NI-MPI přednáška 18

Numerika: odhady zaokrouhlovacích chyb – příklady

Štěpán Starosta

16. 12. 2024

FIT ČVUT

Připomenutí

Nechť $x, y \in F$ a $\odot$ značí operaci sčítání, odečítání, násobení nebo dělení. Pokud nedojde k přetečení nebo podtečení (zůstali jsme v intervalu normalizovaných čísel), tak platí

$$\mathrm{fl}(x \odot y) = (x \odot y)(1 + \delta), \qquad \text{kde } |\delta| \leq \mathbf{u}.$$

Mez $\mathbf{u}$ je **zaokrouhlovací jednotka**.

Lemma 22.4
Pokud $|\delta_i| \leq \mathbf{u}$ *a* $|\rho_i| = 1$ *pro všechna* $i \in \{1, \ldots, n\}$, $n\mathbf{u} < 1$, *tak platí*

$$\prod_{i=1}^{n}(1+\delta_i)^{\rho_i} = 1 + \Theta_n,$$

kde $|\Theta_n| \leq \dfrac{n\mathbf{u}}{1-n\mathbf{u}}$.

Důkaz.

Dokážeme indukcí.

$n = 1$ a $\rho_1 = 1$, pak $|\Theta_1| = |\delta_1| \leq \mathbf{u} < \dfrac{\mathbf{u}}{1-\mathbf{u}}$.

$n = 1$ a $\rho_1 = -1$, pak $\frac{1}{1+\delta_1} = 1 - \frac{\delta_1}{1+\delta_1}$ a tedy $|\Theta_1| = \frac{|\delta_1|}{|1+\delta_1|} \leq \dfrac{\mathbf{u}}{1-\mathbf{u}}$. □

Načítání chyb - pokračování

pokračování.

Předpokládejme, že tvrzení platí pro $n = j - 1$.

Nechť $\rho_j = 1$, pak $\prod_{i=1}^{j}(1+\delta_i)^{\rho_i} = (1+\Theta_{j-1})(1+\delta_j) = 1 + \underbrace{\delta_j + \Theta_{j-1} + \delta_j\Theta_{j-1}}_{\Theta_j}$.

$$|\Theta_j| \le |\delta_j| + |\Theta_{j-1}| + |\delta_j\Theta_{j-1}| \le \mathbf{u} + \frac{(j-1)\mathbf{u}}{1-(j-1)\mathbf{u}} + \frac{(j-1)\mathbf{u}^2}{1-(j-1)\mathbf{u}} = \frac{j\mathbf{u}}{1-(j-1)\mathbf{u}} \le \frac{j\mathbf{u}}{1-j\mathbf{u}}.$$

Nechť $\rho_j = -1$, pak $\prod_{i=1}^{j}(1+\delta_i)^{\rho_i} = \frac{1+\Theta_{j-1}}{1+\delta_j} = 1 + \underbrace{\frac{-\delta_j + \Theta_{j-1}}{1+\delta_j}}_{\Theta_j}$.

$$|\Theta_j| \le \frac{|\delta_j|}{|1+\delta_j|} + \frac{|\Theta_{j-1}|}{|1+\delta_j|} \le \frac{\mathbf{u}}{1-\mathbf{u}} + \frac{1}{1-\mathbf{u}} \cdot \frac{(j-1)\mathbf{u}}{1-(j-1)\mathbf{u}} = \frac{j\mathbf{u} - (j-1)\mathbf{u}^2}{1-j\mathbf{u}+(j-1)\mathbf{u}^2} \le \frac{j\mathbf{u}}{1-j\mathbf{u}}.$$ □

Značení

Budeme používat výhodné značení $\langle n \rangle = \prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$ pro počítání počtu nakumulovaných relativních chyb.

Platí

$$\langle j \rangle \cdot \langle k \rangle = \langle j+k \rangle \qquad \text{a} \qquad \frac{\langle j \rangle}{\langle k \rangle} = \langle j+k \rangle.$$

Pozor: toto značení nám velmi zjednoduší zápis (pomůže jednoduchému počítání nakumulovaných relativních chyb), ale formálně není korektní, protože oba výrazy $\prod_{i=1}^{n}(1+\delta_i)^{\rho_i}$ a $\prod_{i=1}^{n}(1+\delta_i')^{\rho_i'}$ označíme $\langle n \rangle$ a přitom se obecně nerovnají. Při jejich používání je tedy třeba dávat pozor.

Skalární součin

Základní cvičení 24.2

Mějme pevně danou množinu strojových čísel F (např. v jednoduché přesnosti) a uvažujme standardní model aritmetických operací.

Uvažujme algoritmus $V : F^n \to F$, který počítá skalární součin, tedy

$$V(x_1, x_2, \ldots, x_n) = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n,$$

kde $\alpha_i \in F$ jsou pevně zvolené parametry.

1. Odhadněte dopřednou chybu.
2. Odhadněte zpětnou chybu.

Předpokládáme, že nedojde k podtečení, přetečení apod.

Označme $s_i = \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_i x_i$.

$\widehat{s_1} = \mathrm{fl}(\alpha_1 x_1) = \alpha_1 x_1 (1 + \delta_1)$, kde $|\delta_1| \leq \mathbf{u}$. Tedy $\widehat{s_1} = \alpha_1 x_1 \langle 1 \rangle$.

$\widehat{s_2} =$

$\widehat{s_3} =$

$\widehat{s_n} = \widehat{V}(x_1, x_2, \ldots, x_n) = \alpha_1 x_1 \langle n \rangle + \alpha_2 x_2 \langle n \rangle + \alpha_3 x_3 \langle n-1 \rangle + \cdots + \alpha_n x_n \langle 2 \rangle$ pro $n > 1$.

$\left| \widehat{V}(x_1, x_2, \ldots, x_n) - V(x_1, x_2, \ldots, x_n) \right| \leq$

Funkční hodnota polynomu

Základní cvičení 24.3

Mějme pevně danou množinu strojových čísel F (např. v jednoduché přesnosti) a uvažujme standardní model aritmetických operací.

Uvažujme zobrazení $p : x \mapsto (x-2)^9$ a 3 způsoby jeho výpočtu:

a) $p_a(x) = (x-2)^9$;

b) $p_b(x) = x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512$;

c) $p_c(x) = -512 + x\,(2304 + x\,(-4608 + x\,(\ldots)))$ (Hornerova metoda/pravidlo).

Uvažujme algoritmus $V_z : F \to F : x \mapsto p_z(x)$ pro $z \in \{a, b, c\}$ (tedy počítá fukční hodnotu $p(x)$ 3 výše uvedenými způsoby). Pro všechny 3 varianty

1. odhadněte dopřednou chybu, a
2. odhadněte zpětnou chybu.

Předpokládáme, že nedojde k podtečení, přetečení apod.

Metoda a

$p_a(x) = (x-2)^9$

$\widehat{p_a}(x) =$

$|\widehat{p_a}(x) - p_a(x)| \leq$

$\widehat{p_a}(x) = p_a(x + \Delta x)$, kde $|\Delta x| \leq$

Metoda b

$$p_b(x) = x^9 - 18x^8 + 144x^7 - 672x^6 + 2016x^5 - 4032x^4 + 5376x^3 - 4608x^2 + 2304x - 512$$

$$\widehat{p_b}(x) =$$

$$|\widehat{p_b}(x) - p_b(x)| \leq$$

$$\widehat{p_b}(x) = \widehat{p_b}(x + \Delta x), \text{ kde } |\Delta x| \leq$$

Hornernova metoda

Chceme vyhodnotit funkční hodnotu polynomu

$$p(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

Hornerova metoda spočívá ve vyhodnocení

$$p(x) = \Big(\big(\cdots((a_n)x + a_{n-1})x + a_{n-2}\big)x + \cdots + a_2\Big)x + a_1\Big)x + a_0,$$

kde je potřeba $2n$ operací s plovoucí čárkou[1] pro $n > 0$.

[1] *flops*

Hornerova metoda - chyba

Spočítáme chybu vzniklou použitím operací se strojovými čísly.

$$(a_n x\langle 1\rangle + a_{n-1})\langle 1\rangle = a_n x\langle 2\rangle + a_{n-1}\langle 1\rangle.$$

$$\begin{aligned}\big((a_n x\langle 2\rangle + a_{n-1}\langle 1\rangle)x\langle 1\rangle + a_{n-2}\big)\langle 1\rangle = \\ (a_n x^2\langle 3\rangle + a_{n-1}x\langle 2\rangle + a_{n-2})\langle 1\rangle = \\ a_n x^2\langle 4\rangle + a_{n-1}x\langle 3\rangle + a_{n-2}\langle 1\rangle.\end{aligned}$$

Indukcí dostaneme celkovou chybu napočítané hodnoty polynomu p v bodě x označené $\widehat{p}(x)$:

$$\widehat{p}(x) = a_0\langle 1\rangle + a_1 x\langle 3\rangle + \cdots a_{n-1}x^{n-1}\langle 2n-1\rangle + a_n x^n\langle 2n\rangle.$$

Hornerova metoda - dopředná chyba

Platí $\langle i \rangle = 1 + \Theta_i$, kde $|\Theta_i| \leq \dfrac{i\mathbf{u}}{1 - i\mathbf{u}} = \gamma_i$.

Odhadneme dopřednou chybu následovně

$$|p(x) - \widehat{p}(x)| \leq \gamma_{2n} \sum_{i=0}^{n} |a_i||x|^i.$$

Pro relativní chybu pak platí

$$\frac{|p(x) - \widehat{p}(x)|}{|p(x)|} \leq \gamma_{2n} \frac{\sum_{i=0}^{n} |a_i||x|^i}{|p(x)|}$$

Toto je *apriorní teoretický odhad* dopředné chyby, a v některých případech je odhad velice nadsazený. Lehce nahlédneme, že pravá strana může být jakkoliv velká.

Místo tohoto horního odhadu si spočítáme k jaké zaokrouhlovací chybě došlo přesněji (tzv. *Running error analysis*).

Hornernova metoda - aposteriorní odhad dopředné chyby (1/2)

Pro dané x definujme posloupnost (q_i) takto: $q_n = a_n$ a $q_i = q_{i+1}x + a_i$ pro všechna $i \in \{0, \ldots, n-1\}$.

Tedy $q_0 = p(x)$.

V i-tém kroku Hornerovy metody platí

$$(1+\epsilon_i)\widehat{q_i} = \widehat{q}_{i+1}x(1+\delta_i) + a_i, \qquad \text{kde } |\delta_i|, |\epsilon_i| \leq \mathbf{u}.$$

Označme $\widehat{q_i} = q_i + f_i$, dostaneme

$$(1+\epsilon_i)\widehat{q_i} = (q_{i+1} + f_{i+1})x + x\widehat{q}_{i+1}\delta_i + a_i.$$

Vyjádříme f_i:

$$f_i = \underbrace{q_{i+1}x - q_i + a_i}_{=0} + f_{i+1}x + x\widehat{q}_{i+1}\delta_i + \epsilon_i\widehat{q_i}.$$

Platí $f_n = 0$.

Hornernova metoda - aposteriorní odhad dopředné chyby (2/2)

Odhadneme f_i:

$$|f_i| \leq |f_{i+1}||x| + \mathbf{u}(|x||\widehat{q}_{i+1}| + |\widehat{q}_i|).$$

Označme posloupnost (π_i) tak, aby $|f_i| \leq \mathbf{u}\pi_i$, tedy

$$\pi_n = 0 \qquad \text{a} \qquad \pi_i = |x|\pi_{i+1} + |x||\widehat{q}_{i+1}| + |\widehat{q}_i|.$$

Aposteriorní dopřednou chybu π_0 tedy můžeme výhodně napočítat během výpočtu $\widehat{q}_0 = \widehat{p}(x)$ a bude platit

$$|\widehat{p}(x) - p(x)| \leq \pi_0 \mathbf{u}.$$

Hornerova metoda - zpětná chyba

$$\widehat{p}(x) = a_0\langle 1\rangle + a_1 x\langle 3\rangle + \cdots a_{n-1}x^{n-1}\langle 2n-1\rangle + a_n x^n\langle 2n\rangle.$$

Tedy zpětná chyba je:

$\widehat{p}(x) = p(x + \Delta(x))$, kde $|\Delta(x)| \leq \gamma_{2n}|x|$

Výpočet funkční hodnoty polynomu - relativní podmíněnost

Předpokládejme $p(x) \neq 0$.

Podle věty o střední hodnotě platí

$$|p(x + \delta x) - p(x)| = |p'(\epsilon)\delta x|$$

pro nějaké $\epsilon \in (x, x + \delta x)$ nebo $\epsilon \in (x + \delta x, x)$.

Platí

$$\frac{|p(x + \delta x) - p(x)|}{|p(x)|} = \frac{|p'(\epsilon)\delta x|}{|p(x)|}$$

a tedy

$$C_r = \frac{|p(x + \delta x) - p(x)|/|p(x)|}{|\delta x|/|x|} = \frac{|xp'(\epsilon)|}{|p(x)|}$$