Notatki z Metod numerycznych

Jacek Olczyk

October 2018

Część I
Wykład

1 Rozwiązywanie układów równań liniowych

Znane metody

- $Ax = b, A \in \mathbb{R}^{n \times n}$
- Algorytm rokładem LU (elim. Gaussa) z wybraniem el. gł $O(\frac{2}{3}N^3)$.
- 1. złota myśl numeryka: Co zrobić jeśli zadanie jest za trudne? Zmienić zadanie
- Zamiast rozwiązywać układ równań, przybliżamy go
- Czy da się szybciej niż Gauss, który jest $O(n^3)$? To jak nisko da się zejść to problem otwarty, ale istnieją algorytmy lepsze niż sześcian.

2 Przybliżone rozwiązywanie układów równań

- Niech $A = M - Z$, wtedy $Ax = Mx - Zx = b$, zatem $Mx = Zx + b$
- TODO Metoda iteracji prostej Banacha $Mx_{n+1} = Zx_n + b$
- Jeśli wybierzemy M tak, by układ z macierzą M można było tanio rozwiązać, wtedy iteracja też będzie tania
- Chcemy, żeby M było dobrym przybliżeniem A, ale nie aż tak łatwo że
- Metoda Jacobiego
$$a_{kk}x_k^{n+1} = b - \sum_{j \neq k} a_{kj}x_j^n$$
- inny pomysł, metoda Gaussa-Seidela: $a_{kk}x_k^{(n+1)} = b_k - \sum_{j<k} a_{kj}x_j^{(n+1)} - \sum_{j>k} a_{kj}x_j^{(n)}$

- Uwaga: fakt życiowy. Gdy n jest bardzo duże, wówczas w A jest zazwyczaj bardzo dużo zer, o ile układ pochodzi z REAL LIFE^{TM}.
- To oznacza, że ilość elementów różnych od 0 jest rzędu $O(n)$. Mówimy wtedy że macierz jest rzadka.
- Wniosek: Jeśli A ma $O(n)$ niezerowych elementów, to mnożenie Ax kosztuje też $O(n)$. Ponadto, rozwiązanie układu z macierzą dolnotrójkątną też jest $O(n)$

3 Normy macierzowe i wektorowe

Normy wektorowe

$$||x||_p := (\sum_{i=1}^{N} |x_i|^p)^{\frac{1}{p}}$$

$$||x||_\infty := \max_i |x_i|$$

Norma macierzowa

$$||A||_p := \max_{x \neq 0} \frac{||Ax||_p}{||x||_p} = \max_{||x||_p} ||Ax||_p$$

Własności normy macierzowej

1. $$||Ax|| \leqslant ||A|| \cdots ||x|| \forall_{x \in \mathbb{R}^n}$$
2. $$||Ax||$$

   - nie dało się przeczytać tablicy
3. tu też coś było :(

4 Warunek wystarczający zbieżności klasycznej metody iteracyjnej ($A = M - Z$)

$$Mx_{k+1} = b + Zx_k \qquad (*)$$

Niech x^* będzie dokładnym rozwiązaniem $Ax^* = b$

$$x_{k+1} = M^{-1}(b - Zx_k)$$

$$x_{k+1} - x^* = M^{-1}(b - Zx_k) - x^*$$
$$= M^{-1}(Ax^* - Zx_k) - x^*$$

$$= M^{-1}(Ax^* - (M-A)x_k) - x^*$$
$$= M^{-1}Ax^* + (I - M^{-1}A)x_k - x^*$$
$$= -(I - M^{-1}A)x^* + (I - M^{-1}A)x_k$$
$$= (I - M^{-1}A)(x_k - x^*)$$

Czyli B pomnożony błąd k-ty.
Czyli $x_{n+1} - x^* = B(x_k - x^*) = B^2(x_{k-1} - x^*) \ldots = B^{k+1}(x_0 - x^*)$

Wniosek: Jeśli $||B|| < 1$, to $(*)$ zbieżna do $x*$ dla dow. $x_0 \in \mathbb{R}^N$

Twierdzenie: Metoda $(*)$ jest zbieżna do x^* z dowolnego x_0 wtw gdy $\rho(B) < 1$ gdzie $\rho(B) = \max\{|\lambda| : \lambda \text{ jest wartością własną } B\}$ - promień spektralny macierzy B Dowód pominięty

Twierdzenie: Jeśli macierz A jest ściśle diagonalnie dominująca, tzn zachodzi $|a_n| > \sum_{j \neq i} |a_{ij}|$ dla $i = 1..N$ to metoda Jacobiego jest zbieżna (dla dowolnych $x_n \in R^n$)

Dowód. Zbadajmy macierz iteracji.

$$||B||_\infty = ||I - M^{-1}A||_\infty$$

$M^{-}1$ dla macierzy diagonalnej to podnoszenie wszystkich elementów do -1.
$M^{-1}A = I +$ macierz z zerami na diagonali i ułamkami na reszcie, pierwszy wiersz to $0, a_{12}/a_{11}, a_{13}/a_{11} \ldots$
Żeby uzyskać B odejmujemy I.
$||B||_\infty = \max_i w_i$
$w_i = \sum_j |b_{i,j}| = \sum_{j \neq i} |a_{ij}/a_{ji}| = \frac{1}{|a_{ii}|} \sum_{j \neq i} |a_{ij}| < 1$ zatem norma B jest mniejsza od 1 więc normy są zbieżne. □

5 Metody iteracyjne oparte na normalizacji w przestrzeni Kryłowa

k-ta przestrzeń Kryłowa

$$K_k = r_0, Ar_0, ..., A^{k-1}r_0$$

gdzie $r_k := b - Ax_k$ - reszta na k-tej iteracji

Metoda iteracyjna

- $x_k + 1 \in K_k$ przesunięta o x_0
- $x_k + 1$ normalizuje pewną miarę błędu na $x_0 + K_k$
- Na przykład:
$$||x_k - X^*||_C \leqslant ||y - x^*||_C \forall_{y \in x_0 + K_k}$$
lub
$$||r_k|| \leqslant ||b - Ay||_C \forall_{y \in x_0 + K_k}$$
gdzie $C = C^T > 0$

5.1 Metoda gradientów sprzężonych (CG - Conjugate Gradient) dla macierzy $A = A^T > 0$

5.1.1 Fakty o macierzach symetrycznych i dodatnio określonych

Niech $A = A^T > 0$ (symetryczna i dodatnio określona, a co za tym idzie $x^T Ax > 0$ dla $x \neq 0$). Wtedy:

1. Wartości własne są rzeczywiste a wektory własne są ortogonalne (czyli $A = Q\Lambda Q^T$, gdzie Q jest ortogonalna, a Λ jest diagonalna)
2. $||x||_A := \sqrt{x^T Ax}$ określa normę wektorową (norma energetyczna indukowana przez A)

Iterację metody gradientów sprzężonych definiujemy następująco:

$$x_{k+1} \in x_0 + K_k$$

$$||x_{k+1} - x^*||_A \leqslant ||y - x^*||_A \forall_{y \in x_0 + K_k}$$

Ale przecież potrzebujemy mieć rozwiązanie żeby to zrobić!

Fakt. Można stąd wyprowadzić algorytm iteracyjny, który na podstawie kilku poprzednio wyznaczonych wektorów wyznaczy x_{k+1} kosztem jednego mnożenia przez macierz A i $O(N)$

Twierdzenie. Po k iteracjach metody CG błąd $||x_k - x^*||_A \leqslant 2(\frac{\sqrt{\alpha}-1}{\sqrt{\alpha}+1})^k ||x_0 - x^*||_A$ gdzie $\alpha = \lambda_{max}(A)/\lambda_{min}(A)$.

6 Zagadnienia własne

Dla $A \in R^{NxN}$ znaleźć parę własną (λ, x), że $Ax = \lambda x$ oraz $x \neq 0$. λ pierwiastkiem wielomianu charakterystycznego: $det(A - \lambda I) = 0$ Gdy $A = A^T$ to wartości i wektory własne rzeczywiste, istnieje Q ortogonalna $A = Q * L * Q^T$ (L to tylko lambdy na przekątnej)

3 podstawowe klasy zadań obliczeniowych dla zagadnień własnych:

1. ekstremalne wartości własne (największa, najmniejsza, etc) i odp. wektory (PageRank)

2. wartości własne bliskie zadanej wartości (wieżowce w Japonii)

3. pełne zadanie własne

Wyznaczanie wektora odpowiadającego dominującej wartości własnej (zakładamy że istnieje dokładnie jedna wartość własna że jej moduł ostro większy od innych modułów)

$$\|A_x\| = \|\lambda * x\| = |\lambda| * \|x\| = |\lambda|$$

(bo $\|x\| = 1$)

Metoda potęgowa x_0 startowy o normie 1

$$x_{n+1} = Ax_n$$

$$x_{n+1} := \frac{x_{n+1}}{\|x_{n+1}\|}$$

skąd nazwa:

$$x_{n+1} = Ax_n = AAx_{n-1} = A^2x_{n-1} = \cdots = A^{n+1}x_0$$

nie robić tego w ten sposób, bo A jest duże (ale rzadkie) i będzie coraz mniej rzadkie! Lepiej iteracyjnie, bo tanio mnożyć przez rzadką macierz

Twierdzenie o zbieżności tej metody: Załóżmy, że A diagonalizowalna - istnieje Y nieosobliwe że YAY^{-1} tworzy macierz diagonalną

$Ay_i = \lambda y_i$ gdzie y_i to kolumna Y

$$x_0 = \sum_1^n \alpha_i y_i$$

$$x_n = A^n x_0 = A^{n-1}(Ax_0) =$$

$$= A^{n-1} \sum_1^n \alpha_i y_i =$$

$$= A^{n-1} \sum_1^n \alpha_i \lambda_i y_i =$$

$$= \sum_1^n \alpha_i \lambda_i^n y_i =$$

$$= \lambda_1^n * \sum_1^n \alpha_i (\frac{\lambda_i}{\lambda_1})^n y_i$$

Jeżeli λ_1 dominujące, to $\frac{\lambda_i}{\lambda_1^n} \to 0$ ($\lambda_1 \neq 0$)

Odzyskanie wartości własnej na podstawie przybliżenia (znaleźć takie przybliżenie lambdy że norma przybliżenia $A*x - \lambda*x$ minimalna) - jest to zadanie najmniejszych kwadratów iloraz Rayleigh TODO Transformacje spektrum: 1. Jeżeli λ ww A to $\lambda - \mu$ ww $A - \mu * I$ 2. Jeżeli λ ww A nieosobliwego to $1/(\lambda)$ ww $A^(-1)$

Odwrotna metoda potęgowa na zadania typu 2:

Wartosci wlasne $(A - \mu * I)^{-1}$ to $\frac{1}{\lambda - \mu}$ Kiedy największe? Kiedy μ blisko λ_i to wtedy $\frac{1}{\lambda_i - \mu}$ dominującą ww

RQI raileigh quotient iteration, bardzo szybko zbieżne ale niekoniecznie do najbliższego oryginałowi ww TODO

tak naprawdę metoda potęgowa nie na jednym wektorze a na wszystkich, zazwyczaj słabo działa, modyfikacja "raz dodajemy a raz odejmujemy"

3. pełny problem - metoda QR

Skorzystamy z następujących faktów z GAL-u:

- Macierze A, B są podobne jeśli istnieje macierz M spełniająca $A = MBM^{-1}$
- Ortogonalna macierz Q spełnia $Q^{-1} = Q^T$
- Macierze podobne mają te same wartości własne
- W macierzy trójkątnej wartości własne są na przekątnej

Zatem jeśli sprowadzimy macierz A do podobnej do niej macierzy trójkątnej, to będziemy znać wszystkie wartości własne A. Można wykorzystać do tego rozkład QR. Definiujemy metodę iteracyjną:

$$A_{k+1} = R_k Q_k, \text{ gdzie } A_k = Q_k R_k$$

Aby nie wyznaczać całego rozkładu można rozpisać to:

$$A_{k+1} = R_k Q_k = Q_k^T A_k Q_k$$

7 Wykład 5 - Arytmetyka zmiennoprzecinkowa

7.1 Notacja wykładnicza

Np. 6.63×10^{-34} jest przybliżeniem stałej Plancka. U nas będzie tak:

$$x = -1^s m \beta^e$$

gdzie:

1. β - podstawa (typowo $\beta = 2$)

2. e - wykładnik spełniający $e_{min} \leqslant e \leqslant e_{max}$
3. m - mantysa, $1 \leqslant m < \beta$
4. $s \in \{0, 1\}$ - znak liczby

Tak można zapisać każdą liczbę rzeczywistą. Jak to zrobić żeby zmieściło się? Zakładamy, że $m = (f_0 \cdot f_1 \cdot f_2 \cdot f_3 \cdot \ldots \cdot f_{p-1})_\beta$, gdzie $f_i \in \{0, 1, \ldots, \beta - 1\}$ oraz $\_ \cdot \_$ jest konkatenacją. To oznacza, że m jest liczbą w systemie o podstawie β.

Zatem:

$$x = -1^s \cdot (\sum_{i=0}^{p-1} f_i \beta^{-i}) \beta^e$$

Zakres wykładników określa nam

7.2 Liczby znormalizowane (maszynowe)

$$m = (1 \cdot f_1 \cdot f_2 \cdot f_3 \cdot \ldots \cdot f_{p-1})_2$$

To nam daje $1 \leqslant m < 2$.

Reprezentacja liczby maszynowej w pamięci Musi być za pomocą sekwencji bitów. Jak? Do e dodajemy *bias*, żeby nie musieć pamiętać znaku.

$$\left| \; s \; \right| \quad e + bias \quad \left| \; f_1 \quad f_2 \quad f_3 \quad \ldots \quad f_{p-1} \; \right|$$

$e + bias$ będzie zatem miało $size - p$ bitów.

7.3 Standard arytmetyki zmiennoprzecinkowej

IEEE-754 - (pierwsza wersja - 1985, najnowsza 2008) definiuje następujące typy (dla $\beta = 2$):

Nazwa potoczna	p	e_{min}	e_{max}	$size$	$bias$	Oficjalna nazwa
half precision	11	-14	15	16	15	binary16
single precision	24	-126	127	32	127	binary32
double precision	53	-1022	1023	64	1023	binary64
quad precision	113	-16382	16383	128	16383	binary128
double extended precision	64	-16382	16383	80	16383	IA-32

Kiedy dodamy *bias* do e_{min} dostajemy 1, a nie 0! Zarezerwowane wartości $e + bias$ to "0"^{n} oraz "1"^{n}.

Liczba	Znak	$e + bias$	mantysa $(f_1 f_2 f_3 \ldots f_{p-1})$
$+0$	0	$0 \ldots 0$	$0 \ldots 0$
-0	1	$0 \ldots 0$	$0 \ldots 0$
$+\infty$	0	$1 \ldots 1$	$0 \ldots 0$
$-\infty$	1	$1 \ldots 1$	$0 \ldots 0$
NaN	0	1...1	$\{0,1\}^* 1 \{0,1\}^*$

A skąd te liczby? $\frac{1}{+0} = \infty$, $\frac{1}{-0} = -\infty$. Co za tym idzie, $\frac{1}{\infty} = 0$. A skąd NaN? $\frac{0}{0}$, $\infty - \infty$ itd.

Orientacyjne zakresy liczb znormalizowanych:

	najmniejsza	największa
binary32	$\sim 10^{-38}$	$\sim 10^{38}$
binary64	$\sim 10^{-308}$	$\sim 10^{308}$

Bardzo mały system liczb maszynowych

$$\beta = 2, p = 3, e_{min} = -1, e_{max} = 2$$

Liczby znormalizowane są postaci:

$$x = -1^s \cdot (1 f_1 f_2)_2 2^e$$

$e = 0$:
1, 1.25, 1.5, 1.75
$e = 1$:
2, 2.5, 3, 3.5
$e = 2$:
4, 5, 6, 7, 8
$e = -1$:
0.5, 0.625, 0.75, 0.875

Ale nie ma nic między 0 i 0.5!

Liczby zdenormalizowane

$$x = -1^s (0 f_1 f_2 \dots f_{p-1}) 2^{e_{min}}$$

Wypełniają pustkę wokół zera.

liczba	znak	$e + bias$	mantysa
+subnormal	0	...	$\{0,1\}^* 1 \{0,1\}^*$
-subnormal	1	...	$\{0,1\}^* 1 \{0,1\}^*$

Co w naszym małym systemie daje dodatkowe 3 liczby wokół 0.

7.4 Działania arytmetyczne na liczbach IEEE-754

1. Reprezentacja liczby rzeczywistej x : $fl(x)$ - najbliższa x w sensie zaokrąglenia do najbliższej (default, da się zmienić by było do zera lub od zera)
2. Jeśli x jest reprezentowana przez znormalizowaną l. maszynową, to zachodzi $\frac{|fl(x)-x|}{|x|} \leqslant 2^{-p}$, gdzie p jest ilością bitów w mantysie - ν - precyzja arytmetyki. Inaczej mówiąc, $fl(x) = x(1+\varepsilon), |\varepsilon| \leqslant \nu$
3. Dla $\cdot \in \{+, -, \times, \div\}$ standard wymaga, by:

$$\underset{\substack{\text{wynik obliczenia } a \cdot b \text{ w arytmetyce } fl \\ \text{gdy } a, b \text{ są liczbami maszynowymi}}}{fl(a \cdot b)} = \underset{\substack{\text{reprezentacja dokładnego} \\ \text{wyniku } a \cdot b \text{ w arytmetyce } fl}}{fl(a \cdot b)}$$

4. Zatem możemy uznać, że $fl(a \cdot b) = (a \cdot b)(1+\varepsilon), |\varepsilon| \leqslant \nu$, z tym że ε jest inny dla różnych działań, oraz pod warunkiem, że wynik jest reprezentowalny.

8 Wykład 6 - 15/11

Na kolokwium nie będzie arytmetyki zmiennoprzecinkowej ani dzisiejszego wykładu.

8.1 Jeszcze trochę fl - a

Arytmetyka zmiennopozycyjna nie jest zwykłą arytmetyką.

- Nieprawdą jest że $\sim (a == b) \iff a$ nie jest b, bo NaN jest nieporównywalny
- nie musi być, że jeśli $1+x == 1$ to $x = 0$. (np. dla $x =$ precyzja arytmetyki)
- unikać testów z równością
- wynik $x - x + 1$ może być różny w zależności od kolejności działań.
- stawiać nawiasy by wymusić kolejność
- kumulacja błędów zaokrągleń w obliczeniach
- wzmocnienie wcześniejszych błędów może być katastrofalne

Przykład (redukcja cyfr przy odejmowaniu) . Obliczmy $s := x+y, x, y \in \mathbb{R}$. Ale zamiast x, y w komputerze mamy tylko $\tilde{x} = fl(x) = x(1+\varepsilon_x), \tilde{y} = fl(y) = y(1+\varepsilon_y)$! Zamiast s obliczymy w fl wartość $\tilde{s} = fl(\tilde{x}+\tilde{y}) = (\tilde{x}+\tilde{y})(1+\delta)$. Zatem błąd wyniku to $\frac{|s-\tilde{s}|}{|s|} = \frac{|x+y-fl(\tilde{x}+\tilde{y})(1+\delta)|}{|x+y|} = x(1+\varepsilon_x)(1+\delta)+y(1+\varepsilon_y)(1+\delta) = x(1+E_x)+y(1+E_y)$ gdzie $E_x = \varepsilon_x + \delta + \varepsilon_x\delta$ Zatem $|E_x| \leqslant |\varepsilon_x| + |\delta| + |\varepsilon_x\delta| \leqslant 2ni + ni^2 \approx 2ni$ Zatem błąd $= \frac{|s-\tilde{s}|}{|s|} = \frac{|x+y-x(1+E_x)-y(1+E_y)|}{|x+y|} = \frac{|xE_x-yE_y|}{|x+y|} \leqslant 2\frac{|x|+|y|}{|x+y|}ni$ Zauważmy, że:

- jeśli x, y - tego samego znaku, to $\frac{|x|+|y|}{|x+y|} = 1$ i wtedy błąd szacuje się przez $2ni$
- jeśli $x \approx y$ to $\frac{|x|+|y|}{|x+y|} >> 1$ i wtedy błąd ograniczamy przez bardzo dużą liczbę

8.2 Katastrofy spowodowane fl - em

- zatonięcie całej platformy wiertniczej spowodowanej niewłaściwym użyciem pakietu numerycznego.

- Rakiety patriot - im dłużej stały włączone, tym więcej celności traciły. Systemowy zegar liczył ticki co $\frac{1}{10}$ sekundy. Więc $t * \frac{1}{10}$ daje tym gorszy błąd im więcej tickow.
- rakiety arian - double został przepisany na short inta
- rakieta nie poleciała na marsa bo system imperialny vs. metryczny
- partia zielonych w niemczech dostała 4.76% głosów ale excel zaokrąglił w górę i przeszli próg wyborczy
- indeks giełdowy jechał w dół a powinien był w górę, to przez częstą aktualizację z błędem

8.3 Błędy w obliczeniach numerycznych

Mamy zadanie obliczeniowe, algorytm i dane. Niech $P_i : \mathbb{R}^n \supseteq D \to \mathbb{R}^m$
Problem obliczeniowy: mając dane $x \in D$, wyznacz $y = P(x)$.

Przykłady

- Dla zadanej funkcji rzeczywistej oblicz $y = f(x)$
- dla zadanych $A \in \mathbb{R}^{n \times n}$ oblicz $y = Ax$ $(P = Ax)$
- dla $A \in \mathbb{R}^{n \times n}$ nieosobliwej oraz $b \in \mathbb{R}^n$ oblicz $b \in \mathbb{R}^n$ t. że $Ay = b$ $(P = A^{-1}x)$
- dla $f : \mathbb{R} \to \mathbb{R}$ oblicz y t. że $f(y) = 0$ $(P(x) = f^{-1}(x))$

Zwykle szukamy nie algorytmu dla danego zadania, tylko klasy zadań.

Definicja Błąd bezwzględny: $\|x - \tilde{x}\|$. Błąd względny: $\frac{\|x-\tilde{x}\|}{\|x\|}(x \neq 0)$. Nieuniknione w fl są błąd reprezentacji danych oraz wyników Algorytm jest silnie numerycznie poprawny (backward stable) jeśli wynik jego działania w fl można zinterpretować jako wynik zadania obliczeniowego (w arytmetyce dokładnej) na danych lekko zaburzonych.
Algorytm jest numerycznie poprawny, jeśli wynik też może być zaburzony na poziomie reprezentacji.

Uwarunkowanie zadania numerycznego Jak zaburzenie danych wpływa na zaburzenie algorytmu?

$$cond_{abs}(P, x) := \sup_{\Delta \neq 0, \|\Delta\| < \delta} \frac{\|P(x + \Delta) - P(x)\|}{\|\Delta\|}$$

- uwarunkowanie zadania P w punkcie x

Innymi słowy, $\|P(x+\Delta) - P(x)\| \leqslant cond_{abs}(P,x)\|\Delta\|$, gdzie Δ jest najmniejsza możliwa dla danych zaburzeń. Można to idealizować i rozważać przypadek graniczny, gdy $\|\Delta\| \to 0$

$$cond_{abs}(P,x) := \lim_{\|\Delta\|\to 0} \frac{\|P(x+\Delta) - P(x)\|}{\|\Delta\|}$$

Analogicznie, uwarunkowanie względnie:

$$\frac{\|P(x+\Delta) - P(x)\|}{\|P(x)\|} \leqslant cond_{rel}(P,x)\frac{\|\Delta\|}{\|x\|}$$

Przykład

- $P(x) = f(x)$: $cond_{abs}(P,x) = |f'(x)|$
- $Ax = b, A\tilde{x} = \tilde{b}$

Jak błąd wzgl. x zależy od błędu wzgl. b? niech błąd względny $b \leqslant \varepsilon$, wtedy

$$\tilde{x} = A^{-1}\tilde{b} = A^{-1}(b+\Delta) = A^{-1}b + A^{-1}\Delta$$

zatem $\|x - \tilde{x}\| \leqslant \|A^{-1}\Delta\| \leqslant \|A^{-1}\|\|\Delta\|$ stąd

błąd $\leqslant \|A^{-1}\|\frac{\|\Delta\|}{\|b\|}\frac{\|b\|}{\|x\|} <= \|A^{-1}\|\varepsilon\frac{\|Ax\|}{\|x\|} <= \|A^{-1}\|\|A\|\|\varepsilon = cond(A)varepsilon$

Można pokazać ogólniej, że jeśli $Ax = b$ oraz $\tilde{A}\tilde{x} = \tilde{b}$, gdzie zaburzenia wzgl. b i A są na poziomie dostatecznie małego epsilona, to błąd x da się oszacować przez $4cond(A)\varepsilon$.

Część II
Ćwiczenia

9 Układy nadokreślone - kontynuacja

9.1 Zadanie 1.

Macierz Hessenberga - to macierz trójkątna górna, z tym że niezerowe elementy mogą być jeden element pod diagonalą.

$$\begin{bmatrix} x & \dots & & & & x \\ x & x & \dots & & & x \\ & x & x & \dots & & x \\ & & \ddots & & & x \\ & & & x & x & x \\ & & & & x & x \end{bmatrix}$$

Jak najmniejszym kosztem znaleźć rozkład QR tej macierzy?
Metodą Householdera? Nie ma jak wykorzystać zer na dole.

9.1.1 Obroty Givensa - przypomnienie

1. G_{ij} - macierz Givensa
2. $b = G_{ij}a$
3. $b_j = 0$
4. $\cos\phi = \frac{a_i}{\sqrt{a_i^2+a_j^2}}$
5. $\sin\phi = \frac{a_j}{\sqrt{a_i^2+a_j^2}}$

9.1.2 Zamiana macierzy Hessenberga w górnotrójkątną obrotami Givensa

$$(G_{ij}a)_j = -\frac{a_i a_j}{\sqrt{a_i^2+a_j^2}} + \frac{a_i a_j}{\sqrt{a_i^2+a_j^2}} = 0$$

$$G_{n-1n}\dots G_{ii+1}\dots G_{12}A = R$$

9.1.3 Koszt

Robimy $n-1$ iteracji.
Dla $G_{i\ i+1}$ trzeba wykonać jeden pierwiastek, $w_i = cw_i + sw_{i+1}$ oraz $w_{i+1} = -sw_i + cw_{i+1}$, łącznie $4(n-1)$ mnożeń.
Wszystko razem: $4\sum_{i=1}^{n-1} n-i = 4\sum_{i=1}^{n-1} i = \frac{4n(n-1)}{2} \sim 2n^2$.

9.2 Zadanie 2.

Dane są punkty $(-1,-1),(0,2),(1,0),(2,1)$.
Znajdź prostą $y = ax + b$ najlepiej przybliżającą te punkty (w sensie LZNK).
Zadane punkty oznaczamy jako (x_i, y_i).
Zatem to, co chcemy zminimalizować to $y(x_i) - y_i$.
Policzmy normę: $\min_{a,b}\sum_{i=1}^{4}(y(x_i)-y_i)^2$
Niewiadome to a oraz b, więc niech:

$$z = \begin{bmatrix} a \\ b \end{bmatrix} \quad d = \begin{bmatrix} y_i \end{bmatrix}_{i=1,2,3,4} \quad A = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \end{bmatrix}$$

Teraz wystarczy użyć LZNK aby obliczyć $\min ||Az - d||_2$:

TODO: policzyć to

Uwaga: w ten sposób odległość między punktami a prostą liczymy w pionie, a nie najbliższą (to dobrze, tak działą LZNK).

Uwaga 2: LZNK nie działa dla równania $y = a + e^{bx}$, ale dla $y = a + be^x$ już tak!

10 Normy

Przypomnienie definicji: Norma $|| \cdot || : V \to \mathbb{R}^+$ spełnia następujące warunki:

1. $||u + v|| \leqslant ||u|| + ||v||$
2. $||\alpha v|| = ||av||$
3. $||v|| = 0 \implies v = 0$ - wektor zerowy

p-te normy wektorowe:

$$||x||_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

$$||x||_\infty = \max_i |x_i|$$

Normy macierzowe Niech $A \in \mathbb{R}^{n \times n}$.
Macierzowe normy indukowane są postaci

$$||A|| = \sup_{x \neq 0} \frac{||Ax||}{||x||} = \sup_{||x||=1} ||Ax||$$

p-te normy macierzowe

$$||A||_p = \sup_{||x||_p=1} ||Ax||_p, p = 1, 2, \ldots, \infty$$

wszystkie poza $1, 2, \infty$ zwykle się pomija

10.1 Własności norm indukowanych macierzy

1. $||Ax|| \leqslant ||A||||x||$ - z definicji mamy $||A|| \geqslant \frac{||Ax||}{||x||}$
2. $||AB|| \leqslant ||A||||B||, A, B \in \mathbb{R}^{n \times n}$ - bo

$$||ABx|| \leqslant ||A||||Bx|| \leqslant ||A||||B||||x||$$

oraz

$$||AB|| = \sup_{x \neq 0} \frac{||ABx||}{||x||} \leqslant ||A||||B||$$

Fakt. W przestrzeniach skończonego wymiaru wszystkie normy spełniają równanie: $\exists_{c_1,c_2>0}\forall_x c_1||x||_1 \leqslant ||x||2 \leqslant c_2||x||_1$, gdzie normy są dowolne (niekoniecznie pierwsza i druga)

10.2 Zależności między normami

Niech $x \in \mathbb{R}^n$, a normy będą p-te.

$$||x||_1^2 = (\sum_i |x_i|)^2 \geqslant ||x||_2^2$$

$$||x||_1 \leqslant \alpha||x||_2$$

Jakie α wybrać?

$$||x||_1 \geqslant ||x||_\infty$$

$$n||x||_\infty \geqslant ||x||_1$$

$$||x||_\infty \leqslant ||x||_2$$

$$\sqrt{n}||x||_\infty \geqslant ||x||_2$$

$$||x||_1 \leqslant n||x||_\infty \leqslant n||x||_2$$

Zatem $\alpha = n$.

Nierówności $\frac{1}{n}||A||_2 \leqslant \frac{1}{\sqrt{n}}||A||_\infty \leqslant ||A||_2 \leqslant \sqrt{n}||A||_1 \leqslant n||A||_2$

10.3 Wzory na normy macierzowe

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

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

Zatem $||A^T||_1 = ||A||_\infty$.

Norma druga (spektralna)

$$||A||_2 = \max_{\lambda \in \delta(A^T A)} \sqrt{\lambda}$$

Gdzie $\delta(M)$ jest zbiorem wartości własnych macierzy M.
Jeśli Q ortogonalna: $||Q||_2 = 1$, co za tym idzie $||I||_2 = 1$ λ- wartość własna oraz v - wektor własny spełnieją $Av = \lambda v$

Norma Frobeniusa (Euklidesowa)

$$||A||_F = \sqrt{\sum_{i,j} |a_{ij}|}$$

Nie jest normą indukowaną, bo dla wszystkich norm indukowanych zachodzi $||I|| = \sup_{x \neq 0} \frac{||Ix||}{||x||} = 1$, a $||I||_F = \sqrt{n}$ - zatem nie pochodzi od drugiej normy wektorowej!

11 Ćwiczenia 26/10

11.1 Dalsze własności norm

$$||A||_2 = \max_{\lambda \in \delta(A^T A)} \sqrt{\lambda}$$

Jeśli $A = A^T$ to $||A||_2 = \max_{\lambda \in \delta(A)} |\lambda|$.

Twierdzenie. Jeśli λ jest wartością własną A, to λ^2 jest wartością własną A^2.

Dowód. $Av = \lambda v \implies A^2 v = \lambda A v = \lambda^2 v$ □

$$||A||_2 = \sup_{||x||=1} ||Ax||_2$$

$$||Ax||_2 = \sqrt{(Ax)^T Ax} = \sqrt{x^T A^T A x} = (*)$$

$A^T A$ jest macierzą symetryczną, oraz $A^T A = Q^T \Lambda Q$, gdzie Λ jest macierzą diagonalną $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ oraz Q jest macierzą ortogonalną wektorów własnych.

UZUPEŁNIĆ TODO

12 Ćwiczenia 9/11

12.1 Metoda Richardsona - kontynuacja zadania

Metoda Richardsona - $x_{k+1} = x_k + \tau(b - Ax_k)$ Szukaliśmy parametru τ t. że metoda Richardsona jest zbieżna. A ma wartości własne $\lambda_1, \ldots, \lambda_n > 0$. Jest zbieżna dla $\tau \in (0, \frac{2}{\lambda_{max}})$. Dla jakich τ jest zbieżna najszybciej?

$$\rho(I - \tau A) = \max\{|1 - \tau\lambda_{min}|, |1 - \tau\lambda_{max}|\}$$

Powiedzieliśmy, że

$$||x_{k+1} = x^*|| \leqslant ||I - Q^{-1}A||||x_k - x^*||$$

Zatem szukamy τ realizującego:

$$\arg\min_{\tau\in(0,\frac{2}{\lambda_{max}})}\rho(I-\tau A)$$

Czyli:

$$\arg\min_{\tau\in(0,\frac{2}{\lambda_{max}})}\max\{|1-\tau\lambda_{min}|,|1-\tau\lambda_{max}|\}$$

Pierwsz funkcja ma miejsce zerowe w $\frac{1}{\lambda_{min}}$, a druga w $\frac{1}{\lambda_{max}}$. Można narysować obie funkcje, one przecinają się w zwykle dwóch punktach, czyli $|1-\tau\lambda_{min}| = |1-\tau\lambda_{max}|$. Zatem albo $\tau = 0$, co daje nam rozbieżność, albo $\tau = \frac{2}{\lambda_{max}+\lambda_{min}}$ Żeby wystartować z metodą, to wystarczyłoby ograniczenie górne na λ_{max}

12.2 Metoda Gaussa-Seidela

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

Wykaż, że metoda Gaussa-Seidela jest zbieżna dla macierzy A.

Fakt. Metoda Gaussa-Seidela jest zbieżna dla macierzy diagonalnie dominującej ($\forall_i |a_{ii}| > \sum_{j\neq i} |a_{ij}|$). Metoda iteracyjna jest zbieżna, jeśli promień spektralny macierzy jest mniejszy od 1.

Promień spektralny

- $$\rho(I-Q^{-1}A)$$
- kres dolny po wszystkich normach indukowanych:

$$\inf_{||\cdot||\text{jest indukowana}} ||I-Q^{-1}A||$$

Wystarczy pokazać, że dla pewnej normy indukowanej $||\cdot||$(jakiej?) zachodzi:

$$||I-Q^{-1}A|| < 1$$

Dowód. Tutaj,

$$Q = \begin{bmatrix} 2 & & & & & & \\ -1 & 2 & & & & & \\ & -1 & 2 & & & & \\ & & \ddots & \ddots & & & \\ & & & \ddots & \ddots & & \\ & & & & -1 & 2 & \\ & & & & & -1 & 2 \end{bmatrix}$$

Korzystamy z tego, że $Q^{-1}Q = I$, i prostym spostrzeżeniem jest:

$$Q = \begin{bmatrix} 2^{-1} & & & & & & \\ 2^{-2} & 2^{-1} & & & & & \\ \vdots & & 2^{-1} & & & & \\ \vdots & & \ddots & \ddots & & & \\ \vdots & & & \ddots & \ddots & & \\ \vdots & & & & & 2^{-1} & \\ 2^{-n} & & & & & 2^{-2} & 2^{-1} \end{bmatrix}$$

Teraz

$$Q^{-1}A = \begin{bmatrix} 1 & -\frac{1}{2} & & & & & \\ 0 & \frac{3}{4} & -\frac{1}{2} & & & & \\ & -2^{-3} & \frac{3}{4} & -\frac{1}{2} & & & \\ & \vdots & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \frac{3}{4} & -\frac{1}{2} \\ 0 & -2^{-n} & \dots & \dots & & -2^{-3} & \frac{3}{4} \end{bmatrix}$$

$$G = I - Q^{-1}A = \begin{bmatrix} 1 & \frac{1}{2} & & & & & \\ 0 & \frac{1}{4} & \frac{1}{2} & & & & \\ & 2^{-3} & \frac{1}{4} & \frac{1}{2} & & & \\ & \vdots & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \frac{1}{4} & \frac{1}{2} \\ 0 & 2^{-n} & \dots & \dots & & 2^{-3} & \frac{1}{4} \end{bmatrix}$$

$||G||_1 = \sum_{i=1}^{n} (\frac{1}{2})^i = 1 - \frac{1}{2^n} < 1$ □

12.3 Błędy numerycznych rozwiązań układów równań

Rozwiązujemy układ $Ax^* = b$, x^* jest dokładnym rozwiązaniem, x - wynik obliczeń numerycznych. Wtedy $x^* = x + A^{-1}(b - Ax) = x + A^{-1}r = x + e$,

gdzie r jest wektorem residualnym ($r = b - Ax$), a e - błędem. Rozwiązując układ równań $Ae = r$ otrzymamy poprawkę rozwiązania. Tylko czy to ma sens? Układy z macierzą A już umiemy łatwo rozwiązywać, tylko to wciąż nie będzie idealne, bo znów mamy przybliżenie.

Iteracyjne poprawianie rozwiązań

$$\begin{array}{rcl} x^0 & = & x \\ x^{(k+1)} & = & x^{(k)} + A^{-1}r^{(k)}, \quad k = 0, 1, \ldots \end{array}$$

Żeby poprawianie poprawiało, obliczanie wektora residualnego musi być wykonane w jak największej precyzji.

12.4 Wartości i wektory własne

λ, v - para własna dla A, spełnia $Av = \lambda v$

$$||Av|| = ||\lambda v||$$
$$||Av|| = |\lambda|||v||$$
$$\frac{||Av||}{||v||} = |\lambda|$$

Czyli dla dowolnej wartości własnej i dowolnej normy indukowanej zachodzi:

$$\sup_{v \neq 0} \frac{||Av||}{||v||} \geqslant |\lambda| \implies |\lambda| \leqslant ||A||$$

To się przydaje w metodzie Richardsona.

Twierdzenie Gerszgorina - o lokalizacji wartości własnych. Każda wartość własna macierzy A leży co najmniej w jednym z kół na płaszczyźnie zespolonej:

$$D_i = \{z \in \mathbb{C} : |z - a_{ii}| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}|\} \text{ dla } i = 1, 2, \ldots, n$$

Dowód. Weźmy dowolną wartość własną λ z wektorem v. Pokażemy, że istnieje wiersz i macierzy A t. że $\lambda \in D_i$. Niech $||v||_\infty = 1$ co implikuje że $|v_i| = 1$. Teraz $(Av)_i = \lambda v_i = \sum_{j=1}^{n} a_{ij} v_j$

$$(\lambda - a_{ii})v_i = \sum_{j=1, j\neq i}^{n} a_{ij} v_j$$
$$|(\lambda - a_{ii})v_i| = |\sum_{j=1, j\neq i}^{n} a_{ij} v_j|$$
$$|(\lambda - a_{ii})v_i| \leqslant \sum_{j=1, j\neq i}^{n} |a_{ij}||v_j| \leqslant \sum_{j\neq i} |a_{ij}|$$

□

Wniosek. Macierz diagonalnie dominująca nie ma zerowych wartości własnych, czyli jest nieosobliwa.

13 Ćwiczenia 16/11

13.1 Wyznaczanie wektorów i wartości własnych - c.d.

Metody wyznaczania - dla $Av_i = \lambda_i v_i$

- Metoda potęgowa - Zaczynamy od wektora x_0 i mnożymy z lewej przez A. Zakładamy $|\lambda_1| > |\lambda_2| \geqslant \ldots \geqslant |\lambda_n|$ oraz A ma n wektorów własnych. Robimy:

$$x_0, \|x_0\|_2 = 1$$

$$y_{k+1} = Ax_k, x_{k+1} = \frac{y_{k+1}}{\|y_{k+1}\|_2}$$

  Na koniec wartość własną dostajemy:

$$\sigma_k = \frac{x_k^T A x_k}{x_k^t x_k} = x_k^T y_{k+1}$$

- Odwrotna metoda potęgowa.

$$x_0, \|x_0\|_2 = 1$$

$$(A - \mu I) y_{k+1} = x_k$$

$$x_{k+1} = \frac{y_{k+1}}{\|y_{k+1}\|_2}$$

$$\sigma_k = x_k^T y_{k+1}$$

  Jeśli $\mu = 0$ to $|\lambda_n|$ musi być ostro mniejsze, bo w normalnej $|\lambda_0|$ musiało być ostro większe, a $v_i = \lambda_i A^{-1} v_i \implies A^{-1} v_i = \frac{1}{\lambda_i} v_i$ Wartości własne $(A - \mu I)^{-1}$ to $\frac{1}{\lambda_i - \mu}$

- Co jeśli $|\lambda_1| = |\lambda_2| > |\lambda_3| \geqslant \ldots$? Dostaniemy wektor będący kombinacją liniową wektorów własnych odpowiadających λ_1 i λ_2.

- Mamy $|\lambda_1| > |\lambda_2| > |\lambda_3| \geqslant \ldots \geqslant |\lambda_n|$. Wyznaczyliśmy λ_1 z metody potęgowej. Jak wyznaczyć $|\lambda_2|$? Przyjmujemy $x_0 = \sum_{i=2}^{n} \alpha_i v_i$ jeśli $\{v_i\}$ baza ortogonalna: wybieramy $x_0 \perp v_1$. Wtedy co kilka kroków trzeba x_k ortogonalizować, żeby zachować ortogonalność utraconą ze względu na błędy zmiennoprzecinkowe.

Dane są:

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}, x_0 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

Czy metoda potęgowa dla A jest zbieżna dla x_0? Policzmy ręcznie i zobaczmy XDD
Wartości własne: rozwiążmy $\det(A - \lambda I) = 0$

$$\det(A - \lambda I) = (2 - \lambda)(2 - \lambda) - 1 = 0$$

$$\lambda_1 = 3, \lambda_2 = 1$$

Czyli wektory własne:

$$v_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, v_2 = \begin{bmatrix} 2 \\ 1 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 1 \\ -1 + \varepsilon \end{bmatrix}$$

$$Ax_0 = \begin{bmatrix} 1 + \varepsilon \\ -1 + 2\varepsilon \end{bmatrix}$$

Z epsilona zrobiły się dwa epsilony, w następnym 4 i 5, pottem 13 i 14. Iloraz współczynników przy epsilonach zbiega do jedynki, więc wynikowy wektor będzie w dobrym kierunku, ale możemy zbiec do wektora do którego mieliśmy dalej.

13.2 Metoda QR

$$A_0 = A$$

$$A_k = Q_k R_k - \text{rozkład QR}$$

$$A_{k+1} = R_k Q_k - \text{mnożymy na odwrót}$$

Wyznaczanie rozkładów jest kosztowne, ale da się łatwiej używając macierz Hessenberga która jest podobna do macierzy A (czyli ma te same wartości własne), a metoda QR zachowuje hessenbergowość.

Dowód. Mamy pokazać, że jeśli macierz Hessenberga $A = QR$ to macierz RQ też jest Hessenberga.

1. Jeśli A - Hessenberga to Q też:

$$\begin{bmatrix} x & \dots & & & x \\ x & x & \dots & & x \\ & x & x & \dots & x \\ & & \ddots & & x \\ & & & x & x \end{bmatrix} = \begin{bmatrix} & & & \\ & & & \\ & & & \\ & & & \\ & & & \end{bmatrix} \begin{bmatrix} x & \dots & & & x \\ & x & \dots & & x \\ & & x & \dots & x \\ & & & \ddots & x \\ & & & x & x \end{bmatrix}$$

   Każda kolejna kolumna Q to kolumna A odpowiednio przemnożona.

2. Iloczyn RQ tj, trójkątna górna razy Hessenberga daje macierz Hessenberga. Rozpisać tak samo.

□

Jak sprowadzić macierz do postaci Hessenberga przy pomocy podobieństw? Używamy Householdera. Dlaczego by nie do trójkątnej? Pomnożymy z lewej strony i dostaniemy zera w pierwszej kolumnie, ale potem pomnożymy z prawej żeby było podobieństwo i rozwali nam to pierwszą kolumnę. Jeśli zrobimy tak, żeby nie tykać pierwszego wiersza, to z prawej ten sam householder nie zmieni nam pierwszej kolumny. Dowód poprawności:

$$A_k = \begin{bmatrix} B & F^T \\ D & E \end{bmatrix}$$

Gdzie B jest Hessenberga $k \times k$, a D ma zera we wszystkich kolumnach poza ostatnią, w której jest wektor d. Dobieramy $\tilde{H}_k$ tak aby $\tilde{H}_k d = \alpha \tilde{e_1}$ Wtedy $H_k = \begin{bmatrix} I & 0 \\ 0 & \tilde{H}_k \end{bmatrix}$ Teraz

$$A_{k+1} = H_k A_k H_k = A_k = \begin{bmatrix} B & F^T \\ \tilde{H}_k D & \tilde{H}_k E \end{bmatrix} H_k = \begin{bmatrix} B & F^T \tilde{H}_k \\ \tilde{H}_k D \tilde{H}_k & \tilde{H}_k E \tilde{H}_k \end{bmatrix}$$

W ten sposób otrzymujemy $H_{n-2} \dots H_1 A H_i \dots H_{n-2} = T$, T jest postaci Hessenberga i jest podobna do A.

14 Ćwiczenia n+2

14.1 Arytmetyka fl

W arytmetyce fl nie ma łączności w działaniach, np. $1 + \gamma + \gamma, \gamma = \frac{3}{2}10^{-16}$ Precyzja arytmetyki to około $2.2 \cdot 10^{-16}$. Policzmy:

$$fl(1 + \gamma + \gamma) = fl((1 + \gamma) + \gamma) = fl(fl(1 + \gamma) + \gamma)$$

$$fl(1 + \gamma) = 1$$

, bo najmniejsza reprezentowalna liczba większa od 1 to $1 + \varepsilon$. Zatem wynikiem jest 1. Ale co jeśli inaczej znawiasujemy? Mnożenie przez 2 jest dokładne.

$fl(1 + \gamma + \gamma) = fl(1 + (\gamma + \gamma)) = fl(1 + fl(\gamma + \gamma)) = fl(1 + 2\gamma) \neq 1$, bo $2\gamma > \varepsilon$

14.2 Utrata cyfr znaczących przy odejmowaniu

$$x = 0.3721478693,\ y = 0.3720230572,\ x - y = ?$$

Mamy system który obsługuje tylko 5 cyfr znaczących.

$$rd(x) = 0.37215,\ rd(y) = 0.37202$$

W naszym systemie uzyskamy wynik $fl(x-y) = 0.00013$, podczas gdy w idealnej arytm. $x - y = 0.0001248121$. Błąd bezwględny nie jest duży, ale względny: $\frac{fl(x-y)-(x-y)}{x-y} \simeq 4 \cdot 10^{-2}$. Arytmetyka ma dokładność rzędu 10^{-5}, a nasz błąd jest rzędu aż 10^{-2}! Jest to związane z tym, że liczby które od siebie odejmujemy są bliskie sobie.

Jak policzyć $a^2 - b^2$? Wersja 1:

```
s := a * a;
t := b * b;
w := s - t;
```

Wersja 2:

```
u := a + b;
v := a - b;
w := u * v;
```

Mamy zagwarantowane, że wszystkie działania spełniają

$$fl(x \cdot y) = (x \cdot y)(1 + \nu), |\nu| \leqslant \varepsilon, \cdot \in \{+, -, *, /\}$$

Ale to zakłada, że liczby są dokładnie reprezentowane! Jakie są błędy naszych "algorytmów"?

1. $fl(a^2 - b^2) = [a^2(1 + \delta_1) - b^2(1 + \delta_2)](1 + \delta_3) = a^2(1 + \nu_1) - b^2(1 + \nu_2)$, gdzie $\nu_1 = \delta_1 + \delta_3 + \delta_1\delta_3, \nu_2 = \delta_2 + \delta_3 + \delta_2\delta_3$. $\frac{fl(a^2-b^2)-(a^2-b^2)}{a^2-b^2} = \frac{a^2\nu_1 - b^2\nu_2}{a^2-b^2}$ Błąd względny zależy od danych! Jeśli a^2, b^2 są bliskie i duże i dodatkowo błędy ν_1, ν_2 są przeciwnego znaku, to błąd względny jest DUŻY!
2. $fl(a^2 - b^2) = [(a + b)(1 + \delta_1)(a - b)(1 + \delta_2)](1 + \delta_3)$
   $= (a^2 - b^2)(1 + \delta_1)(1 + \delta_2)(1 + \delta_3) = (a^2 - b^2)(1 + E)$
   $E = \delta_1 + \delta_2 + \delta_3 + \delta_1\delta_2 + \delta_2\delta_3 + \delta_1\delta_3 + \delta_1\delta_2\delta_3$ Poza pierwszymi trzema składnikami, reszta jest grubo poniżej dokładności arytmetyki, zatem $|E| \tilde{\leqslant} 3\varepsilon$!

Zatem licząc różnicę kwadratów, zawsze liczmy wzorem skróconego mnożenia.

14.3 Uwarunkowanie zadania

Wcześniej były dokładne dane i niedokładne obliczenia, teraz mamy dokładne obliczenia na niedokładnych danych. Policzmy uwarunkowanie zadania różnicy kwadratów.
Zaburzone dane:

$$\tilde{a} = a(1 + \delta_1),\ \tilde{b} = b(1 + \delta_2),\ |\delta_1| \leqslant \varepsilon$$

$$|\frac{\tilde{a}^2 - \tilde{b}^2 - (a^2 - b^2)}{a^2 - b^2}| = |\frac{a^2(1 + \delta_1)^2 - b^2(1 + \delta_2)^2 - (a^2 - b^2)}{a^2 - b^2}| \tilde{\leqslant} 2\frac{a^2 + b^2}{|a^2 + b^2|}$$ - wskaźnik uwarunkowania

Wskaźnik uwarunkowania jest wysoki, co oznacza że niedokładne dane mają duży wpływ na błąd, czyli zadanie jest źle uwarunkowane. Uwaga, tutaj nie miało znaczenia którego algorytmu użyliśmy! Dla niedokładnych danych oba algorytmy dadzą niedokładne wyniki.

14.4 Numeryczna poprawność algorytmu

Jak obliczyć $f(x) = 1 - \cos x$? Dla $x \approx 0$ mamy $\cos x \approx 1$. Jak przekształcić?

$$\cos x = \cos 2\frac{x}{2} = \cos^2 \frac{x}{2} - \sin^2 \frac{x}{2}$$

Wtedy

$$1 - \cos x = \cos^2 \frac{x}{2} + \sin^2 \frac{x}{2} - \cos^2 \frac{x}{2} + \sin^2 \frac{x}{2} = 2\sin^2 \frac{x}{2}$$

Dzielenie przez 2 jest dość dokładne, a f. tryg. i tak musimy policzyć.

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

Mamy utratę precyzji gdy $x >> 1$

$$g(x) = \sqrt{x^2 + 1} - x = \frac{1}{\sqrt{x^2 + 1} + x}$$

Jak policzyć iloczyn skalarny $x^T y$?

$$x^T y = \sum_{i=1}^{n} x_i y_i$$

Sprawdźmy uwarunkowanie:

$$\tilde{x}_i = x_i(1 + \delta_i)$$

$$\tilde{y}_i = y_i(1 + \gamma_i)$$

$$|\frac{\sum \tilde{x}_i \tilde{y}_i - \sum x_i y_i}{\sum x_i y_i}| \approx |\frac{\sum \tilde{x}_i \tilde{y}_i (\delta_i + \gamma_i) - \sum x_i y_i}{\sum x_i y_i}| \leqslant 2\frac{\sum |x_i||y_i|}{|\sum x_i y_i|}\varepsilon$$

Zadanie jest niestety źle uwarunkowane.

15 Ćwiczenia 30/11

Znajdź uwarunkowanie zadania obilczania Ax ze względu na zaburzenia macierzy A Zaburzenie macierzy Δ - macierz
mamy $\tilde{A} = A + \Delta$
Zaburzenie względne:

$$\frac{\|\tilde{A} - A\|}{\|A\|} = \frac{\|\Delta\|}{\|A\|}$$

uwarunkowanie zadania:

$$\frac{\|\tilde{A}x - Ax\|}{\|Ax\|} = \frac{\|\Delta x\|}{\|Ax\|} \leqslant \frac{\|\Delta\|\|x\|}{\|Ax\|}\frac{\|A\|}{\|A\|} = \frac{\|A\|\|x\|}{\|Ax\|}\frac{\|\Delta\|}{\|A\|}$$

Współczynnik $\frac{\|A\|\|x\|}{\|Ax\|}$ przy zaburzeniu względnym macierzy jest większy lub równy 1.

Eliminacja Gaussa

$$A = \begin{bmatrix} \varepsilon & 1 \\ 1 & 1 \end{bmatrix}, b = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \varepsilon \in (0,1)$$

Wtedy $x = \frac{1}{1-\varepsilon}, y = 1 - \frac{\varepsilon}{1-\varepsilon}$.
Eliminacja Gaussa daje:

$$A = LU, L = \begin{bmatrix} 1 & 0 \\ \frac{1}{\varepsilon} & 1 \end{bmatrix}, U = \begin{bmatrix} \varepsilon & 1 \\ 0 & 1 - \frac{1}{\varepsilon} \end{bmatrix}$$

Co jeśli ε jest bliski 0?

$$fl(U) = \begin{bmatrix} \varepsilon & 1 \\ 0 & -\frac{1}{\varepsilon} \end{bmatrix}$$

$$Lz = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, z = \begin{bmatrix} z_1 \\ z_2 \end{bmatrix}, z_1 = 1, z_2 = 2 - \frac{1}{\varepsilon}$$

Ale $fl(z_2) = -\frac{1}{\varepsilon}$!

$$fl(U) \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ -\frac{1}{\varepsilon} \end{bmatrix} = fl(z)$$

Wyszło $x = 0, y = 1$!

Z wyborem elementu głównego! po przestawieniu:

$$\begin{bmatrix} 1 & 1 \\ \varepsilon & 1 \end{bmatrix} \begin{bmatrix} x & y \end{bmatrix} = \begin{bmatrix} 2 & 1 \end{bmatrix}$$

$$L = \begin{bmatrix} 1 & 0 \\ \varepsilon & 1 \end{bmatrix}, U = \begin{bmatrix} 1 & 1 \\ 0 & 1 - \varepsilon \end{bmatrix}$$

$$fl(U) = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

Dla małych ε wynik wychodzi $x = 1, y = 1$ - znacznie lepiej! Wybór elementu głównego daje nam numeryczną poprawność.

Uwarunkowanie zadania

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

$$\|A\|_1 = 2$$

$$\|A^{-1}\|_1 =$$

$$A^{-1} = \frac{1}{\varepsilon - 1} \begin{bmatrix} 1 & -1 \\ -1 & \varepsilon \end{bmatrix}$$

$$\|A^{-1}\|_1 = \frac{2}{1 - \varepsilon}$$

Dla małego ε macierz jest dobrze uwarunkowana, jednak bez wyboru el. gł. wynik jest słaby, więc algorytm nie jest numerycznie poprawny.

15.1 UwUarunkowanie LZNK

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

Układem równań normalnych

$$A^T Ax = A^T b$$

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

Ale float!

$$fl(A^T A) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Trzeba używać Householderae albo Givensa.

15.2 Interpolacja

Dane: $f : \mathbb{R} \to \mathbb{R}$
Szukamy wielomianu, $w(x)$ stopnia n t. że $w(x_i) = f(x_i)$ dla $i = 0, 1, \ldots, n$.

Bazy wielomianów:

- potęgowa (standardowa) - $\{1, x, x^2, \ldots, x^n\}$
- Newtona - $\{1, x - x_0, (x - x_0)(x - x_1), \ldots, \prod_{i=0}^{n-1}(x - x_i)\}$
- Langranża - $\{l_i\} : l_i(x) = \prod_{j=0, j\neq i}^{n-1} \frac{x - x_j}{x_i - x_j}$

Wartości wielomianów w bazie standardowej Schemat Hornera!

$$w(x) = \sum_{i=0}^{n} a_i x^i = (a_n x^{n-1} + \cdots + a_1)x + a_0$$

Można wyłączać kolejne x aż do $(\ldots(a_n x + a_{n-1})x + \cdots + a_1)x + a_0$

```
w = a[n];
for k = n - 1 downto 0 do
    w = w*x+a[k]
done
```

#taniejniema - n dodawań, n mnożeń

Co dla bazy Newtona? Też schemat Hornera, tylko zamiast wyciągać x wyciągamy $x - x_i$ - $2n$ dodawań, n mnożeń

Co dla bazy Lagaraża? Nie jest to dobry pomysł żeby jej używać w obliczeniach numerycznych.

Czy Lagrażyna to wgl baza?

$$l_i(x_k) = \begin{cases} 0 & i \neq k \\ 1 & \text{wpp} \end{cases}$$

To co wgl robimy z Lagryzoniem? Postać barycentryczna wielomianu interpolacyjnego Lagrożnego

$$p(x) = \prod_{i=0}^{n} (x - x_i)$$

Nie jest on w bazie Newtona!!!1!

$$l_i(x) = \frac{p(x)}{x - x_i} \prod_{j=0, j\neq i}^{n} \frac{1}{x_i - x_j}$$

Niech $L(x)$ - wielomian interpolujący funkcję $f(x)$.

$$L(x) = \sum_{i=0}^{n} c_i L_i(x)$$

Dla x_k z całej sumy zostaje tylko $c_k L_k(x_k) = c_k$.
Łatwo wyznaczyć interpolację, trudno nią interpolować!

$$L(x) = \sum_{i=0}^{n} \frac{p(x)}{x - x_i} \prod_{j=0, j\neq i}^{n} \frac{1}{x_i - x_j} f(x_i) = p(x) \sum_{i=0}^{n} \frac{w_i}{x - x_i} f(x_i) \text{gdzie} w_i = \prod_{j\neq i} \frac{1}{x_i - x_j}$$

Pierwsza postać barycentryczna! Niech $\tilde{L}$ - wielomian interpolujący funkcję $f(x) \equiv 1$. Jasne jest, że $\tilde{L} \equiv 1$.

$$\tilde{L}(x) = p(x) \sum_{i=0}^{n} \frac{w_i}{x - x_i}$$

$$L(x) = \frac{L(x)}{\tilde{L}(x)} = \frac{\sum_{i=0}^{n} \frac{w_i}{x - x_i} f(x_i)}{\sum_{i=0}^{n} \frac{w_i}{x - x_i}}$$

Druga postać barycentryczna - pozbyliśmy się $p(x)$, co jest dobre question mark?

16 Ćwiczenia 7/12

16.1 Ko-loss

16.2 Interpollacja

Mamy wielomian w bazie newtona:

$$w(x) = \sum_{i=0}^{n} b_i p_i(x) \quad p_i(x) = \begin{cases} 1 & i = 0 \\ \prod_{j=0}^{i-1}(x - x_i) & i > 0 \end{cases}$$

$$w(x_i) = f(x_i) i = 0, \ldots, n$$

A $f(x_i)$ mamy dane!

$$b_i = f[x_0, x_1, \ldots, x_i]$$

$$f[x_0, \ldots, x_i] = \frac{f[x_1, \ldots, x_i] - f[x_0, \ldots, x_{i-1}]}{x_i - x_0}$$

oraz

$$f[x_i] = f(x_i)$$

Ręczna robótka - mamy dane węzły $\{2, 4, 0\}$, wartości w węzłach $\{11, 63, 7\}$. Liczymy od najmniejszych przypadków.

2	11		
4	63	26	
0	7	14	6
	$f[x_j]$	$f[x_j, x_j + 1]$	$f[x_0, x_1, x_2]$

Interesują nas tylko wartości na przekątnej, więc można to robić w miejscu. Używamy tych współczynników, które zaczynają się od indeksu zerowego. Wtedy $w(x) = 11 + 26(x - 2) + 6(x - 2)(x - 4)$

16.3 Błąd interpolacji

$$f(x) - w(x) = f[x_0, x_1, \ldots, x_n, x]p(x), \text{ dla } x \neq x_i$$

Mało przydatne, ale okazuje się że:

$$f[x_1, x_{i+1}, \ldots, x_{i+k}] = \frac{f^{(k)}(\zeta)}{k!} \text{ o ile } f \in C^k, \zeta \in [x_i, x_{i+k}]$$

Jak więc zmierzyć ten błąd? Normą supremum.

$$\|g\|_\infty = \sup_{x \in [a,b]} g(x)$$

U nas:

$$\sup_{x \in [x_0, x_n]} |f(x) - w(x)| = \sup |\frac{f^{(n+1)}(\zeta)}{(n+1)!} p(x)| \leqslant \sup |\frac{f^{(n+1)}(\zeta)}{(n+1)!}| \sup |p(x)|$$

Zatem:

$$\|f - w\|_{\infty,[x_0,x_n]} \leqslant \frac{\|f^{(n+1)}\|_{\infty,[x_0,x_n]}}{(n+1)!}\|p\|_{\infty,[x_0,x_n]}$$

Wybór węzłów mocno wpływa na błąd!

Węzły równo odległe

$$x_i = x_0 + ih, h = \frac{x_n - x_0}{n}, i = 0, 1, \ldots, n$$

Wtedy mamy:

$$\|p\|_{\infty,[x_0,x_n]} \leqslant \frac{n!h^{n+1}}{4}$$

Dowód. dla $n = 1$:

$$p(x) = (x - x_0)(x - x_1), h = x_1 - x_0$$

A skoro p jest stopnia 2 i nieujemne, to najwyższy punkt między miejscami zerowymi (x_0, x_1) jest dokładnie w połowie, czyli:

$$\sup_{x\in[x_0,x_1]} |(x - x_0)(x - x_1)| = \frac{h^2}{4}$$

Krok indukcyjny:

$$\sup_{x\in[x_0,x_n]} |(x-x_0)\ldots(x-x_n)| = \max\{\sup_{x\in[x_0,x_{n-1}]} |\prod_i (x-x_i)|, \sup_{x\in[x_1,x_n]} |\prod_i (x-x_i)|\}$$

To pierwsze szacujemy przez:

$$\sup_{x\in[x_0,x_{n-1}]} |\prod_i^{n-1}(x - x_i)| \sup_{x\in[x_0,x_{n-1}]} |(x - x_n)| \leqslant \frac{(n-1)!h^n}{4} nh$$

□

16.4 Oszacuj błąd interpolacji

Lagrenżola o normie supremeum dla $f(x) = \sin x$ na $[-\frac{\pi}{2}; \frac{\pi}{2}]$ na węzłach $\{-\frac{\pi}{2}, 0, \frac{\pi}{2}\}$. Dla $n = 2$: Trzecia pochodna to minus cosinus, jej norma supremum to 1, a norma p jest mniejsza od $\frac{1}{2}(\frac{\pi}{2})^3$