Počítačové metody v teoretické fyzice I

- informace o přednášce na utf.mff.cuni.cz/~houfek/

Úvod – reprezentace čísel v počítači, zaokrouhlovací chyba, stabilita algoritmů a podmíněnost úlohy

- nejběžnější je reprezentace pomocí pohyblivé řádové čárky (floating-point arithmetic) daná IEEE standardem

např. čísla $\pm 0{,}01$ není reprezentováno přesně, neboť

$$\pm 0{,}01\ (\mathrm{DEC}) = \pm 0{,}000\ 000\ 101\ 000\ 111\ 101\ldots\ (\mathrm{BIN})$$
$$= \pm \underbrace{1.01000111101\ldots}_{t \text{ cifer, určuje přesnost (precision)}} \times 2^{-7}\ (\mathrm{BIN})$$
$$= \pm m \times \beta^{e-t+1}$$

[pozn. neplést s angl. accuracy]

m ... mantisa (significand)
β ... základ (base) – obvykle 2, někdy 10, 16
e ... exponent

- mantisa je celočíselná (ovšem může být někdy brána i jako desetinné číslo)

$$m = d_1 d_2 \ldots d_t$$

kde d_i nabývají hodnot $0 \le d_i < \beta$, ovšem $d_1 \neq 0$

- číslo 0,01 je ve 32-bitovém IEEE standardu = single precision, neboli v REAL*4 (4 byty), reprezentováno takto:

$0{,}01 \approx$ | 0 | 0 1 1 1 1 1 0 0 | 0 1 0 0 0 1 1 | 1 1 0 1 0 1 1 1 | 0 0 0 0 1 0 1 0 |

znaménkový bit s: $(-1)^s$ — $127+e$ (8 bitů) — mantisa m bez první 1 (23 bitů) ~ 7 platných míst

v double precision: 11 bitů — 52 bitů ~ 16 platných míst v desítkové soust.

- rozsah exponentů je $e_{min} \le e \le e_{max}$

přičemž pro single precision je $e_{min} = -126$ a $e_{max} = 127$
$(2^{-126} \approx 1{,}18 \cdot 10^{-38},\ 2^{127} \approx 3{,}4 \cdot 10^{38})$

a pro double precision $e_{min} = -1022$ a $e_{max} = 1023$
$(2^{-1022} \approx 2{,}2 \cdot 10^{-308},\ 2^{1023} \approx 1{,}8 \cdot 10^{308})$

- pozn. speciální exponenty $e = -127$ (samé nuly) a $e = 128$ (samé jedničky) se používají na tzv. subnormální čísla a též pro $\pm\infty$ a NaN

- pokud reálné číslo $x \in \mathbb{R}$ leží v rozsahu systému F čísel s pohyblivou řádovou čárkou, pak

$$fl(x) = x(1+\delta), \text{ kde } |\delta| < u = \frac{\varepsilon_M}{2}$$

kde $fl(x)$ značí číslo x vyjádřené nejbližším číslem s pohyblivou řádovou čárkou, tj. $f(x) \in F$

a u je tzv. unit roundoff $u = \frac{1}{2}\beta^{1-t}$

a ε_M je tzv. <u>strojové epsilon</u>.

Maximální relativní chyba vyjádření lib. čísla x v rozsahu F je tedy dána u.

- je-li $fl(x) = x + \Delta x = x(1+\varepsilon_x)$, hovoříme o Δx jako o <u>zaokrouhlovací chybě</u> a o ε_x jako o <u>relativní zaokrouhlovací chybě</u>

[Pozn: přestože je reprezentace čísel pomocí pohyblivé řádové čárky zdaleka nejběžnější, není jediná. Existují i další systémy jako např. pevná řádová čárka (fixed-point) nebo logaritmický číselný systém (logarithmic number system), a pod. (viz wikipedia nebo Higham: Accuracy and stability ...]

- <u>vliv zaokrouhovacích chyb na výsledek</u>
  - při výpočtech na počítači je nutné mít na paměti, že zaokrouhlovací chyba je všudypřítomná
  - čísla na vstupu jsou zaokrouhlena
  - výsledek lib. aritmetické operace je zaokrouhlen

    Pozn: dobrý systém by měl mít implementovány tyto operace tak, že výsledek bude

    $$fl(x \, op \, y) = (x \, op \, y)(1+\delta), \quad |\delta| < u$$

    pro $op = +, -, *, /$ a druhou odmocninu též
  - numerická matematika se z velké části zabývá právě vlivem zaokrouhlovací chyby na správnost výpočtu

1) pozor! - některé axiomy reálné aritmetiky nejsou v systémech s pohyblivou řádovou čárkou splněny

např: <u>výsledek závisí na pořadí sčítanců</u>

a to i tehdy, jsou-li všechny sčítance kladné

Př. uvažujme jednoduchou sumu (viz notebook roundoff.errors.nb)

$$S_n = 1 + \sum_{k=1}^{n} \frac{1}{k^2+k} = 1 + \sum_{k=1}^{n} \left(\frac{1}{k} - \frac{1}{k+1}\right) = 2 - \frac{1}{n+1}$$

- sčítáme-li odpředu, objeví se zaokrouhlovací chyby, zvláště pro velká n
- sčítáme-li odzadu (od nejmenších příspěvků po největší), zaokrouhlovací chyby se neprojeví

nebo: <u>rovnice nemusí mít jednoznačné řešení</u>

<u>případně nemusí mít žádné řešení</u>

(i když v přesné aritmetice existuje jediné)

Př. rovnice $1 + x = 1$ má řešení $x = 0$

ovšem v double precision bude rovnice splněna pro lib. $x \lesssim 10^{-17}$, které patří do systému čísel double prec.

Př. rovnice $f(x) = 0$ většinou nemá v daném systému čísel s pohyblivou řádovou čárkou žádné řešení, lze hledat pouze řešení, pro které je $|f(x)|$ nejmenší, případně menší než jisté ε, když nám stačí určitá přesnost

2) <u>odečítání blízkých (ne nutně velkých) čísel</u> - cancellation problem

- operace odečítání může vnést do numerických výpočtů značné chyby, jsou-li odečítaná čísla blízká

- máme-li $fl(x_1) = x_1 + \Delta x_1$ a $fl(x_2) = x_2 + \Delta x_2$ pak $fl(x_1 \pm x_2) = x_1 \pm x_2 + \Delta x_1 \pm \Delta x_2 + \varepsilon(x_1 \pm x_2)$ kde $\varepsilon \lesssim \varepsilon_M$

a relativní chyba je

$$\delta = \left| \frac{fl(x_1 \pm x_2) - (x_1 \pm x_2)}{x_1 \pm x_2} \right| \leq \frac{|\Delta x_1| + |\Delta x_2|}{|x_1 \pm x_2|} + |\varepsilon|$$

- pokud je např. $x_1 \approx 1$, $x_2 \approx 1$, $\Delta x_1 \approx 10^{-8}$ a $\Delta x_2 \approx 10^{-8}$ a v přesné aritmetice $x_1 - x_2 \approx 10^{-4}$, pak $\delta \approx 2 \cdot 10^{-4}$
- často se s tímto problémem nedá moc dělat, ale někdy ze znalosti problému lze změnit výpočet tak, že se odečítání lze vyhnout

[Př. výraz $\sqrt{x+1} - \sqrt{x}$ lze vyčíslit též pomocí $\frac{1}{\sqrt{x+1} + \sqrt{x}}$ a pro velká x pak velká zokrouhlovací chyba nevzniká]

[Př. řešení kvadratické rovnice $ax^2 + bx + c = 0$ spočtené pomocí $x_{1,2} = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$ může být zatíženo chybou, pokud $b^2 \gg |4ac|$, neboť pak $\sqrt{b^2 - 4ac} \approx |b|$ a jeden z kořenů dostaneme odečítáním dvou blízkých čísel, lze obejít tak, že tento kořen spočteme jako $x_2 = \frac{c}{a x_1}$]

3) <u>součet řady „velkých" čísel, jehož výsledek je „malý"</u>

- jde o skryté odečítání blízkých čísel, neboť až mezivýsledky můžou být blízké (smearing problém)
- opět lze někdy obejít jiným přístupem k danému problému

[Př. výpočet e^x pomocí Taylorovy řady $1 + \frac{x}{1!} + \frac{x^2}{2!} + \dots$ pro $x > 0$ více méně v pořádku, jen pro větší x máme pomalou konvergenci

ovšem pro $x < 0$ výpočet selže (výsledek je malé číslo jako součet velkých s opačnými znaménky)

trik: pro $x < 0$ lze psát $e^x = \frac{1}{e^{-x}}$]

- často nastává u řad typu binomického rozvoje

Př. učebnicový příklad ilustrující řešení přes rekurentní vztahy:

integrál $I_n(a) = \int_0^1 \frac{x^n}{x+a}\,dx = \left(\text{trik } x^n = [(x+a)-a]^n\right)$

$$= \sum_{k=0}^{n-1} (-a)^k \binom{n}{k} \left[(1+a)^{n-k} - a^{n-k}\right] \frac{1}{n-k} + (-a)^n \ln \frac{a+1}{a}$$

- použití této řady v double precision dá např. pro $a=10$ a $n \gtrsim 10$ řešení se značnými chybami, nebo vyjdou přímo očividné nesmysly ($I_n<0$) (koeficienty jsou např. pro $n=10$ a $k=5$ řádu 10^{11} a přitom výsledek leží v $0<I_n<1$!)

- řešení tohoto problému spočívá ve využití rekurentního vztahu

$$I_n = \int_0^1 \frac{x^{n-1}(x+a-a)}{x+a}\,dx = \int_0^1 x^{n-1}\,dx - a\int_0^1 \frac{x^{n-1}}{x+a}\,dx = \frac{1}{n} - aI_{n-1}$$

ovšem dopředná rekurence, kdy startujeme s $I_0 = \ln\frac{a+1}{a}$, <u>není stabilní</u> !

- libovolná chyba (i zaokrouhlovací) na začátku bude v n-tém kroku násobena faktorem $(-a)^n$ a výsledky pro $n \gtrsim 15$ znehodnotí úplně

- co to zkusit obráceně? $I_{n-1} = \frac{1}{a}\left(\frac{1}{n} - I_n\right)$

ovšem jak začít? lze libovolně, neboť počáteční chyba bude potlačovaná faktorem $\frac{1}{a^k}$, po k iteracích, např. pro $n=10$ a $a=10$ zvolíme $I_{n+k} = 1$ pro vhodně velké $k \gtrsim n + \log\frac{1}{\varepsilon_n}$ a dostaneme I_n s přesností ε_n

- <u>stabilita algoritmu</u>

Řekneme, že algoritmus výpočtu veličiny $f = f(x)$, závisející na vstupních datech x, je stabilní, pokud jeho provedením v aritmetice s pohyblivou řádovou čárkou dá pro všechna x výsledek $\tilde{f}$, který se od přesného f liší relativní chybou řádu $O(\varepsilon_M)$, tj. při zmenšení ε_M lze chybu odhadnout výrazem $c\varepsilon_M$, kde konstanta c nezávisí na x.

- v praxi navíc požadujeme, aby pro konkrétní ε_M (např. pro $\varepsilon_M \sim 10^{-16}$ v double prec.) nebylo $c \sim \varepsilon_M^{-1}$ (tj. $\sim 10^{16}$), neboť pak by chyba byla srovnatelná s výsledkem!

Př. při výpočtu $\sqrt{x+1} - \sqrt{x}$ lze pro velká kladná x odhadnout absolutní chybu jako $2\sqrt{x}\,\varepsilon_M$ a tedy relativní chyba bude $\delta \leq \frac{2\sqrt{x}\,\varepsilon_M}{\sqrt{x+1}-\sqrt{x}} = 2\varepsilon_M\sqrt{x}\left(\sqrt{x+1}+\sqrt{x}\right) \sim 4\varepsilon_M x$

a nelze odhadnout výrazem $c\varepsilon_M$ (pro velká x roste)

$\Rightarrow$ výpočet je nestabilní

ovšem pro $\frac{1}{\sqrt{x+1}+\sqrt{x}}$ dostaneme pro odhad relativní chyby jmenovatele $\delta \leq \frac{2\sqrt{x}\,\varepsilon_M}{\sqrt{x+1}+\sqrt{x}} \sim \varepsilon_M$ a pro celý výraz tedy také ε_M $\Rightarrow$ stabilní výpočet

- <u>podmíněnost úlohy</u>

  - v některých úlohách, bez ohledu na použitý algoritmus, může výsledek silně záviset na vstupních datech
  - klasickým příkladem takové úlohy je určení kořenů polynomu zadaného jeho koeficienty (vstupní data)

[Př. polynom $p(x) = x^{10} - 10x^9 + 45x^8 - \dots - 10x + 1 = (x-1)^{10}$

má desetinásobný kořen $x = 1$

avšak $q(x) = p(x) - 10^{-10}$ (pouze a_0 se změnilo o 10^{-10})

má kořeny $x_k = 1 + \frac{1}{10} e^{\frac{2\pi i k}{10}}$ pro $k = 0, \dots, 9$

a tedy změna výsledku je $\sim 10^{-1}$ (v přesné aritmetice!)]

- obecně uvažujme problém, který řešíme, jako funkci $f: X \to Y$, kde X je normovaný vektorový prostor vstupních dat a Y je normovaný vektorový prostor řešení s tím, že f je obecně (a též obvykle) nelineární.

Řekneme, že <u>úloha</u> (problém) <u>je dobře podmíněná</u> (well-conditioned), když malá změna vstupních dat x vede k malým změnám výstupních dat $f(x)$

(př. je např. výpočet vlastního integrálu, či určení vlastních čísel a vektorů hermitovské matice)

Naopak <u>úloha je špatně podmíněná</u>, (ill-conditioned) když malá změna x vede k velké změně $f(x)$.

(hledání kořenů polynomu, řešení soustavy lin. rovnic pro nevhodné matice či počáteční úloha pro dif. rovnice)

Podmíněnost se kvantifikuje buď pomocí <u>absolutního čísla podmíněnosti</u>

$$\hat{\varkappa}(x) = \lim_{\varepsilon \to 0} \sup_{\|\delta x\| \le \varepsilon} \frac{\|\delta f\|}{\|\delta x\|},$$

kde tedy δx je malá změna x a $\delta f = f(x + \delta x) - f(x)$

nebo pomocí <u>relativního čísla podmíněnosti</u>

$$\varkappa(x) = \lim_{\varepsilon \to 0} \sup_{\|\delta x\| \le \varepsilon} \frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|}$$

- pro diferencovatelné f je $\hat{\varkappa}(x)$ dáno pomocí Jacobiánu zobr. $f(x)$: $J_{ij}(x) = \frac{\partial f_i}{\partial x_j}\Big|_x$, $\delta f \approx J(x)\,\delta x$

  a tedy $\hat{\varkappa}(x) = \|J(x)\|$, kde maticová norma $J(x)$ je indukovaná normami v X a Y

  a $$\varkappa(x) = \frac{\|J(x)\|}{\frac{\|f(x)\|}{\|x\|}}$$

- relativní číslo podmíněnosti je obvykle důležitější při výpočtech v pohyblivé řádové čárce, kde pracujeme s relativní chybou $\sim \varepsilon_M$

- problém je (typicky) dobře podmíněný pro $\varkappa \lesssim 10^2$ a špatně podmíněný pro $\varkappa \gtrsim 10^{10}$

  ovšem přesná hodnota $\varkappa$ závisí na zvolené normě

Př. výpočet $f(x) = \sqrt{x}$ je dobře podmíněná úloha, neboť $$\varkappa = \frac{\frac{1}{2\sqrt{x}}}{\frac{\sqrt{x}}{x}} = \frac{1}{2}$$

ovšem pro $f(x) = x_1 - x_2$ dostaneme $J = (1, -1)$ a pro $\|\cdot\|_\infty$ normu pak $$\varkappa = \frac{2}{\frac{|x_1 - x_2|}{\max\{|x_1|, |x_2|\}}} = \frac{2\max\{|x_1|, |x_2|\}}{|x_1 - x_2|}$$ což může být libovolně velké pro $x_1 \sim x_2$

Pozn: $\|J\|_\infty = 2$ = maximum součtů řádkových prvků (viz poznámky k vektorovým a maticovým normám)

Př. pro hledání kořenů polynomu jde $\varkappa \to \infty$

neboť pro $\delta x = 10^{-10}$ je $\frac{\|\delta f\|}{\|\delta x\|} \approx 10^{9}$

pro $\delta x = 10^{-20}$ je $\frac{\|\delta f\|}{\|\delta x\|} \approx 10^{18}$

atd.

tj. $\frac{\|\delta f\|}{\|\delta x\|}$ roste nade všechny meze a přitom $\|f(x)\| \sim 10^2$ a $\|x\| \sim 10^0$

Normy na vekt. prostorech a normy matic

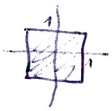
Def: Norma $\|\cdot\|$ je funkce zobrazující z $\mathbb{C}^n \to \mathbb{R}$ splňující pro $\forall x, y \in \mathbb{C}^n$ a $\alpha \in \mathbb{C}$:

1) $\|x\| \geq 0$ a $\|x\| = 0$ právě, když $x = 0$
2) $\|x+y\| \leq \|x\| + \|y\|$
3) $\|\alpha x\| = |\alpha| \|x\|$

p-normy na $\mathbb{C}^n$ a jejich jednotkové koule $\{x \in \mathbb{C}^n, \|x\| \leq 1\}$

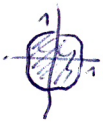
$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|$$

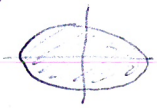


$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{1/2}$$

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$$



vážené p-normy $\|x\|_W = \|Wx\|_p = \left(\sum_{i=1}^{n} |w_i x_i|^p\right)^{1/p}$, kde W je diagonální matice



- lze zobecnit na lib. nesingulární matice W

Maticová norma indukovaná vekt. normou

- uvažujme matici $A \in \mathbb{C}^{m \times n}$ jako zobrazení z vekt. prostoru s normou $\|\cdot\|_{(n)}$ do prostoru s $\|\cdot\|_{(m)}$

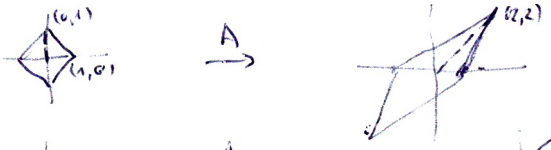
pak indukovaná matic. norma $\|A\|_{(m,n)}$ je nejmenší C pro které platí $\|Ax\|_{(m)} \leq C \|x\|_{(n)}$

neboli

$$\|A\|_{(m,n)} = \sup_{\substack{x \in \mathbb{C}^n \\ x \neq 0}} \frac{\|Ax\|_{(m)}}{\|x\|_{(n)}} = \sup_{\substack{x \in \mathbb{C}^n \\ \|x\|_{(n)} = 1}} \|Ax\|_{(m)}$$

Př. $A = \begin{pmatrix} 1 & 2 \\ 0 & 2 \end{pmatrix}$ - obecně A zobrazí $e_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ na $\begin{pmatrix} a_{11} \\ a_{21} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ a $e_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$ na $\begin{pmatrix} a_{12} \\ a_{22} \end{pmatrix} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$

1-norm



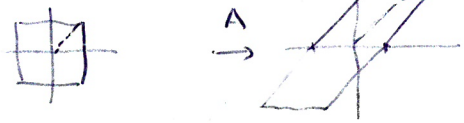
$\|A\|_1 = 4$

2-norm

$\|A\|_2 = \sqrt{\frac{1}{2} + \frac{7}{2\sqrt{65}}} \approx 2{,}9208$

$= \sqrt{\frac{1}{2}(9 + \sqrt{65})}$

∞-norm



$\|A\|_\infty = 3$ (uvidíme později, že $\|A\|_2 = \max \sigma_i$, kde σ_i jsou singulární čísla)

Př. pro lib. diagonální matici $D = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_m \end{pmatrix}$ se jednotková koule

ztransf. na hyperelipsu s poloosami ~~poloo~~ o délce $|d_i|$

a tedy $\|D\|_2 = \max_{1 \le i \le n} |d_i| = \|D\|_p$

- ~~lze ukázat~~ platí, že $\|A\|_1$ = maximum součtů absolutních hodnot sloupcových prvků

$= \max_{1 \le j \le n} \|a_j\|_1$ kde a_j je j-tý sloupec matice

neboť $\|Ax\|_1 = \|\sum_{j=1}^{n} x_j a_j\|_1 \le \sum_{j=1}^{n} |x_j| \|a_j\|_1 \le \|x\|_1 \max_{1 \le j \le n} \|a_j\|_1$, což platí i pro $\|x\|_1 = 1$

$$Ax = \left(a_1 \middle| a_2 \middle| \cdots \middle| a_n \right) \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$$

a volbou $x = e_j$, pro které je $\|a_j\|_1$ maximální, dostáváme rovnost a tedy

$$\|A\|_1 = \sup_{\|x\|_1 = 1} \|Ax\|_1 = \max_{1 \le j \le n} \|a_j\|_1$$

- podobně $\|A\|_\infty$ = maximum součtů řádkových prvků

$= \max_{1 \le i \le n} \|a_i^*\|_1$, kde a_i^* je i-tý řádek

neboť pro $\|x\|_\infty = 1$ $\|Ax\|_\infty = \max_{1 \le i \le m} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \le \max_i \sum_{j=1}^{n} |a_{ij}|$ neboť $|x_j| \le 1$

a rovnost nastává pro jeden z vektorů $x = (\pm 1, \pm 1, \dots \pm 1)$

Pozn. měli by studenti znát z LA Hölderovu nerovnost $|x^* y| \le \|x\|_p \|y\|_q$ kde $1 \le p, q \le \infty$ a $\frac{1}{p} + \frac{1}{q} = 1$

a spec. případ Cauchyovu-Schwarzovu nerovnost $|x^* y| \le \|x\|_2 \|y\|_2$

- Pro odhad normy součinu dvou matic platí

$$\|AB\|_{(p,n)} \le \|A\|_{(p,-)} \|B\|_{(-,n)}$$

ale obecně zde není rovnost, např. $\|A^n\| \le \|A\|^n$, ale pro $n \ge 2$ obecně neplatí $\|A^n\| = \|A\|^n$

<u>Obecné maticové normy</u> - splňují stejné podmínky jako normy vektorů jen nyní maticové (jako kdyby měli max rozměrný vekt. prostor)

- nejpoužívanější je Frobeniova norma (též Hilbertova-Schmidtova norma)

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |a_{ij}|^2 \right)^{1/2}$$

, tj. jde o 2-normu v max. rozm. prostoru

i zde platí odhad pro součin

$$\|AB\|_F \le \|A\|_F \|B\|_F$$

(lze ukázat, že $\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \dots \sigma_r^2}$