}{|f'(\xi_k)|} \leqslant \frac{M - m}{m}|x_{k+1} - x_k|,$$

gdzie

$$m = \inf_{x \in \langle a,b \rangle} |f'(x)|, \quad M = \sup_{x \in \langle a,b \rangle} |f'(x)|$$

# Regula falsi

- oszacowanie pesymistyczne
- wymaga znajomości $m$ i $M$
- dla przybliżeń w niewielkim otoczeniu $\alpha$:

$$|\alpha - x_{k+1}| \sim \left|\frac{f(x_{k+1})}{f'(x_{k+1})}\right| \sim \left|\frac{x_{k+1} - x_k}{f(x_{k+1}) - f(x_k)}\right| \cdot |f(x_{k+1})|$$

# Regula falsi

- metoda zbieżna dla dowolnej funkcji ciągłej na przedziale $\langle a, b \rangle$, o ile tylko spełniony jest warunek $f(a)f(b) < 0$ i pierwsza pochodna tej funkcji jest ograniczona i różna od zera w otoczeniu pierwiastka
- jeżeli druga pochodna nie zmienia znaku w rozpatrywanym przedziale, to punktem stałym iteracji jest punkt, w którym

$$ff'' > 0$$

- stosunkowo wolno zbieżna

# Regula falsi - przykład

Szukamy pierwiastka równania

$$x^3 + x^2 - 3x - 3 = 0$$

Z wykresu funkcji $f(x) = x^3 + x^2 - 3x - 3$ wynika, że pierwiastek dodatni leży w przedziale $(1, 2)$. Ponadto,

$$\begin{aligned} f'(x) &= 3x^2 + 2x - 3 \\ f''(x) &= 6x + 2 \end{aligned}$$

zatem obie pochodne są dodatnie w tym przedziale.

# Regula falsi - przykład

| $x$ | $f(x)$ |
|---|---|
| $a = 1$ | -4 |
| $b = 2$ | 3 |
| $x_1 = 1,57142$ | -1,36449 |
| $x_2 = 1,70540$ | -0,24784 |
| $x_3 = 1,72788$ | -0,03936 |
| $x_4 = 1,73140$ | -0,00615 |

```
def regula(func, a, b, delta, epsilon, maxit):
    fa = func(a); fb = func(b)
    if fa * fb > 0: raise ValueError("Wrong interval!")

    for k in range(maxit):
        dx = fb*(b-a)/(fb-fa)
        x = b-dx; ac = x-a; fx = func(x)

        if fx == 0:  break
        elif fb * fx > 0:
            b = x; fb = fx
        else:
            a = x; fa = fx

        dx = min(abs(dx), abs(ac))

        if abs(dx) < delta: break
        if abs(fx) < epsilon: break

    err = abs(b - a) / 2
    return x, err, fx
```

```
def fun(x):
    return x**3 + x**2 -3*x-3

regula(fun,1,2,10e-6, 10e-6,100)

(1.7320504374844243, 0.13397478125778783, -3.5025160194379623e-06)
```

# Metoda siecznych

- metodę regula falsi możemy ulepszyć rezygnując z założenia, aby funkcja miała w punktach wytyczających następną cięciwę różne znaki
- kolejne przybliżenie pierwiastka wyznaczamy z poprzednich:

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

- na ogół szybsza niż regula falsi
- zdarzają się przypadki, dla których nie jest ona zbieżna (np. gdy początkowe przybliżenia nie leżą dostatecznie blisko pierwiastka)
- gdy różnica $(x_k - x_{k-1})$ jest rzędu oszacowania błędu, następne przybliżenie nie do przyjęcia

# Metoda siecznych - przykład

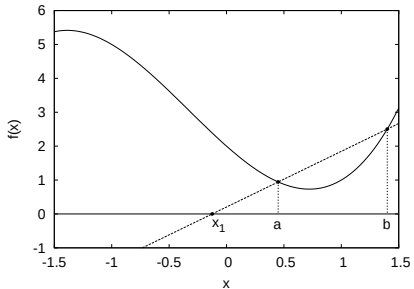
$$x^3 + x^2 - 3x - 3 = 0$$

| $x$ | $f(x)$ |
|---|---|
| $a = 1$ | -4 |
| $b = 2$ | 3 |
| $x_1 = 1,57142$ | -1,36449 |
| $x_2 = 1,70540$ | -0,24784 |
| $x_3 = 1,73513$ | 0,02920 |
| $x_4 = 1,73199$ | 0,000576 |

# Metoda siecznych

## Uwaga!

W pierwszym kroku warunek $f(a)f(b) < 0$ ciągle jest potrzebny!



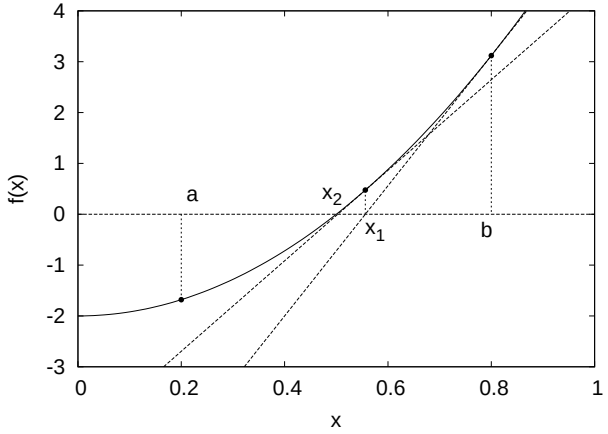
# Metoda Newtona (stycznych)

- $f(x)$ jest ciągła na przedziale $\langle a, b \rangle$
- $f(a)f(b) < 0$
- $f'(x)$ i $f''(x)$ mają stały znak w rozpatrywanym przedziale
- dzięki wykorzystaniu dodatkowej informacji zawartej w pochodnych funkcji zbieżność na ogół lepsza niż omówionych już algorytmów

# Metoda Newtona

- od końca przedziału, w którym funkcja $f(x)$ ma ten sam znak co jej druga pochodna, prowadzimy styczną do wykresu funkcji
- punkt $x_1$ przecięcia stycznej z osią $X$ pierwszym przybliżeniem pierwiastka
- z punktu $(x_1, f(x_1))$ prowadzimy następną styczną i określamy kolejne przybliżenie
- powtarzamy aż do uzyskania żądanej dokładności

# Metoda Newtona



# Metoda Newtona

Z równania stycznej

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

Stąd łatwo otrzymać iteracyjny wzór Newtona na kolejne przybliżenie pierwiastka równania $f(x) = 0$:

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

# Metoda Newtona

Niech $f'(x) > 0$ i $f''(x) > 0$ dla $\forall x \in \langle a, b \rangle$.
Ze wzoru Newtona mamy

$$x_1 = b - \frac{f(b)}{f'(b)}$$

Z rozwinięcia w szereg Taylora otrzymujemy

$$f(\alpha) = f(b) + f'(b)(\alpha - b) + \frac{1}{2} f''(c)(\alpha - b)^2$$

gdzie $c \in \langle a, b \rangle$

# Metoda Newtona

Jeśli $\alpha$ ma być pierwiastkiem równania, to

$$\alpha = b - \frac{f(b)}{f'(b)} - \frac{1}{2}\frac{f''(c)}{f'(b)}(\alpha - b)^2$$

czyli

$$\alpha - x_1 = -\frac{1}{2}\frac{f''(c)}{f'(b)}(\alpha - b)^2 < 0$$

(bo pochodne są dodatnie)

# Metoda Newtona

Stąd wynika

$$x_1 > \alpha$$

Ponieważ zachodzi również $x_1 - b < 0$, więc

$$x_1 < b$$

- $x_1$ leży bliżej szukanego pierwiastka, niż punkt startowy $b$
- kontynuując, można pokazać, że ciąg przybliżeń jest malejącym ciągiem ograniczonym z dołu poprzez $\alpha \rightarrow$ jest ciągiem zbieżnym (do pewnej liczby $g$)
- przechodząc obustronnie do granicy $n \rightarrow \infty$, mamy

$$g = g - \frac{f(g)}{f'(g)} \quad \Rightarrow \quad f(g) = 0 \quad \Rightarrow \quad g = \alpha$$

# Metoda Newtona

Błąd $n$–tego przybliżenia szacujemy z twierdzenia Lagrange'a o przyrostach

$$|x_n - \alpha| \leqslant \left| \frac{f(x_n)}{m} \right|, \quad m = \inf_{x \in \langle a,b \rangle} |f'(x)|$$

Ze wzoru Taylora wynika

$$\begin{aligned} f(x_n) &\equiv f(x_{n-1} + (x_n - x_{n-1})) \\ &= f(x_{n-1}) + f'(x_{n-1})(x_n - x_{n-1}) + \frac{1}{2} f''(\xi_{n-1})(x_n - x_{n-1})^2 \end{aligned}$$

co po uwzględnieniu wzoru Newtona prowadzi do

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

# Metoda Newtona

Stąd wynika

$$|f(x_n)| \leqslant \frac{1}{2}M(x_n - x_{n-1})^2, \quad M = \sup_{x\in\langle a,b\rangle} |f''(x)|$$

czyli również

$$|\alpha - x_n| \leqslant \frac{1}{2}\frac{M}{m}(x_n - x_{n-1})^2 = \frac{M}{2m}\left[\frac{f(x_n)}{f'(x_n)}\right]^2$$

Podobnie, jak w przypadku metody regula falsi, dla $x_n$ dostatecznie bliskich $\alpha$ możemy korzystać z oszacowania

$$|\alpha - x_n| \approx \left|\frac{f(x_{n-1})}{f'(x_{n-1})}\right|$$

# Metoda Newtona - przykład

$$x^3 + x^2 - 3x - 3 = 0$$

| $x$ | $f(x)$ | $x$ | $f(x)$ |
|---|---|---|---|
| $x_0 = 2$ | 3 | $x_0 = 1$ | -4 |
| $x_1 = 1,76923$ | 0,36048 | $x_1 = 3$ | 24 |
| $x_2 = 1,73292$ | 0,00823 | $x_2 = 2,2$ | 5,88800 |
| $x_3 = 1,73205$ | -0,000008 | $x_3 = 1,83015$ | 0,98899 |
| | | $x_4 = 1,7578$ | 0,24782 |
| | | $x_5 = 1,73195$ | -0,00095 |

# Metoda Newtona

## Twierdzenie

*Jeżeli mamy przedział $\langle a, b\rangle$ taki, że:*

i. *$f(a)$ i $f(b)$ mają przeciwne znaki,*

ii. *$f''$ jest ciągła i nie zmienia znaku na $\langle a, b\rangle$,*

iii. *styczne do krzywej $y = f(x)$ poprowadzone w punktach o odciętych a i b przecinają oś OX wewnątrz przedziału $\langle a, b\rangle$,*

*wówczas równanie $f(x) = 0$ ma dokładnie jeden pierwiastek $\alpha$ w przedziale $\langle a, b\rangle$ i metoda Newtona jest zbieżna do $\alpha$ dla dowolnego punktu startowego $x_0 \in \langle a, b\rangle$.*

# Metoda Newtona - przykład

Metodę Newtona można zastosować np. do obliczania pierwiastka kwadratowego z liczby dodatniej $c$. Jest on bowiem rozwiązaniem równania

$$x^2 - c = 0$$

Na podstawie wzoru Newtona otrzymujemy (dla $x_n \neq 0$):

$$x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right)$$

Warunki twierdzenia są spełnione na przedziale $\langle a, b\rangle$ takim, że

$$0 < a < c^{1/2}, \quad b > \frac{1}{2}\left(a + \frac{c}{a}\right)$$

# Metoda Newtona - implementacja

```
def newton(f,Df,x0,epsilon,maxit):
    xn = x0
    for n in range(0,maxit):
        fxn = f(xn)
        if abs(fxn) < epsilon:
            print('Found solution after',n,'iterations.')
            return xn
        Dfxn = Df(xn)
        if Dfxn == 0:
            print('Zero derivative. No solution found.')
            return None
        xn = xn - fxn/Dfxn
    print('Exceeded maximum iterations. No solution found.')
    return None
```

# Metoda iteracyjna Eulera

Niech $x_n$ będzie aktualnym przybliżeniem poszukiwanego pierwiastka

$$f(x_n + h) = 0$$

gdzie $h$ jest pewną małą liczbą. Rozwijając funkcję $f(x)$ w szereg Taylora wokół punktu $x_n$ i pomijając wyrazy rzędu wyższego niż drugi otrzymamy

$$f(x_n) + hf'(x_n) + \frac{h^2}{2}f''(x_n) = 0, \quad h = x - x_n$$

Jeśli $[f'(x_n)]^2 \geqslant 2f(x_n)f''(x_n)$, to powyższy trójmian kwadratowy ma pierwiastki rzeczywiste

$$h_n = -\frac{f'(x_n)}{f''(x_n)}\left(1 \pm \sqrt{1 - 2\frac{f(x_n)f''(x_n)}{[f'(x_n)]^2}}\right)$$

# Metoda iteracyjna Eulera

Biorąc mniejszy z nich, znajdziemy

$$x_{n+1} = x_n - u(x_n)\frac{2}{1+\sqrt{1-2t(x_n)}}$$

$$u(x) = \frac{f(x)}{f'(x)}, \quad t(x) = \frac{f(x)f''(x)}{[f'(x)]^2}$$

# Metoda iteracyjna Eulera

W przypadku $|t| \ll 1$ możemy jeszcze skorzystać z przybliżenia

$$\frac{2}{1-\sqrt{2t(x_n)}} \simeq 1 + \frac{1}{2}t(x_n)$$

Wówczas

$$x_{n+1} = x_n - u(x_n)\left(1 + \frac{1}{2}t(x_n)\right)$$

# Rząd metod

## Definicja

Mówimy, że metoda jest rzędu $p$, jeżeli istnieje stała $K$ taka, że dla dwu kolejnych przybliżeń $x_k$ i $x_{k+1}$ zachodzi nierówność

$$|x_{k+1} - \alpha| \leqslant K|x_k - \alpha|^p.$$

# Rząd metod

| Metoda | Rząd |
|---|---|
| bisekcji | 1 |
| regula falsi | 1 |
| siecznych | $\frac{1}{2}(1+\sqrt{5}) \simeq 1,62$ |
| Brenta | $\simeq 1,8$ |
| Newtona | 2 |
| Eulera | 3 |

# Pierwiastki wielokrotne

## Definicja

Liczbę $\alpha$ nazywamy $r$–krotnym pierwiastkiem równania $f(x) = 0$, jeżeli

$$0 < |g(\alpha)| < \infty, \quad g(x) = \frac{f(x)}{(x-\alpha)^r}$$

# Pierwiastki wielokrotne

- metoda bisekcji, regula falsi i metoda siecznych mogą być stosowane do pierwiastków o krotności nieparzystej ($f(a)f(b) < 0$)
- rząd metody siecznych obniża się
- metoda Newtona może być stosowana do wszystkich krotności, jeśli tylko istnieje odpowiednie lewo- lub prawostronne sąsiedztwo szukanego pierwiastka, w którym znak $f'(x)$ i $f''(x)$ pozostaje stały
- rząd metody Newtona obniża się

# Pierwiastki wielokrotne

Jeżeli krotność pierwiastka $r$ jest znana, to metodę Newtona można zmodyfikować w sposób następujący:

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

# Pierwiastki wielokrotne

Jeżeli krotność nie jest znana, postępujemy inaczej. Zakładamy, że $f(x)$ ma $r$–tą pochodną ciągłą w otoczeniu pierwiastka $\alpha$ o krotności $r$. Wówczas

$$f^{(i)}(\alpha) = 0, \quad i < r$$

i ze wzoru Taylora wynika

$$\begin{aligned} f(x) &= \frac{1}{r!}(x-\alpha)^r f^{(r)}(\xi) \\ f'(x) &= \frac{1}{(r-1)!}(x-\alpha)^{r-1} f^{(r)}(\xi') \end{aligned}$$

przy czym $\xi$ i $\xi'$ leżą w przedziale między $x$ i $\alpha$.

# Pierwiastki wielokrotne

Niech

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

Zachodzi

$$\lim_{x \to \alpha} \frac{u(x)}{x - \alpha} = \frac{1}{r},$$

czyli równanie $u(x) = 0$ ma pierwiastek pojedynczy $\alpha$. Równanie $u(x) = 0$ można już rozwiązać wszystkimi omówionymi metodami.

# Pierwiastki wielokrotne

Metoda Newtona daje w tym przypadku

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

natomiast wzór siecznych prowadzi do

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

# Przyspieszanie zbieżności

## Definicja

Dla ciągu $\{x_n\}_{n=0}^{\infty}$ różnicą skończoną w przód nazywamy

$$\Delta x_n = x_{n+1} - x_n, \quad n \geqslant 0.$$

Różnice wyższego rzędu określone są wzorem rekurencyjnym:

$$\Delta^k x_n = \Delta^{k-1}(\Delta x_n), \quad k \geqslant 2.$$

# Przyspieszanie zbieżności

## Twierdzenie

*Niech ciąg* $\{x_n\}_{n=0}^{\infty}$ *zbiega liniowo do granicy* $\alpha$ *i niech* $\alpha - x_n \neq 0$ *dla wszystkich* $n \geqslant 0$. *Jeśli istnieje liczba rzeczywista A taka, że* $|A| < 1$ *i*

$$\lim_{n\to\infty} \frac{\alpha - x_{n+1}}{\alpha - x_n} = A,$$

*wówczas ciąg* $\{y_n\}_{n=0}^{\infty}$ *zdefiniowany jako*

$$y_n = x_n - \frac{(\Delta x_n)^2}{\Delta^2 x_n} = x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}$$

*zbiega do* $\alpha$ *szybciej niż* $\{x_n\}_{n=0}^{\infty}$, *tzn.*

# Przyspieszanie zbieżności

## Twierdzenie
*(c.d.)*

$$\lim_{n\to\infty} \frac{\alpha - y_n}{\alpha - x_n} = 0.$$

# Przyspieszanie zbieżności

- technika ta nazywana jest procesem $\Delta^2$ Aitkena
- w połączeniu z metodą Newtona otrzymujemy algorytm Steffensena
- stosuje się go do poprawienia zbieżności metody Newtona w przypadku pierwiastków wielokrotnych

# Równania algebraiczne

$$w(z) \equiv a_0 z^n + a_1 z^{n-1} + \ldots + a_{n-1} z + a_n = 0$$

- dokładnie $n$ pierwiastków
- pierwiastki mogą być rzeczywiste lub zespolone
- jeśli współczynniki $a_0, a_1, \ldots, a_n$ są rzeczywiste, to ewentualne pierwiastki zespolone występują w parach sprzężonych
- pierwiastków rzeczywistych wielomianów z rzeczywistymi współczynnikami możemy szukać opisanymi wcześniej metodami
- jeśli zależy nam na znalezieniu pierwiastków zespolonych, lepiej używać metod dedykowanych wielomianom

# Metoda Laguerre'a

- bardzo popularna
- pozwala wyliczyć pierwiastki rzeczywiste i zespolone, pojedyncze i wielokrotne
- wymaga arytmetyki zespolonej

# Metoda Laguerre'a

Aby znaleźć pierwiastek wielomianu $w_n(z)$ stopnia $n$, przybliżamy go wielomianem

$$r(z) = a(z - p_1)(z - p_2)^{n-1}$$

Parametry $a$, $p_1$ i $p_2$ dobieramy tak, aby

$$w_n(z_k) = r(z_k), \quad w_n'(z_k) = r'(z_k), \quad w_n''(z_k) = r''(z_k)$$

Jeśli $z_k$ jest pewnym przybliżeniem pojedynczego pierwiastka $\alpha$, wówczas parametr $p_1$ będzie jego następnym oszacowaniem.

# Metoda Laguerre'a

Zauważmy, że dla

$$w(z) = (z - \alpha_1)(z - \alpha_2) \dots (z - \alpha_n)$$

zachodzi

$$S_1 \equiv \frac{w'(z)}{w(z)} = \sum_{i=1}^{n} \frac{1}{z - \alpha_i}$$

Różniczkując obustronnie otrzymamy

$$-\frac{\mathrm{d}S_1(z)}{\mathrm{d}z} \equiv S_2(z) = \left(\frac{w'(z)}{w(z)}\right)^2 - \frac{w''(z)}{w(z)} = \sum_{i=1}^{n} \frac{1}{(z - \alpha_i)^2}$$

# Metoda Laguerre'a

Stąd

$$\begin{aligned} S_1(z_k) &= \frac{1}{z_k - p_1} + \frac{(n-1)}{z_k - p_2} \\ S_2(z_k) &= \frac{1}{(z_k - p_1)^2} + \frac{(n-1)}{(z_k - p_2)^2} \end{aligned}$$

czyli ($p_1 = z_{k+1}$)

$$z_{k+1} = z_k - \frac{nw(z_k)}{w'(z_k) \pm \sqrt{H(z_k)}}$$

$$H(z_k) = (n-1)^2 \left[w'(z_k)\right]^2 - n(n-1)w(z_k)w''(z_k)$$

# Metoda Laguerre'a

- znak we wzorze iteracyjnym wybieramy tak, aby poprawka $|z_{k+1} - z_k|$ była jak najmniejsza
- dla wielomianów, które mają tylko pierwiastki rzeczywiste, metoda Laguerre'a jest globalnie zbieżna
- w przypadku wielomianów z pierwiastkami zespolonymi zbieżności dla dowolnego punktu startowego nie można zagwarantować, ale w praktyce w większości przypadków metoda działa dobrze
- w szczególności dla punktu startowego $z_0 = 0$ metoda pozwoli zazwyczaj znaleźć pierwiastek o najmniejszym module
- rząd metody Laguerre'a wynosi 3 dla pierwiastków pojedynczych i 1 dla wielokrotnych

# Macierz towarzysząca (ang. *companion matrix*)

- wartości własne macierzy **A** to pierwiastki wielomianu charakterystycznego

$$w(x) = \det[\mathbf{A} - x\mathbf{I}]$$

- są lepsze sposoby na numeryczne wyznaczenie wartości własnych
- problem można jednak odwrócić i poszukać takiej macierzy, której wartości własne byłyby pierwiastkami interesującego nas wielomianu

# Macierz towarzysząca

Wystarczy, że weźmiemy

$$\mathbf{A} = \begin{pmatrix} -\frac{a_{n-1}}{a_n} & -\frac{a_{n-2}}{a_n} & \dots & -\frac{a_1}{a_n} & -\frac{a_0}{a_n} \\ 1 & 0 & \dots & 0 & 1 \\ 0 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & 0 \end{pmatrix}$$

# Macierz towarzysząca

Łatwo sprawdzić, że wielomian charakterystyczny macierzy **A** ma postać

$$x^n + \frac{a_{n-1}}{a_n}x^{n-1} + \frac{a_{n-2}}{a_n}x^{n-2} + \cdots + \frac{a_1}{a_n}x + \frac{a_0}{a_n} = 0$$

a więc rzeczywiście jest równoważny interesującemu nas wielomianowi.

# Liczba pierwiastków rzeczywistych

Aby oszacować liczbę pierwiastków rzeczywistych wielomianu $w(x)$ stopnia $n$, tworzymy ciąg

$$w(x), w'(x), w''(x), \ldots, w^{(n)}(x)$$

Oznaczmy przez $M(x_0)$ liczbę zmian znaku w ciągu w punkcie $x_0$.

## Twierdzenie

*(Fouriera) Jeżeli $w(x)$ jest wielomianem stopnia $n$ określonym w przedziale $(a, b)$ oraz $w(a)w(b) \neq 0$, to liczba zer wielomianu w tym przedziale wynosi*

$$M(a) - M(b),$$

*lub jest od tej liczby mniejsza o liczbę parzystą.*

# Liczba pierwiastków rzeczywistych - przykład

Niech

$$w(x) = x^3 - 2x^2 - 5x + 5$$

Tworzymy ciąg pochodnych:

$$\begin{aligned} w'(x) &= 3x^2 - 4x - 5 \\ w''(x) &= 6x - 4 \\ w'''(x) &= 6 \end{aligned}$$

# Liczba pierwiastków rzeczywistych - przykład

| $x$ | $-\infty$ | $\infty$ | 0 | 1 | 3 |
|---|---|---|---|---|---|
| $w$ | $-$ | $+$ | $+$ | $-$ | $+$ |
| $w'$ | $+$ | $+$ | $-$ | $-$ | $+$ |
| $w''$ | $-$ | $+$ | $-$ | $+$ | $+$ |
| $w'''$ | $+$ | $+$ | $+$ | $+$ | $+$ |
| $M(x)$ | 3 | 0 | 2 | 1 | 0 |

# Liczba pierwiastków rzeczywistych - przykład

Ponieważ

$$M(-\infty) - M(\infty) = 3$$

z twierdzenia Fouriera wynika, że równanie $w(x) = 0$ ma jeden lub trzy pierwiastki rzeczywiste. Ponadto

$$\begin{aligned} M(-\infty) - M(0) &= 1 \\ M(0) - M(1) &= 1 \\ M(1) - M(3) &= 1 \end{aligned}$$

czyli wielomian ma trzy pierwiastki, po jednym w każdym z przedziałów $(-\infty, 0)$, $(0, 1)$ i $(1, 3)$.

# Układy równań nieliniowych

Poszukujemy rozwiązania $\vec{\alpha} \in \mathbb{R}^n$ równania

$$F(\vec{x}) = \mathrm{o}$$

$$F : \mathbb{R}^n \to \mathbb{R}^n$$

- gdy odwzorowanie $F$ jest afiniczne, np. $F(\vec{x}) = \mathbf{A}\vec{x} + \vec{b}$, uzyskujemy układ równań liniowych
- jeżeli $F$ jest odwzorowaniem nieliniowym, rozwiązanie układu $F(\vec{x}) = \mathrm{o}$ komplikuje się
- brak ogólnego kryterium istnienia rozwiązania
- będziemy a priori zakładać, że rozwiązanie takie istnieje

# Ogólne metody iteracyjne

Konstruujemy ciąg punktów

$$\vec{x}^{(0)}, \quad \vec{x}^{(1)}, \quad \vec{x}^{(2)}, \quad \ldots$$

zbieżny do rozwiązania $\vec{\alpha}$ układu $F(\vec{x}) = 0$. Jeżeli można wskazać odwzorowanie $G$ takie, że

$$\vec{x}^{(i)} = G(\vec{x}^{(i-1)}, \ldots, \vec{x}^{(i-p)})$$

to metodę iteracyjną nazywamy metodą stacjonarną $p$–punktową. Jeżeli natomiast $G$ ulega modyfikacjom podczas iteracji, to mówimy o metodzie niestacjonarnej (zmiennego operatora).

# Ogólne metody iteracyjne

## Definicja

Niech $G: \ D \subset \mathbb{R}^n \times \ldots \times \mathbb{R}^n \to \mathbb{R}^n$. Punkt $\vec{\alpha}$ nazywamy punktem przyciągania metody iteracyjnej, jeżeli istnieje takie otoczenie $U_{\vec{\alpha}}$ tego punktu, że obierając punkty $\vec{x}^{(-p+1)}, \ldots, \vec{x}^{(0)}$ z tego otoczenia uzyskamy ciąg punktów $\vec{x}^{(1)}, \vec{x}^{(2)}, \ldots \in D$, zbieżny do $\vec{\alpha}$.

# Ogólne metody iteracyjne

## Definicja

Odwzorowanie $G: \mathbb{R}^n \to \mathbb{R}^n$ nazywamy różniczkowalnym w sensie Frécheta w punkcie $\vec{x}$, jeżeli istnieje taka macierz $A: \mathbb{R}^n \to \mathbb{R}^n$, że

$$\lim_{\vec{h} \to o} \frac{\|G(\vec{x} + \vec{h}) - G(\vec{x}) - A\vec{h}\|}{\|\vec{h}\|} = o$$

przy dowolnym sposobie wyboru wektorów $\vec{h} \to o$. Macierz $A$ nazywamy pochodną Frécheta odwzorowania $G$ w punkcie $\vec{x}$ i oznaczamy ją przez $G'(\vec{x})$.

# Ogólne metody iteracyjne

## Twierdzenie

*Jeżeli pochodna Frécheta odwzorowania* $G : \mathbb{R}^n \to \mathbb{R}^n$ *w punkcie* $\vec{\alpha}$ *ma promień spektralny* $\varphi(G'(\vec{\alpha})) = \beta < 1$ *oraz* $G(\vec{\alpha}) = \vec{\alpha}$*, to* $\vec{\alpha}$ *jest punktem przyciągania metody iteracyjnej* $\vec{x}^{(i+1)} = G(\vec{x}^{(i)})$.

- pozwala uzasadnić lokalną zbieżność wielu metod iteracyjnych

# Metoda Newtona

## Twierdzenie

*Niech funkcja $F(\vec{x})$ będzie różniczkowalna w sensie Frécheta w pewnym otoczeniu $K(\vec{\alpha}, r)$ punktu $\vec{\alpha}$, w którym $F(\vec{\alpha}) = 0$. Załóżmy, że $F'(\vec{x})$ jest ciągła w punkcie $\vec{\alpha}$, a $F'(\vec{\alpha})$ jest nieosobliwa. Wówczas $\vec{\alpha}$ jest punktem przyciągania metody Newtona*

$$\vec{x}^{(i+1)} = \vec{x}^{(i)} - [F'(\vec{x}^{(i)})]^{-1} F(\vec{x}^{(i)})$$

# Metoda Newtona

## Twierdzenie

*(c.d.) Ponadto, jeżeli ciągłość pochodnej w punkcie $\vec{\alpha}$ jest typu Höldera*

$$\|F'(\vec{x}) - F'(\vec{\alpha})\| \leqslant H\|\vec{x} - \vec{\alpha}\|^p, \quad \vec{x} \in K(\vec{\alpha}, r), \quad p \in (0, 1\rangle$$

*to*

$$\|\vec{x}^{(i+1)} - \vec{\alpha}\| \leqslant C\|\vec{x}^{(i)} - \vec{\alpha}\|^{1+p},$$

*gdzie* $C = 4H\|[F'(\vec{\alpha})]^{-1}\|$.

# Metoda siecznych

Przybliżamy $F$ odwzorowaniem afinicznym w taki sposób, aby

$$F(\vec{y}^{(j)}) \simeq \mathbf{A}\vec{y}^{(j)} + \vec{b}, \quad j = 0, 1, \ldots, n$$

a następnie przyjmujemy rozwiązanie liniowego układu równań

$$\mathbf{A}\vec{x} + \vec{b} = 0$$

jako przybliżenie poszukiwanego pierwiastka.

# Metoda siecznych

Rozwiązaniem tego układu będzie

$$\vec{x} = -\mathbf{A}^{-1}\vec{b} = -\mathbf{A}^{-1}(F(\vec{y}) - \mathbf{A}\vec{y}) = \vec{y} - \mathbf{A}^{-1}F(\vec{y})$$

Niech $\Delta\mathbf{Y}$ będzie macierzą o kolumnach $\vec{y}^{(1)} - \vec{y}^{(0)}$, $\vec{y}^{(2)} - \vec{y}^{(2)}$, ..., $\vec{y}^{(n)} - \vec{y}^{(n-1)}$, a $\Delta\mathbf{F}$ macierzą o kolumnach $F(\vec{y}^{(1)}) - F(\vec{y}^{(0)})$, $F(\vec{y}^{(2)}) - F(\vec{y}^{(2)})$, ..., $F(\vec{y}^{(n)}) - F(\vec{y}^{(n-1)})$. Wówczas

$$\Delta\mathbf{F} = \mathbf{A}\Delta\mathbf{Y}$$

# Metoda siecznych

1. wybieramy punkty $\vec{x}^{(-n)}, \vec{x}^{(-n+1)}, \ldots, \vec{x}^{(0)}$ i przyjmujemy $i = 0$
2. obliczamy $\Delta\mathbf{F}^{(i)}$ oraz $\Delta\mathbf{Y}^{(i)}$ przyjmując

$$\vec{y}^{(0)} = \vec{x}^{(i-n)}, \quad \vec{y}^{(1)} = \vec{x}^{(i-n+1)}, \quad \ldots, \quad \vec{y}^{(n)} = \vec{x}^{(i)}$$

3. wyznaczamy $\vec{x}^{(i+1)}$ ze wzoru

$$\vec{x}^{(i+1)} = \vec{x}^{(i)} - \Delta\mathbf{Y}^{(i)}[\Delta\mathbf{F}^{(i)}]^{-1}F(\vec{x}^{(i)})$$

4. zwiększamy $i$ o jeden i wracamy do kroku 2

# Metoda siecznych

- metodę można stosować, o ile tylko macierz $\Delta\mathbf{F}^{(i)}$ nie jest osobliwa
- w przeciwnym razie musimy inaczej wybierać punkty $\vec{y}^{(j)}$, np.

$$\begin{aligned} \vec{y}^{(n)} &= \vec{x}^{(i)} \\ \vec{y}^{(n-j)} &= \vec{x}^{(i)} + h\vec{e}^{(j)}, \quad j = 1, 2, \ldots, n \end{aligned}$$

  gdzie $h$ jest pewną stałą, a $\vec{e}^{(j)}$ to wektory przestrzeni $\mathbb{R}^n$
- w niektórych przypadkach niemożliwe takie dobranie punktów $\vec{y}^{(j)}$, aby macierz była nieosobliwa

# Metody numeryczne

## Wykład 5 - Interpolacja

Janusz Szwabiński

# Plan wykładu

1. Zagadnienie interpolacyjne

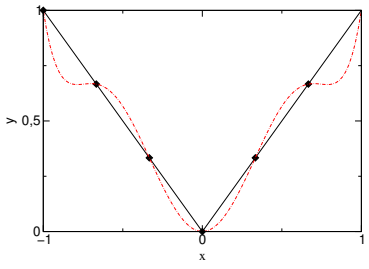
2. Interpolacja wielomianowa

3. Interpolacja wymierna

# Zagadnienie interpolacyjne

$F(x_i) = y_i, \quad i = 0, 1, \ldots, n$

- $x_i$ – węzły interpolacji
- $y_i \equiv f(x_i)$ – wartości funkcji interpolowanej
- $F(x)$ – funkcja interpolująca



# Interpolacja wielomianowa

$\Pi_n$ – zbiór wielomianów stopnia $\leqslant n$

## Twierdzenie

*Dla dowolnych $n+1$ punktów węzłowych $(x_i, y_i)$, $i = 0, 1, \ldots, n$, $x_i \neq x_k$ dla $i \neq k$, istnieje dokładnie jeden wielomian $W_n \in \Pi_n$ taki, że*

$$W_n(x_i) = y_i, \quad i = 0, 1, \ldots, n.$$

$\Rightarrow$ interpolacja wielomianowa jest zagadnieniem jednoznacznym!

# Interpolacja wielomianowa

## Dowód.

Mamy $n+1$ punktów węzłowych $(x_i, y_i)$. Szukamy wielomianu interpolującego w postaci

$$W_n(x) = a_0 + a_1 x + \ldots + a_n x^n, \quad W_n(x_i) = y_i, \quad i = 0, \ldots, n$$

Stąd

$$\begin{aligned} a_0 + a_1 x_0 + \quad & \ldots \quad + a_n x_0^n = y_0 \\ & \vdots \\ a_0 + a_1 x_n + \quad & \ldots \quad + a_n x_n^n = y_n \end{aligned}$$

# Interpolacja wielomianowa

## Dowód.

Macierz układu (niewiadomymi są współczynniki $a_i$!)

$$A = \begin{pmatrix} 1 & x_0 & x_0^2 & \dots & x_0^n \\ 1 & x_1 & x_1^2 & \dots & x_1^n \\ & & \vdots & & \\ 1 & x_n & x_n^2 & \dots & x_n^n \end{pmatrix}$$

$\Rightarrow \det A \neq 0$ (wyznacznik Vandermonde'a z $x_i \neq x_k$ dla $i \neq k$)

$\Rightarrow$ układ równań jest układem Cramera

# Interpolacja wielomianowa

## Dowód.

$\Rightarrow$ istnieje dokładnie jedno rozwiązanie

$$a_i = \frac{\det A_i}{\det A} = \frac{1}{\det A} \sum_{j=0}^{n} y_j A_{ij}$$

gdzie $A_{ij}$ są kolejnymi dopełnieniami algebraicznymi elementów $i$–tej kolumny macierzy $A$

□

# Wzór interpolacyjny Lagrange'a

$$\left.\begin{array}{l} W_n(x) = a_0 + \ldots + a_n x^n \\ a_i = \frac{1}{\det A} \sum_{j=0}^{n} y_j A_{ij} \end{array}\right\} \quad W_n(x) = y_0\Phi_0(x) + \ldots + y_n\Phi_n(x)$$

$$\Phi_i(x) \quad - \quad \text{wielomiany stopnia} \leqslant n$$

Ponieważ

$$W_n(x_i) = y_0\Phi_0(x_i) + y_1\Phi_1(x_i) + \ldots + y_n\Phi_n(x_i) \equiv y_i$$

$$\Rightarrow \Phi_i(x_j) = \begin{cases} 0, & \text{gdy } i \neq j \\ 1, & \text{gdy } i = j \end{cases}$$

# Wzór interpolacyjny Lagrange'a

Niech

$$\Phi_i(x) = \lambda(x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n)$$

Z warunku $\Phi_i(x_i) = 1$ otrzymujemy

$$\Phi_i(x) = \frac{(x - x_0)(x - x_1)\dots(x - x_{i-1})(x - x_{i+1})\dots(x - x_n)}{(x_i - x_0)(x_i - x_1)\dots(x_i - x_{i-1})(x_i - x_{i+1})\dots(x_i - x_n)}$$

# Wzór interpolacyjny Lagrange'a

Stąd

$$W_n(x) = \sum_{j=0}^{n} y_j \frac{\omega_n(x)}{(x - x_j)\left.\frac{\omega_n(x)}{x - x_j}\right|_{x=x_j}} = \sum_{j=0}^{n} y_j \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)}$$

$$\omega_n(x) = (x - x_0)(x - x_1)\dots(x - x_n)$$

Z twierdzenia o jednoznaczności

$\Rightarrow$ wielomian Lagrange'a jest jedynym wielomianem interpolacyjnym stopnia $\leqslant n$

# Wzór interpolacyjny Lagrange'a

## Przykład

Szukamy wielomianu interpolacyjnego, który przechodzi przez następujące punkty węzłowe:

| $x_i$ | -2 | 1 | 2 | 4 |
|---|---|---|---|---|
| $y_i$ | 3 | 1 | -3 | 8 |

$$
\begin{aligned}
W_3(x) &= \frac{3(x-1)(x-2)(x-4)}{(-2-1)(-2-2)(-2-4)} + \frac{(x+2)(x-2)(x-4)}{(1+2)(1-2)(1-4)} \\
&\quad \frac{-3(x+2)(x-1)(x-4)}{(2+2)(2-1)(2-4)} + \frac{8(x+2)(x-1)(x-4)}{(4+2)(4-1)(4-2)} \\
&= \frac{2}{3}x^3 - \frac{3}{2}x^2 - \frac{25}{6}x + 6
\end{aligned}
$$

# Wzór interpolacyjny Lagrange'a

```
def lagrange(x,xData,yData):
    n = len(xData)
    y = 0
    for i in range(n):
        w = 1.0
        for j in range(n):
            if i != j:
                w = w*(x-xData[j])/(xData[i]-xData[j])
        y = y + w*yData[i]
    return y
```

# Metoda Neville'a

$W_{i_0,i_1,\ldots,i_k}(x)$ - wielomian stopnia $k$ przechodzący przez $(x_{i_j}, y_{i_j})$, $j = 0, 1, \ldots, k$

## Twierdzenie

*Wielomian $W_{i_0,i_1,\ldots,i_k}$ daje się przedstawić wzorem rekurencyjnym*

$$W_{i_0,i_1,\ldots,i_k}(x) = \frac{(x - x_{i_0})W_{i_1,\ldots,i_k}(x) - (x - x_{i_k})W_{i_0,\ldots,i_{k-1}}(x)}{x_{i_k} - x_{i_0}}$$

# Metoda Neville'a

## Dowód.

Niech $P(x)$ będzie prawą stroną powyższego równania. Stopień $P(x)$ jest $\leqslant k$. Ponadto, zgodnie z definicją wielomianów $W_{i_1,\ldots,i_k}(x)$ i $W_{i_0,\ldots,i_{k-1}}(x)$ mamy

$$P(x_{i_0}) = W_{i_0,\ldots,i_{k-1}}(x_{i_0}) = y_{i_0}, \quad P(x_{i_k}) = W_{i_1,\ldots,i_k}(x_{i_k}) = y_{i_k} \tag{1}$$

a dla $j = 1, 2, \ldots, k-1$

$$P(x_{i_j}) = \frac{(x_{i_j} - x_{i_0})y_{i_j} - (x_{i_j} - x_{i_k})y_{i_j}}{x_{i_k} - x_{i_0}} = y_{i_j}$$

$P(x)$ ma zatem cechy $W_{i_0,i_1,\ldots,i_k}(x)$ □

# Metoda Neville'a

$$
\begin{array}{llllll}
x_0 & y_0 & & & & \\
 & & W_{0,1}(x) & & & \\
x_1 & y_1 & & W_{0,1,2}(x) & & \\
 & & W_{1,2}(x) & & W_{0,1,2,3}(x) & \\
x_2 & y_2 & & W_{1,2,3}(x) & & \\
 & & W_{2,3}(x) & & & \\
x_3 & y_3 & & & &
\end{array}
$$

# Metoda Neville'a

## Przykład

| $x_i$ | 0 | 1 | 3 |
|---|---|---|---|
| $y_i$ | 1 | 3 | 2 |

Schemat Neville'a dla $W_{0,1,2}(2)$ ma postać:

$$
\begin{array}{cc}
0 & 1 \\
 & \\
1 & 3 \\
 & \\
3 & 2
\end{array}
\qquad
\begin{array}{l}
W_{0,1}(2) = \frac{(2-0)*3-(2-1)*1}{1-0} = 5 \\
\\
W_{1,2}(2) = \frac{(2-1)*2-(2-3)*3}{3-1} = \frac{5}{2}
\end{array}
\qquad
W_{0,1,2}(2) = \frac{10}{3}
$$

# Oszacowanie błędu wzoru interpolacyjnego

## Twierdzenie (Rolle'a)

*Niech:*

1. *funkcja będzie określona na przedziale domkniętym* $< a, b >$*;*
2. *istnieje pochodna skończona* $f'(x)$ *przynajmniej w przedziale otwartym* $(a, b)$*;*
3. *na końcach przedziału funkcja przyjmuje wartości* $f(a) = f(b)$*.*

*Wówczas między a i b można znaleźć taki punkt c* $(a < c < b)$*, że*

$$f'(c) = 0.$$

# Oszacowanie błędu wzoru interpolacyjnego

Niech $f(x)$ będzie funkcją $n+1$ razy różniczkowalną oraz $\epsilon(x) = W_n(x) - f(x)$:

## Twierdzenie

$$|\epsilon(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_n(x)|$$

*gdzie*

$$M_{n+1} = \sup_{x \in <a,b>} |f^{(n+1)}(x)|$$

# Oszacowanie błędu wzoru interpolacyjnego

## Dowód.

Wprowadźmy funkcję pomocniczą ($K$ - pewna stała)

$$\phi(u) = W_n(u) - f(u) + K(u - x_0)(u - x_1)\dots(u - x_n)$$

$$\phi(x_i) = 0, \quad i = 0, 1, \dots, n$$

Współczynnik $K$ dobieramy tak, aby pierwiastkiem funkcji $\phi(u)$ był również punkt $\tilde{x}$, różny od węzłów interpolacji:

$$K = \frac{f(\tilde{x}) - W_n(\tilde{x})}{\omega_n(\tilde{x})} \quad (\omega_n(\tilde{x}) \neq 0 \text{ dla } \tilde{x} \neq x_i)$$

# Oszacowanie błędu wzoru interpolacyjnego

## Dowód.

$\phi(u)$ ma w sumie $n+2$ miejsc zerowych $x_0, x_1, \ldots, x_n, \tilde{x}$.
Twierdzenia Rolle'a:

- $\Rightarrow$ $\phi'(u)$ ma w każdym z podprzedziałów położonych między pierwiastkami co najmniej jedno miejsce zerowe
- $\Rightarrow$ w przedziale $(\min(x, x_0), \max(x, x_n))$ jest co najmniej $n+1$ miejsc zerowych $\phi'(u)$
- $\Rightarrow$ co najmniej $n$ miejsc zerowych drugiej pochodnej $\phi''(u)$

$\vdots$

# Oszacowanie błędu wzoru interpolacyjnego

## Dowód.

$\Rightarrow$ istnieje co najmniej jeden punkt $\xi$ w przedziale $(\min(x, x_0), \max(x, x_n))$ taki, że $\phi^{(n+1)}(\xi) = 0$

Ponieważ

$$\begin{aligned} W_n^{(n+1)}(x) &= 0 \\ \omega_n^{(n+1)}(x) &= (n+1)! \end{aligned}$$

więc

$$\phi^{(n+1)}(u) = -f^{(n+1)}(u) + K(n+1)! \quad \Rightarrow \quad K = \frac{f^{(n+1)}(\xi)}{(n+1)!}$$

# Oszacowanie błędu wzoru interpolacyjnego

## Dowód.

Stąd wynika

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

zatem

$$|\epsilon(x)| \leqslant \frac{M_{n+1}}{(n+1)!}|\omega_n(x)|,$$

gdzie

$$M_{n+1} = \sup_{x \in <a,b>} |f^{(n+1)}(x)|$$

# Wzór interpolacyjny Newtona

- wzór Lagrange'a i metoda Neville'a pozwalają wyznaczyć wartość wielomianu w punkcie

- niepraktyczne, jeśli punktów jest dużo

# Ilorazy różnicowe

## Definicja

Ilorazami różnicowymi pierwszego rzędu nazywamy wyrażenia

$$
\begin{aligned}
f[x_0; x_1] &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} \\
f[x_1; x_2] &= \frac{f(x_2) - f(x_1)}{x_2 - x_1} \\
&\vdots \\
f[x_{n-1}; x_n] &= \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}
\end{aligned}
$$

# Ilorazy różnicowe

## Definicja

Ilorazami różnicowymi drugiego rzędu nazywamy wyrażenia

$$
\begin{aligned}
f[x_0; x_1; x_2] &= \frac{f[x_1; x_2] - f[x_0; x_1]}{x_2 - x_0} \\
&\vdots \\
f[x_{n-2}; x_{n-1}; x_n] &= \frac{f[x_{n-1}; x_n] - f[x_{n-2}; x_{n-1}]}{x_n - x_{n-2}}
\end{aligned}
$$

# Ilorazy różnicowe

## Definicja

Ilorazem różnicowym rzędu $n$ nazywamy

$$f[x_i; x_{i+1}; \ldots; x_{i+n}] = \frac{f[x_{i+1}; \ldots; x_{i+n}] - f[x_i; x_{i+1}; \ldots; x_{i+n-1}]}{x_{i+n} - x_i}$$

dla $n = 1, 2, \ldots$ oraz $i = 0, 1, 2, \ldots$

# Ilorazy różnicowe

| $x_i$ | $f(x_i)$ | Ilorazy różnicowe | | | |
|---|---|---|---|---|---|
| | | rzędu 1 | rzędu 2 | rzędu 3 | rzędu 4 |
| $x_0$ | $f(x_0)$ | | | | |
| | | $f[x_0; x_1]$ | | | |
| $x_1$ | $f(x_1)$ | | $f[x_0; x_1; x_2]$ | | |
| | | $f[x_1; x_2]$ | | $f[x_0; x_1; x_2; x_3]$ | |
| $x_2$ | $f(x_2)$ | | $f[x_1; x_2; x_3]$ | | $f[x_0; x_1; x_2; x_3; x_4]$ |
| | | $f[x_2; x_3]$ | | $f[x_1; x_2; x_3; x_4]$ | |
| $x_3$ | $f(x_3)$ | | $f[x_2; x_3; x_4]$ | | |
| | | $f[x_3; x_4]$ | | | |
| $x_4$ | $f(x_4)$ | | | | |

# Wzór Newtona

## Twierdzenie

$$W_n(x) = f(x_0) + f[x_0; x_1]\omega_0(x) + \ldots + f[x_0; \ldots; x_n]\omega_{n-1}(x),$$

*przy czym* $\omega_i(x) = (x - x_0) \ldots (x - x_i)$.

# Wzór Newtona

## Dowód.

Dla $n = 0$ twierdzenie jest prawdziwe, ponieważ

$$W_0(x) \equiv f(x_0).$$

Załóżmy, że jest ono prawdziwe dla pewnego $k - 1 > 0$. Różnica $W_k(x) - W_{k-1}(x)$ jest wielomianem, który można przedstawić w postaci

$$W_k(x) - W_{k-1}(x) = A(x - x_0) \dots (x - x_{k-1}),$$

gdzie $A$ jest współczynnikiem przy najwyższej potędze $x$ wielomianu $W_k(x)$.

# Wzór Newtona

## Dowód.

Według założenia indukcyjnego, najwyższymi współczynnikami wielomianów $W_{0,1,\ldots,k-1}$ oraz $W_{1,\ldots,k}$ są odpowiednio $f[x_0, \ldots, x_{k-1}]$ i $f[x_1, \ldots, x_k]$. Ze wzoru Neville'a wynika

$$W_k(x) = \frac{(x - x_0)W_{1,\ldots,k}(x) - (x - x_k)W_{0,\ldots,k-1}(x)}{x_k - x_0},$$

zatem

$$A = \frac{f[x_1, \ldots, x_k] - f[x_0, \ldots, x_{k-1}]}{x_k - x_0} = f[x_0, \ldots, x_k].$$

□

# Wzór Newtona

## Przykład

| $x_i$ | 0 | 2 | 3 | 4 | 6 |
|---|---|---|---|---|---|
| $y_i$ | 1 | 3 | 2 | 5 | 7 |

| $x_i$ | $f(x_i)$ | $f[x_i; x_{i+1}]$ | $f[x_i; x_{i+1}; x_{i+2}]$ | $f[x_i; \ldots; x_{i+3}]$ | $f[x_i; \ldots; x_{i+4}]$ |
|---|---|---|---|---|---|
| 0 | 1 | | | | |
| | | 1 | | | |
| 2 | 3 | | $-\frac{2}{3}$ | | |
| | | -1 | | $\frac{2}{3}$ | |
| 3 | 2 | | 2 | | $-\frac{2}{9}$ |
| | | 3 | | $-\frac{2}{3}$ | |
| 4 | 5 | | $-\frac{2}{3}$ | | |
| | | 1 | | | |
| 6 | 7 | | | | |

# Wzór Newtona

$$
\begin{aligned}
W_4(x) &= 1 + 1(x-0) - \frac{2}{3}(x-0)(x-2) \\
&\quad + \frac{2}{3}(x-0)(x-2)(x-3) \\
&\quad - \frac{2}{9}(x-0)(x-2)(x-3)(x-4) \\
&= -\frac{2}{9}x^4 + \frac{8}{3}x^3 - \frac{88}{9}x^2 + \frac{35}{3}x + 1
\end{aligned}
$$

# Wzór Newtona - implementacja

```
n = len(x)-1
W = np.copy(y)
for i in range(1,n+1):
    for j in range(n,i-1,-1):
        W[j]=(W[j]-W[j-1])/(x[j]-x[j-i])
```

$\Rightarrow$ $W$ zawiera potrzebne ilorazy różnicowe w kolejności $W[0] = f(x_0)$, $W[1] = f[x_0, x_1], \ldots, W[n] = f[x_0, \ldots, x_n]$

Sprawdzenie ($n = 2$):

$i = 1, \quad j = 2, \quad W[2] = \frac{y_2 - y_1}{x_2 - x_1} = f[x_1, x_2]$

$\quad\quad\; j = 1, \quad W[1] = \frac{y_1 - y_0}{x_1 - x_0} = f[x_0, x_1]$

$i = 2, \quad j = 2, \quad W[2] = \frac{f[x_1, x_2] - W[1]}{x_2 - x_0} = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0} = f[x_0, x_1, x_2]$

# Wzór Newtona - współczynniki wielomianu

Chcemy zapisać wielomian interpolacyjny w postaci

$$W_n(x) = a_0 + a_1 x + \ldots + a_n x^n$$

## Przykład

$(n = 2)$:

$$\begin{aligned} W(x) &= f[x_0, x_1, x_2]x^2 \\ &\quad + \{f[x_0, x_1] - f[x_0, x_1, x_2](x_1 + x_2)\}x \\ &\quad + \{f(x_0) - f(x_0, x_1)x_0 + f[x_0, x_1, x_2]x_0 x_1\} \end{aligned}$$

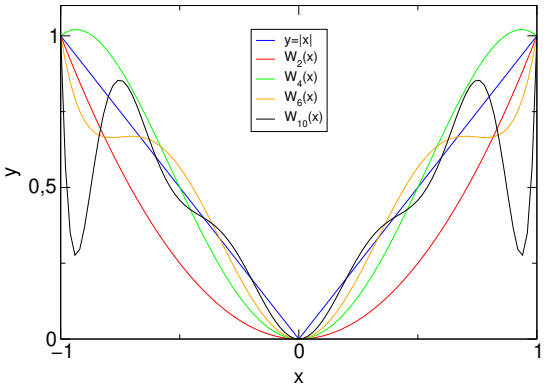
# Wzór Newtona - współczynniki wielomianu

```
a = np.copy(W)
for i in range(n):
    for j in range(n-1,i-1,-1):
        a[j]=a[j]-x[j-i]*a[j+1]
```

Sprawdzenie ($n = 2$):

$$
\begin{array}{llll}
i = 0 & j = 1 & a[1] = f[x_0, x_1] - x_1 f[x_0, x_1, x_2] \\
 & j = 0 & a[0] = f(x_0) - x_0\{f[x_0, x_1] - f[x_0, x_1, x_2]\} \\
i = 1 & j = 1 & a[1] = f[x_0, x_1] - (x_0 + x_1) f[x_0, x_1, x_2] \\
i = 2 & & a[2] = f[x_0, x_1, x_2]
\end{array}
$$

# Zbieżność procesów interpolacyjnych



# Interpolacja wymierna

Mamy $(n+1)$ punktów węzłowych

$$(x_i, y_i = f(x_i)), \quad i = 0, 1, \ldots, n$$

Wartość funkcji $f(x)$ przybliżamy funkcją wymierną

$$\phi^{\mu,\nu}(x) \equiv \frac{P^\mu(x)}{Q^\nu(x)} = \frac{a_0 + a_1 x_1 + \ldots + a_\mu x^\mu}{b_0 + b_1 x_1 + \ldots + b_\nu x^\nu}$$

spełniającą warunek

$$\phi^{\mu,\nu}(x_i) = y_i, \quad i = 0, 1, 2, \ldots, n$$

# Pułapka!

Na podstawie

$$\phi^{\mu,\nu}(x_i) = y_i, \quad i = 0, 1, 2, \ldots, n$$

można wnioskować, że

$$P^{\mu}(x_i) - y_i Q^{\nu}(x_i) = 0, \quad i = 0, 1, \ldots, \mu + \nu$$

określa współczynniki $a_i$, $b_i$

# Pułapka!

## Przykład

Niech $\mu = \nu = 1$ oraz

| $x_i$ | 0 | 1 | 2 |
|---|---|---|---|
| $y_i$ | 1 | 2 | 2 |

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

# Pułapka!

Zakładając, że $b_1 = 1$, otrzymamy

$$b_0 = 0, a_0 = 0, \ a_1 = 2$$

czyli

$$\phi^{1,1}(x) = \frac{2x}{x}$$

$\Rightarrow$ funkcja $\phi^{1,1}(x)$ nie rozwiązuje zadania interpolacji

# Odwrotności ilorazów różnicowych

## Definicja

Odwrotnościami ilorazów różnicowych nazywamy wielkości

$$\begin{aligned}
\varphi[x_i; x_j] &= \frac{x_i - x_j}{y_i - y_j} \\
\varphi[x_i; x_j; x_k] &= \frac{x_j - x_k}{\varphi[x_i; x_j] - \varphi[x_i; x_k]} \\
\varphi[x_i; \ldots; x_l; x_m; x_n] &= \frac{x_m - x_n}{\varphi[x_i; \ldots; x_l; x_m] - \varphi[x_i; \ldots; x_l; x_n]}
\end{aligned}$$

przy czym niektóre z nich mogą być nieskończone ze względu na zerowanie się mianowników.

# Interpolacja wymierna

Spełniona jest następująca własność:

$$\begin{aligned}\frac{P^n(x)}{Q^n(x)} &= y_o + \frac{P^n(x)}{Q^n(x)} - y_o = y_o + \frac{P^n(x)}{Q^n(x)} - \frac{P^n(x_o)}{Q^n(x_o)} \\ &= y_o + (x - x_o)\frac{P^{n-1}(x)}{Q^n(x)} = y_o + \frac{x_i - x_o}{\frac{Q^n(x)}{P^{n-1}(x)}}\end{aligned}$$

Stąd wynika

$$\frac{Q^n(x_i)}{P^{n-1}(x_i)} = \frac{x_i - x_o}{y_i - y_o} = \varphi[x_o; x_i], \quad i = 1, 2, \ldots, 2n$$

# Interpolacja wymierna

czyli

$$\begin{aligned}\frac{Q^n(x)}{P^{n-1}(x)} &= \varphi[x_0; x_1] + \frac{Q^n(x)}{P^{n-1}(x)} - \frac{Q^n(x_1)}{P^{n-1}(x_1)} \\ &= \varphi[x_0; x_1] + (x - x_1)\frac{Q^{n-1}(x)}{P^{n-1}(x)} \\ &= \varphi[x_0; x_1] + \frac{x - x_1}{\frac{P^{n-1}(x)}{Q^{n-1}(x)}},\end{aligned}$$

co prowadzi do

$$\frac{P^{n-1}(x_i)}{Q^{n-1}(x_i)} = \varphi[x_0; x_1; x_i], \quad i = 2, 3, \ldots, 2n$$

# Interpolacja wymierna

Ostatecznie

$$\begin{aligned}\phi^{n,n}(x) &= \frac{P^n(x)}{Q^n(x)} = y_0 + \frac{x - x_0}{\frac{Q^n(x)}{P^{n-1}(x)}} \\ &= y_0 + \frac{x - x_0}{\varphi[x_0; x_1] + \frac{\frac{x - x_1}{P^{n-1}(x)}}{Q^{n-1}(x)}} \\ &= \dots\end{aligned}$$

# Interpolacja wymierna

$\phi^{n,n}(x)$ można więc przedstawić w postaci następującego ułamka łańcuchowego:

$$\begin{aligned}\phi^{n,n}(x) \;=\; & y_0 + \underline{x - x_0}/\overline{\varphi[x_0; x_1]} + \underline{x - x_1}/\overline{\varphi[x_0; x_1; x_2]} \\ & + \underline{x - x_2}/\overline{\varphi[x_0; x_1; x_2; x_3]} + \cdots \\ & + \underline{x - x_{2n-1}}/\overline{\varphi[x_0; x_1; \ldots; x_{2n}]}\end{aligned}$$

# Interpolacja wymierna

## Przykład

Mamy daną tabelę odwrotności ilorazów różnicowych:

| $x_i$ | $y_i$ | $\varphi[x_0; x_i]$ | $\varphi[x_0; x_1; x_i]$ | $\varphi[x_0; x_1; x_2; x_i]$ |
|---|---|---|---|---|
| 0 | 0 | | | |
| 1 | $-1$ | -1 | | |
| 2 | $-\frac{2}{3}$ | -3 | $-\frac{1}{2}$ | |
| 3 | 9 | $\frac{1}{3}$ | $\frac{3}{2}$ | $\frac{1}{2}$ |

$$\phi^{2,1}(x) = 0 + \underline{x}/\overline{-1} + \underline{x-1}/\overline{-\frac{1}{2}} + \underline{x-2}/\overline{\frac{1}{2}} = \frac{x}{-1 + \frac{x-1}{-\frac{1}{2} + \frac{x-2}{\frac{1}{2}}}} = \frac{4x^2 - 9x}{-2x + 7}$$

# Metody numeryczne

## Wykład 6 - Interpolacja, część II

Janusz Szwabiński

# Plan wykładu

1. Interpolacja trygonometryczna

2. Interpolacja funkcjami sklejanymi
   2.1 Interpolacja przedziałami liniowa
   2.2 Interpolacja funkcjami sklejanymi

# Funkcje okresowe

## Definicja

Funkcję $f : \mathbb{R} \to \mathbb{R}$ nazywamy okresową z okresem $h > 0$, jeśli

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

dla każdego $x \in \mathbb{R}$

$\Rightarrow$ funkcja $f$ jest określona jednoznacznie przez jej wartości dla $x \in < 0, h)$

$\Rightarrow$ funkcja $\tilde{f} := f\left(\frac{hx}{2\pi}\right)$ ma okres $2\pi$

# Wielomiany trygonometryczne

## Definicja

(Rzeczywistymi) Wielomianami trygonometrycznymi stopnia co najwyżej $n$ nazywamy elementy zbioru

$$T_n^{\mathbb{R}} := \{T(x) : \quad T(x) = \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx), \quad a_k, b_k \in \mathbb{R}\}$$

$T_n^{\mathbb{R}}$ jest $(2n+1)$ wymiarową przestrzenią wektorową nad $\mathbb{R}$

# Wielomiany trygonometryczne

$$e^{ix} = \cos x + i \sin x \quad \Rightarrow \quad \begin{cases} \cos x = \frac{1}{2}(e^{ix} + e^{-ix}) \\ \sin x = -\frac{i}{2}(e^{ix} - e^{-ix}) \end{cases}$$

Stąd wynika

$$\begin{aligned} T(x) &= \frac{a_0}{2} + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx) \\ &= e^{-inx} \sum_{k=0}^{2n} c_k e^{ikx} = e^{-inx} P(x) \end{aligned}$$

$$c_{n-k} = \frac{a_k + ib_k}{2}, \quad c_{n+k} = \frac{a_k - ib_k}{2}, \quad 1 \leqslant k \leqslant n, \quad c_n = \frac{a_0}{2}$$

# Wielomiany trygonometryczne

## Definicja

Elementy zbioru

$$P_{n-1}^{\mathbb{C}} := \{P(x) : \quad P(x) = \sum_{k=0}^{n-1} c_k e^{ikx}, \quad c_k \in \mathbb{C}\}$$

nazywamy (zespolonymi) wielomianami trygonometrycznymi stopnia co najwyżej $(n-1)$.

# Interpolacja trygonometryczna

## Twierdzenie

*Dla dowolnych liczb zespolonych $y_k = f(x_k)$, $k = 0, 1, \ldots, n-1$, oraz parami różnych $x_k \in [0, 2\pi)$ istnieje dokładnie jeden wielomian trygonometryczny*

$$P(x) = c_0 + c_1 e^{ix} + c_2 e^{2ix} + \ldots + c_{n-1} e^{(n-1)ix}$$

*taki, że*

$$P(x_k) = y_k$$

*dla $k = 0, 1, \ldots, n-1$.*

# Interpolacja trygonometryczna

## Dowód.

Niech

$$\begin{aligned} \omega &= e^{ix} \\ \omega_k &= e^{ix_k} \quad (\Rightarrow\ \omega_j \neq \omega_k \text{ dla } j \neq k, 0 \leqslant j, k \leqslant n-1) \\ P(\omega) &= c_0 + c_1\omega + \ldots c_{n-1}\omega^{n-1} \end{aligned}$$

Zagadnienie interpolacji tryg. jest więc równoważne znalezieniu wielomianu stopnia $\leqslant (n-1)$ takiego, że

$$P(\omega_k) = y_k, \quad k = 0, 1, \ldots, n-1$$

Z twierdzenia o jednoznaczności interpolacji wielomianowej wynika, że istnieje jeden taki wielomian. □

# Interpolacja trygonometryczna - węzły równoodległe

Niech

$$\zeta_n = e^{\frac{2\pi i}{n}}, \quad x_k = \frac{2\pi k}{n}, \quad k = 0, \ldots, n-1$$

## Twierdzenie

*Dla danych punktów węzłowych $(x_k, y_k)$, $k = 0, \ldots, n-1$, współczynniki wielomianu trygonometrycznego*

$$P(x) = \sum_{k=0}^{n-1} c_k e^{ikx}, \quad x \in [0, 2\pi)$$

*mają postać*

$$c_k = \frac{1}{n} \sum_{l=0}^{n-1} y_l \zeta_n^{-kl}, \quad l = 0, \ldots, n-1$$

# Interpolacja trygonometryczna - węzły równoodległe

## Dowód.

Wystarczy pokazać, że powyższy wielomian rzeczywiście interpoluje dane. Mamy:

$$\begin{aligned} P(x_m) &= \sum_{k=0}^{n-1} \frac{1}{n} \sum_{l=0}^{n-1} y_l \zeta_n^{-kl} e^{ikx_m} = \sum_{l=0}^{n-1} y_l \frac{1}{n} \sum_{k=0}^{n-1} \zeta^{k(m-l)} \\ &= \sum_{l=0}^{n-1} y_l \delta_{ml} = y_m \end{aligned}$$

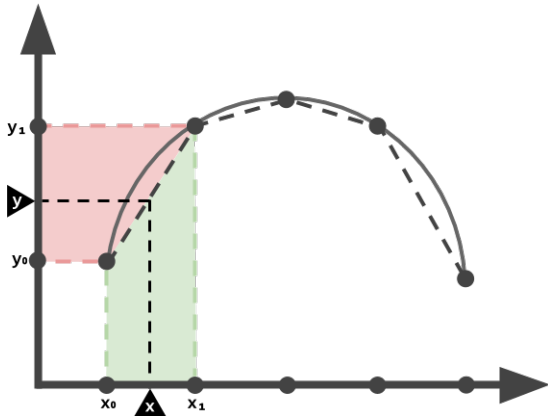
dla $m = 0, \ldots, n-1$ □

# Interpolacja trygonometryczna - węzły równoodległe

Wnioski:

- $\Rightarrow$ $O(n)$ operacji potrzebnych do wyliczenia $c_k$
- $\Rightarrow$ $O(n^2)$ operacji potrzebnych do wyznaczenia wielomianu trygonometrycznego

# Interpolacja przedziałami liniowa



# Interpolacja przedziałami liniowa

$(x_i, y_i = f(x_i)), \quad i = 0, 1, \ldots, n$ - punkty węzłowe

W przedziale $[x_j, x_{j+1}]$, $0 \leqslant j \leqslant n-1$ przybliżamy $f(x)$ funkcją liniową:

$$f(x) = y_j + \frac{y_{j+1} - y_j}{x_{j+1} - x_j}(y_{j+1} - y_j) + \Delta f(x)$$

gdzie

$$\Delta f(x) = \frac{\gamma}{2}(x - x_j)(x - x_{j+1})$$

# Interpolacja przedziałami liniowa

Jeśli np. $f(x)$ jest funkcją gładką w $[x_j, x_{j+1}]$, to

$$\gamma = f''(a), \quad a \in [x_j, x_{j+1}]$$

Zachodzi

$$|\Delta f(x)| \leqslant \frac{\gamma_1}{8}(x_{j+1} - x_j)^2$$

gdzie

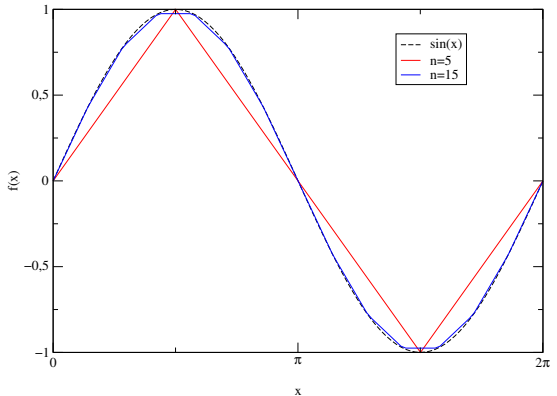
$$\gamma_1 = \max_{x \in [x_j, x_{j+1}]} |f''(x)|$$

$\Rightarrow$ dokładność można zwiększyć, zmniejszając długość przedziału $h = x_{j+1} - x_j$ (zwiększając liczbę węzłów)

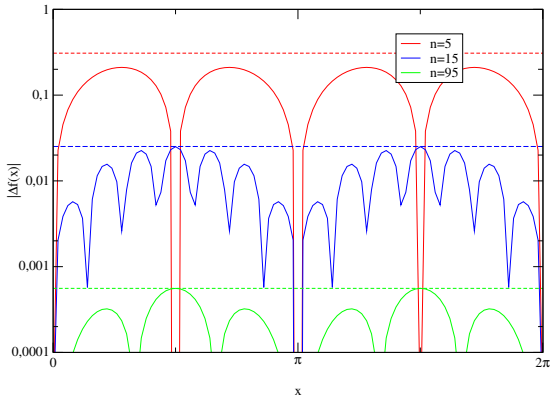
# Interpolacja przedziałami liniowa

```
def linint(XX, fxx, X):
    N = len(XX)
    JL = 0; JU = N
    while JU - JL > 1:
        JM = (JU + JL) // 2
        if (XX[-1] > XX[1]) == (X > XX[JM]):
            JL = JM
        else:
            JU = JM
    J = JL
    dx = XX[J + 1] - XX[J]
    df = fxx[J + 1] - fxx[J]
    fx = df / dx * (X - XX[J]) + fxx[J]
    return fx
```

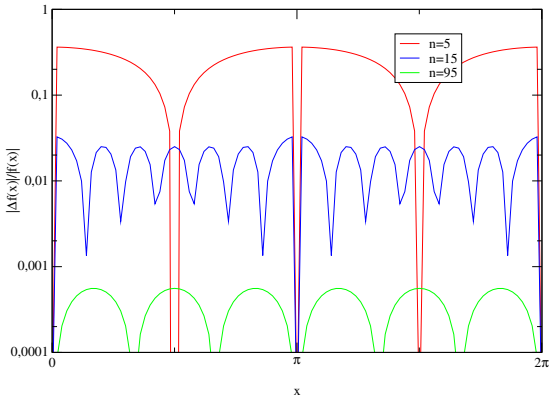
# Interpolacja przedziałami liniowa



# Interpolacja przedziałami liniowa



# Interpolacja przedziałami liniowa



# Funkcje sklejane

$x_i$, $i = 0, 1, \ldots, n$ - punkty węzłowe w przedziale $[a, b]$

$$a = x_0 < x_1 < \ldots < x_{n-1} < x_n = b$$

$\Delta_n$ - podział przedziału $[a, b]$

## Definicja

Funkcję $S(x) = S(x, \Delta_n)$ określoną na przedziale $[a, b]$ nazywamy funkcją sklejaną stopnia $m$ ($m \geqslant 1$), jeżeli

a. $S(x)$ jest wielomianem stopnia co najwyżej $m$ na każdym podprzedziale $(x_j, x_{j+1})$, $j = 0, 1, \ldots, n-1$

b. $S(x) \in C^{m-1}([a, b])$

# Funkcje sklejane

$S_m(\Delta_n)$ - zbiór wszystkich funkcji sklejanych stopnia $m$

$S(x) \in S_m(\Delta_n)$

$\Rightarrow S(x) = c_{i0} + c_{i1}x + \ldots + c_{im}x^m, \quad x \in (x_i, x_{i+1})$

$\Rightarrow n(m+1)$ dowolnych stałych $c_{ij}$

Żądanie ciągłości pochodnych rzędu $0, \ldots, m-1$ w każdym węźle wewnętrznym $x_i$, $i = 1, \ldots, n-1$

$\Rightarrow (n-1)m$ warunków na stałe $c_{ij}$

$\Rightarrow S(x)$ zależy od $n(m+1) - m(n-1) = n + m$ parametrów

# Funkcje sklejane trzeciego stopnia

Niech $f(x) \in C([a, b])$

## Definicja

Funkcję $S(x) \in S_3(\Delta_n)$ nazywamy interpolacyjną funkcją sklejaną stopnia trzeciego dla funkcji $f(x)$, jeżeli

$$S(x_i) = f(x_i) = y_i, \quad i = 0, 1, \ldots, n, \quad n \geqslant 2$$

$S(x)$ stopnia trzeciego zależy od $(n + 3)$ parametrów

$\Rightarrow$ potrzebujemy dodatkowych warunków na $S(x)$

# Funkcje sklejane trzeciego stopnia

Najczęściej zakłada się

$$S'(a+0) = \alpha_1, \;\; S'(b-0) = \beta_1$$

lub

$$S''(a+0) = \alpha_2, \;\; S''(b-0) = \beta_2$$

gdzie $\alpha_1$, $\beta_1$, $\alpha_2$ i $\beta_2$ - ustalone liczby rzeczywiste

Jeżeli $f(x)$ jest funkcją okresową o okresie $(b-a)$, wówczas:

$$S^{(i)}(a+0) = S^{(i)}(b-0), \;\; i = 1, 2$$

# Funkcje sklejane trzeciego stopnia

Pytania:

- czy zagadnienie interpolacji za pomocą funkcji sklejanych stopnia trzeciego z jednym z dodatkowych warunków jest rozwiązywalne?
- czy rozwiązanie jest jednoznaczne?
- jaki jest błąd interpolacji na przedziale $[a, b]$?

# Funkcje sklejane trzeciego stopnia

Niech

$$M_j = S''(x_j), j = 0, 1, \ldots, n$$
$$h_j = x_j - x_{j-1}$$

$S''(x)$ jest funkcją ciągłą na przedziale $[a, b]$ i liniową na każdym podprzedziale $[x_{j-1}, x_j]$

$$\Rightarrow \quad S''(x) = M_{j-1}\frac{x_j - x}{h_j} + M_j\frac{x - x_{j-1}}{h_j}, \quad x \in [x_{j-1}, x_j]$$

# Funkcje sklejane trzeciego stopnia

Całkując stronami otrzymujemy ($x \in [x_{j-1}, x_j]$):

$$\begin{aligned} S'(x) &= -M_{j-1}\frac{(x_j - x)^2}{2h_j} + M_j\frac{(x - x_{j-1})^2}{2h_j} + A_j \\ S(x) &= M_{j-1}\frac{(x_j - x)^3}{6h_j} + M_j\frac{(x - x_{j-1})^3}{6h_j} + A_j(x - x_{j-1}) + B_j \end{aligned}$$

gdzie $A_j$ i $B_j$ - pewne stałe. Z warunku interpolacji wynika:

$$\begin{aligned} B_j &= y_{j-1} - M_j\frac{h_j^2}{6} \\ A_j &= \frac{y_j - y_{j-1}}{h_j} - \frac{h_j}{6}(M_j - M_{j-1}) \end{aligned}$$

# Funkcje sklejane trzeciego stopnia

$S'(x)$ ma być funkcją ciągłą na $[a, b]$:

$$S'(x_j - 0) = S'(x_j + 0), \quad j = 1, \ldots, n-1$$

Otrzymujemy układ $(n-1)$ równań ($j = 1, \ldots, n-1$)

$$\mu_j M_{j-1} + 2M_j + \lambda_j M_{j+1} = d_j, \quad j = 1, \ldots, n-1$$

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

# Funkcje sklejane trzeciego stopnia

Dodatkowy warunek dla pierwszych pochodnych

$$\begin{aligned} \Rightarrow \qquad & 2M_0 + M_1 = d_0 \\ & M_{n-1} + 2M_n = d_n \end{aligned}$$

gdzie

$$\begin{aligned} d_0 &= \frac{6}{h_1}\left(\frac{y_1 - y_0}{h_1} - \alpha_1\right) \\ d_n &= \frac{6}{h_n}\left(\beta_1 - \frac{y_n - y_{n-1}}{h_n}\right) \end{aligned}$$

# Funkcje sklejane trzeciego stopnia

Stąd

$$\begin{pmatrix} 2 & 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & 2 & \lambda_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & 2 & \lambda_2 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 2 & \lambda_{n-1} \\ 0 & \cdots & \cdots & \cdots & 1 & 2 \end{pmatrix} \begin{pmatrix} M_0 \\ M_1 \\ M_2 \\ \vdots \\ M_{n-1} \\ M_n \end{pmatrix} = \begin{pmatrix} d_0 \\ d_1 \\ d_2 \\ \vdots \\ d_{n-1} \\ d_n \end{pmatrix}$$

# Funkcje sklejane trzeciego stopnia

## Twierdzenie

*Macierz*

$$A = \begin{pmatrix} 2 & 1 & 0 & 0 & \cdots & 0 \\ \mu_1 & 2 & \lambda_1 & 0 & \cdots & 0 \\ 0 & \mu_2 & 2 & \lambda_2 & \cdots & 0 \\ \vdots & & & & & \vdots \\ 0 & \cdots & \cdots & \cdots & 2 & \lambda_{n-1} \\ 0 & \cdots & \cdots & \cdots & 1 & 2 \end{pmatrix}$$

*jest nieosobliwa dla każdego podziału* $\Delta_n$ *przedziału* $[a, b]$

# Funkcje sklejane trzeciego stopnia

## Dowód.

Pokażemy najpierw, że dla każdej pary wektorów $x, y \in \mathbb{R}$ zachodzi

$$Ax = y \quad \Rightarrow \quad \max_i |x_i| \leqslant \max_i |y_i|$$

Niech $|x_r| = \max_i |x_i|$. Ponieważ $Ax = y$, więc

$$\mu_r x_{r-1} + 2x_r + \lambda_r x_{r+1} = y_r$$

# Funkcje sklejane trzeciego stopnia

## Dowód.

Z definicji $r$ oraz z tego, że $\mu_r + \lambda_r \leqslant 1$ wynika

$$\begin{aligned} \max_i |y_i| &\geqslant |y_r| \geqslant 2|x_r| - \mu_r|x_{r-1}| - \lambda_r|x_{r+1}| \\ &\geqslant 2|x_r| - \mu_r|x_r| - \lambda_r|x_r| \\ &= (2 - \mu_r - \lambda_r)|x_r| \geqslant |x_r| = \max_i |x_i| \end{aligned}$$

# Funkcje sklejane trzeciego stopnia

**Dowód.**
Jeśli macierz byłaby osobliwa, to istniałoby rozwiązanie $x \neq 0$ układu $Ax = 0$. Ale wówczas musiałoby być

$$0 < \max_i |x_i| \leqslant 0$$

co prowadzi do sprzeczności. □

# Funkcje sklejane trzeciego stopnia

## Definicja

Średnicą $||\Delta_m||$ podziału $\Delta_m$ nazywamy liczbę

$$||\Delta_m|| = \max_i (x_{i+1} - x_i)$$

# Funkcje sklejane trzeciego stopnia

## Twierdzenie

*Niech* $f \in C^4([a,b])$, $|f^{(4)}(x)| \leqslant L$ *dla* $x \in [a,b]$. *Niech* $\Delta_m = \{a = x_0^{(m)} < x_1^{(m)} < \ldots < x_{n_m}^{(m)} = b\}$ *będzie ciągiem podziałów przedziału* $[a,b]$ *takim, że*

$$\sup_{m,j} \frac{||\Delta_m||}{x_{j+1}^{(m)} - x_j^{(m)}} \leqslant K < +\infty$$

*Istnieją wtedy stałe* $C_i$ $(\leqslant 2)$ *niezależne od* $\Delta_m$ *takie, że dla* $x \in [a,b]$ *zachodzi*

$$|f^{(i)}(x) - S^{(i)}(x, \Delta_m)| \leqslant C_i L K ||\Delta_m||^{4-i}$$

*dla* $i = 0, 1, 2, 3$.

# Funkcje sklejane Lagrange'a trzeciego stopnia

$\lambda_0(x)$ - wielomian interpolujący dla funkcji $f(x)$ z węzłami $x_0, x_1, x_2, x_3$

$\lambda_n(x)$ - z węzłami $x_{n-3}, x_{n-2}, x_{n-1}, x_n$

## Definicja

Funkcję $S_L(x) \in S_3(\Delta_n)$ spełniającą dodatkowe warunki

$$\begin{aligned} S_L'(a+0) &= \alpha_1 = \lambda_0'(a), \\ S_L'(b-0) &= \beta_1 = \lambda_n'(b) \end{aligned}$$

nazywamy funkcją sklejaną Lagrange'a stopnia trzeciego

# Interpolacja przy węzłach równoodległych

Niech

$$x_i = x_0 + ih, \quad h = \frac{b-a}{n}, \quad i = 0, 1, \ldots, n.$$

Dodatkowo

$$x_{-3} = x_0 - 3h \quad x_{-2} = x_0 - 2h \quad x_{-1} = x_0 - h$$
$$x_{n+1} = x_n + h \quad x_{n+2} = x_n + 2h \quad x_{n+3} = x_n + 3h$$

# Interpolacja przy węzłach równoodległych

## Definicja

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

# Interpolacja przy węzłach równoodległych

## Twierdzenie

*Funkcje* $\overline{\phi}_i(x)$, $i = -1, \ldots, n+1$ *określone na przedziale* $[a, b]$ *w następujący sposób:*

$$\overline{\phi}_i(x) = \phi_i^3(x), \quad a \leqslant x \leqslant b$$

*stanowią bazę funkcji sklejanych trzeciego stopnia* $S_3(\Delta_n)$.

$\Rightarrow$ każdą funkcję $S(x) \in S_3(\Delta_n)$ można przedstawić w postaci

$$S(x) = \sum_{i=-1}^{n+1} c_i \phi_i^3(x), \quad a \leqslant x \leqslant b, \quad c_i \in \mathbb{R}$$

# Interpolacja przy węzłach równoodległych

Z warunku interpolacji otrzymujemy

$$c_{i-1} + 4c_i + c_{i+1} = y_i, \quad i = 0, \ldots, n.$$

Założenie dodatkowe

$$S'(a+0) = \alpha_1, \quad S'(b-0) = \beta_1$$

$\Rightarrow$

$$\begin{aligned} -c_{-1} + c_1 &= \frac{h}{3}\alpha_1 \\ -c_{n-1} + c_{n+1} &= \frac{h}{3}\beta_1 \end{aligned}$$

# Interpolacja przy węzłach równoodległych

Stąd

$$\begin{pmatrix} 4 & 2 & & & & \\ 1 & 4 & 1 & & 0 & \\ & 1 & 4 & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & 0 & & 1 & 4 & 1 \\ & & & & 2 & 4 \end{pmatrix} \begin{pmatrix} c_0 \\ c_1 \\ c_2 \\ \vdots \\ c_{n-1} \\ c_n \end{pmatrix} = \begin{pmatrix} y_0 + \frac{h}{3}\alpha_1 \\ y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n - \frac{h}{3}\beta_1 \end{pmatrix}$$

# Metody numeryczne

## Wykład 7 - Aproksymacja funkcji

Janusz Szwabiński

# Plan wykładu

1. Pojęcia podstawowe
   1.1 Zagadnienie aproksymacji
   1.2 Funkcje bazowe
   1.3 Typowe normy
   1.4 Rodzaje aproksymacji

2. Aproksymacja średniokwadratowa
   2.1 Aproksymacja wielomianowa
   2.2 Aproksymacja trygonometryczna
   2.3 Aproksymacja za pomocą funkcji sklejanych
   2.4 Aproksymacja funkcji ciągłych

# Zagadnienie aproksymacji

$\mathcal{X}$ - pewna przestrzeń liniowa
$\mathcal{X}_m$ - $m$-wymiarowa podprzestrzeń przestrzeni $\mathcal{X}$
$f(x)$ - funkcja, którą chcemy aproksymować

# Zagadnienie aproksymacji

## Definicja

Aproksymacja liniowa funkcji $f(x)$ polega na wyznaczeniu takich współczynników $a_0, a_1, \ldots, a_m$ funkcji

$$F(x) = a_0\phi_0(x) + a_1\phi_1(x) + \ldots + a_m\phi_m(x)$$

gdzie $\phi_0(x), \ldots, \phi_m(x)$ są funkcjami bazowymi podprzestrzeni $\mathcal{X}_{m+1}$, aby funkcja $F(x)$ spełniała pewne warunki, np. minimalizowała normę różnicy $\|f(x) - F(x)\|$

# Zagadnienie aproksymacji

## Definicja

Aproksymacja wymierna funkcji $f(x)$ polega na znalezieniu takich współczynników $a_0, \ldots, a_n, b_0, \ldots, b_m$ funkcji

$$F(x) = \frac{a_0\phi_0(x) + a_1\phi_1(x) + \ldots + a_n\phi_n(x)}{b_0\psi_0(x) + b_1\psi_1(x) + \ldots + b_m\psi_m(x)}$$

gdzie $\phi_i(x)$ i $\psi_j(x)$ ($i = 0, \ldots, n$, $j = 0, \ldots, m$) są elementami tej samej bazy $k$ wymiarowej podprzestrzeni liniowej ($k = \max(m, n)$), aby funkcja $F(x)$ spełniała pewne warunki, np. minimalizowała normę różnicy $\|f(x) - F(x)\|$

# Przykłady funkcji bazowych

- funkcje trygonometryczne
  $1, \sin x, \cos x, \sin 2x, \cos 2x, \ldots, \sin kx, \cos kx$
- jednomiany
  $1, x, x^2, \ldots, x^m$
- wielomiany
  $1, (x - x_0), (x - x_0)(x - x_1), \ldots, (x - x_0) \cdots (x - x_m)$
- wielomiany Czebyszewa, Legendre'a

Wybór bazy wpływa na dokładność i koszt obliczeń!!!

# Typowe normy

- Czebyszewa

$$\|f\| = \sup_{\langle a,b\rangle} |f(x)|$$

- $L_2$

$$\|f\|_2 = \left( \int_a^b |f(x)|^2 \mathrm{d}x \right)^{1/2}$$

- $L_2$ z wagą

$$\|f\|_{2,w} = \left( \int_a^b w(x)|f(x)|^2 \mathrm{d}x \right)^{1/2}$$

# Typowe normy

- „dyskretna”

$$\|f\| = \left( \sum_{i=0}^{n} [f(x_i)]^2 \right)^{1/2}$$

# Rodzaje aproksymacji

- średniokwadratowa, kiedy szukamy funkcji $F(x)$ minimalizującej całkę

$$\|f(x) - F(x)\| = \int_a^b w(x)\,[F(x) - f(x)]^2\,\mathrm{d}x$$

lub sumę

$$\|f(x) - F(x)\| = \sum_{i=0}^{n} w(x_i)[F(x_i) - f(x_i)]^2$$

$$w(x_i) \geqslant 0, \quad i = 0, 1, \ldots, n$$

# Rodzaje aproksymacji

- jednostajna, kiedy szukamy funkcji $F(x)$ minimalizującej normę

$$\|F(x) - f(x)\| = \sup_{x \in \langle a,b \rangle} |F(x) - f(x)|$$

# Rodzaje aproksymacji

## Twierdzenie

*Jeżeli funkcja $f(x)$ jest ciągła na skończonym przedziale $\langle a, b \rangle$, to dla każdego $\epsilon$ dodatniego można dobrać takie n, że jest możliwe utworzenie wielomianu $P_n(x)$ stopnia n ($n = n(\epsilon)$), który spełnia nierówność*

$$|f(x) - P_n(x)| < \epsilon$$

*na całym przedziale $\langle a, b \rangle$.*

# Rodzaje aproksymacji

## Twierdzenie

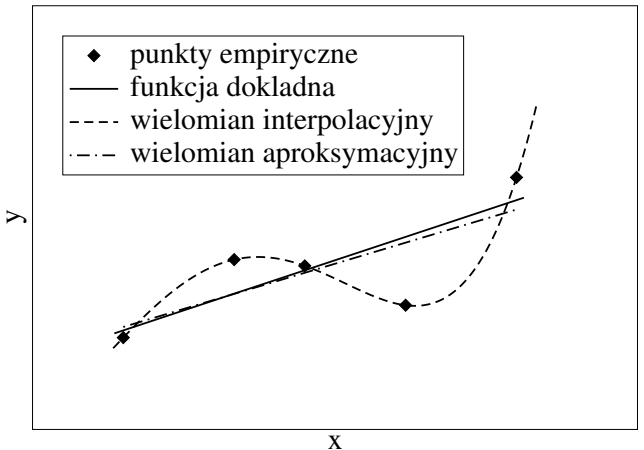
*Jeżeli funkcja $f(x)$ jest ciągła na* $\mathbf{R}$ *i okresowa o okresie* $2\pi$*, to dla każdego* $\epsilon$ *dodatniego istnieje wielomian trygonometryczny*

$$S_n(x) = a_0 + \sum_{k=1}^{n}(a_k \cos kx + b_k \sin kx), \quad n = n(\epsilon)$$

*spełniający dla wszystkich x nierówność*

$$|f(x) - S_n(x)| < \epsilon.$$

# Aproksymacja średniokwadratowa



# Aproksymacja średniokwadratowa

Szukamy wielomianu uogólnionego

$$F(x) = \sum_{i=0}^{m} a_i \phi_i(x)$$

takiego, że suma

$$\|F(x) - f(x)\| = \sum_{i=0}^{n} w(x_i) \left[F(x_i) - f(x_i)\right]^2$$

osiąga minimum.

# Aproksymacja średniokwadratowa

$$H(a_0, \ldots, a_m) = \sum_{j=0}^{n} w(x_j) \left[ f(x_j) - \sum_{i=0}^{m} a_i \phi_i(x_j) \right]^2 = \sum_{j=0}^{n} w(x_j) R_j^n$$

Układ normalny ($k = 0, 1, \ldots, m$)

$$\frac{\partial H}{\partial a_k} = -2 \sum_{j=0}^{n} w(x_j) \left[ f(x_j) - \sum_{i=0}^{m} a_i \phi_i(x_j) \right] \phi_k(x_j) = 0$$

$\phi_j(x)$ tworzą bazę

$\Rightarrow$ wyznacznik różny od zera

$\Rightarrow$ rozwiązanie układu minimalizuje sumę $\|F(x) - f(x)\|$

# Aproksymacja wielomianowa

Niech $\phi_i(x) = x^i$, $i = 0, 1, \ldots, m$ oraz $w(x) \equiv 1$

$\Rightarrow$ układ normalny ma postać

$$\sum_{j=0}^{n} \left[ f(x_j) - \sum_{i=0}^{m} a_i x_j^i \right] x_j^k = 0, \quad k = 0, 1, \ldots, m$$

Stąd

$$\sum_{i=0}^{m} a_i g_{ik} = \rho_k, \quad k = 0, 1, \ldots, m$$

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

# Aproksymacja wielomianowa

Jeśli punkty $x_0, \ldots, x_n$ są różne oraz

- $m \leqslant n$

  $\Rightarrow$ wyznacznik układu jest różny od zera

  $\Rightarrow$ układ ma jednoznaczne rozwiązanie

- $m = n$

  $\Rightarrow$ $F(x)$ pokrywa się z wielomianem interpolacyjnym

  $\Rightarrow$ $H = 0$

# Aproksymacja wielomianowa

Uwaga!

Dla $m \geqslant 6$ układ normalny aproksymacji wielomianowej jest źle uwarunkowany

$\Rightarrow$ aproksymację z jednomianami jako funkcjami bazowymi stosujemy tylko dla małych $m$

$\Rightarrow$ dla dużych $m$ lepiej stosować jako bazę wielomiany ortogonalne

# Aproksymacja trygonometryczna

$f(x)$ jest określona na dyskretnym zbiorze punktów

$$x_i = \frac{\pi i}{L}, \quad i = 0, 1, \ldots, 2L - 1$$

Mamy

$$\sum_{i=0}^{2L-1} \sin mx_i \sin kx_i = \begin{cases} 0, & m \neq k \\ L, & m = k \neq 0 \\ 0, & m = k = 0 \end{cases}$$

# Aproksymacja trygonometryczna

$$\sum_{i=0}^{2L-1} \cos mx_i \cos kx_i = \begin{cases} 0, & m \neq k \\ L, & m = k \neq 0 \\ 2L, & m = k = 0 \end{cases}$$

$$\sum_{i=0}^{2L-1} \cos mx_i \sin kx_i = 0$$

# Aproksymacja trygonometryczna

Szukamy funkcji aproksymującej postaci

$$y_n(x) = \frac{1}{2} + \sum_{j=1}^{n} \left( a_j \cos jx + b_j sinjx \right), \quad n < L$$

Żądanie minimalizacji sumy

$$\sum_{i=0}^{2L-1} [f(x_i) - y_n(x_i)]^2$$

prowadzi do

# Aproksymacja trygonometryczna

$$
\begin{aligned}
a_j &= \frac{1}{L}\sum_{i=0}^{2L-1} f(x_i)\cos jx_i = \frac{1}{L}\sum_{i=0}^{2L-1} f(x_i)\cos\frac{\pi ij}{L} \\
b_j &= \frac{1}{L}\sum_{i=0}^{2L-1} f(x_i)\sin jx_i = \frac{1}{L}\sum_{i=0}^{2L-1} f(x_i)\sin\frac{\pi ij}{L}
\end{aligned}
$$

$$(j = 1, 2, \ldots, n)$$

# Aproksymacja za pomocą funkcji sklejanych

Funkcja jest określona na dyskretnym zbiorze punktów

$$x_i, \quad i = 0, 1, \ldots, n_1, \quad n_1 > n + 3$$

Funkcji aproksymacyjnej szukamy w postaci

$$S(x) = \sum_{i=-1}^{n+1} c_i \phi_i^3(x), \qquad a \leqslant x \leqslant b$$

# Aproksymacja za pomocą funkcji sklejanych

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

# Aproksymacja za pomocą funkcji sklejanych

Niech

$$I = \sum_{k=0}^{n_1} \left[ f(x_k) - \sum_{i=-1}^{n+1} c_i \phi_i^3(x_k) \right]^2$$

Warunek

$$\frac{\partial I}{\partial c_i} = 0, \quad i = -1, 0, 1, \ldots, n+1$$

prowadzi do

$$\sum_{i=-1}^{n+1} b_{ij} c_i = \sum_{k=0}^{n_1} f(x_k) \phi_j^3(x_k), \quad j = -1, 0, \ldots, n+1$$

$$b_{ij} = \sum_{k=0}^{n_1} \phi_i^3(x_k) \phi_j^3(x_k)$$

# Aproksymacja funkcji ciągłych

Szukamy funkcji aproksymującej postaci

$$P(x) = a_0\phi_0(x) + \ldots a_n\phi_n(x)$$

gdzie $\phi_j(x)$ to elementy bazy pewnej podprzestrzeni funkcji całkowalnych z kwadratem

Niech

$$H_n = \int_a^b \mathrm{d}x \, [P(x) - f(x)]^2 = \int_a^b \mathrm{d}x \left[ \sum_{i=0}^{n} a_i\phi_i(x) - f(x) \right]^2$$

# Aproksymacja funkcji ciągłych

Minimum $H_n$ będzie minimalizowało normę

$$\|P(x) - f(x)\|$$

W tym celu rozwiązujemy układ

$$\frac{\partial H_n}{\partial a_i} = 0, \quad i = 0, 1, \ldots, n$$

względem współczynników $a_i$

# Aproksymacja funkcji ciągłych

## Przykład

Funkcję $f(x) = \sin x$ na przedziale $\langle 0, \pi/2 \rangle$ aproksymujemy wielomianem

$$P(x) = a_0 + a_1 x + a_2 x^2$$

Układ równań ma postać

$$\begin{aligned}
a_0 \int_0^{\pi/2} \mathrm{d}x + a_1 \int_0^{\pi/2} x\mathrm{d}x + a_2 \int_0^{\pi/2} x^2\mathrm{d}x &= \int_0^{\pi/2} \sin x\mathrm{d}x \\
a_0 \int_0^{\pi/2} x\mathrm{d}x + a_1 \int_0^{\pi/2} x^2\mathrm{d}x + a_2 \int_0^{\pi/2} x^3\mathrm{d}x &= \int_0^{\pi/2} x \sin x\mathrm{d}x \\
a_0 \int_0^{\pi/2} x^2\mathrm{d}x + a_1 \int_0^{\pi/2} x^3\mathrm{d}x + a_2 \int_0^{\pi/2} x^4\mathrm{d}x &= \int_0^{\pi/2} x^2 \sin x\mathrm{d}x
\end{aligned}$$

# Aproksymacja funkcji ciągłych

czyli

$$\begin{aligned} \frac{\pi}{2}a_0 + \frac{\pi^2}{8}a_1 + \frac{\pi^3}{24}a_2 &= 1 \\ \frac{\pi^2}{8}a_0 + \frac{\pi^3}{24}a_1 + \frac{\pi^4}{64}a_2 &= 1 \\ \frac{\pi^3}{24}a_0 + \frac{\pi^4}{64}a_1 + \frac{\pi^5}{160}a_2 &= -2 \end{aligned}$$

Stąd

$$P(x) \simeq 0,134 + 0,59x + 0,05x^2$$

# Aproksymacja funkcji ciągłych

Średni błąd aproksymacji

$$M^2 = (b-a)^{-1} H_n(a_0, a_1, a_2) \simeq 0,00797$$

## Przykład

Funkcję $f(x) = \sin x$ na przedziale $\langle 0, \pi/2 \rangle$ aproksymujemy posługując się wielomianami Legendre'a

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n, \quad n = 0, 1, 2, \ldots$$

Wprowadzamy zmienną

$$t = \frac{4}{\pi} x - 1$$

# Aproksymacja funkcji ciągłych

Aproksymować będziemy funkcję

$$\hat{f}(t) = \sin \frac{\pi(t+1)}{4}$$

wielomianem

$$W(t) = a_0 P_0(t) + a_1 P_1(t) + a_2 P_2(t)$$

# Aproksymacja funkcji ciągłych

Współczynniki wynoszą

$$\begin{aligned} a_0 &= \frac{1}{2}\int_{-1}^{1} \mathrm{d}t \sin\frac{\pi}{4}(t+1) = \frac{2}{\pi} \\ a_1 &= \frac{3}{2}\int_{-1}^{1} \mathrm{d}t t \sin\frac{\pi}{4}(t+1) = \frac{24}{\pi^2} - \frac{6}{\pi} \\ a_2 &= \frac{5}{2}\int_{-1}^{1} \mathrm{d}t \left(\frac{3}{2}t^2 - \frac{1}{2}\right) \sin\frac{\pi}{4}(t+1) = -\frac{480}{\pi^3} + \frac{120}{\pi^2} + \frac{10}{\pi} \end{aligned}$$

Stąd

$$W(t) \simeq 0,6366197 + 0,5218492x - 0,1390961x^2$$

$$M^2 \simeq 0,0000704$$