Poglavje 2

Numerična stabilnost

Algoritem je **direktno stabilen**, če vrne rezultat, ki se malo razlikuje od prave vrednosti. Ponavadi preverjamo relativno direktno stabilnost. Direktna absolutna in relativna napaka sta

$$|y - \widehat{y}| \text{ in } \frac{|y - \widehat{y}|}{|y|}.$$

Algoritem je **obratno stabilen**, če za izračunan rezultat obstajajo taki malo zmoteni podatki, da iz njih s točnim izračunom dobimo izračunano vrednost. Obratna absolutna napaka je najmanjši $|\Delta x|$, tako da velja $f(x + \Delta x) = \widehat{y}$. Obratna relativna napaka je $\frac{|\Delta x|}{|x|}$.

Pri obratni napaki je izračunana vrednost enaka $\widehat{y} = f(\widehat{x})$. Če je funkcija f *zvezno odvedljiva* v x, potem velja

$$|\widehat{y} - f(x)| = |f(\widehat{x}) - f(x)| \leq |f'(x)||\widehat{x} - x|.$$

Če f ni absolutno občutljiva, bo pri majhnih vrednostih $\Delta x = |\widehat{x} - x|$ napaka absolutno obratno stabilne metode majhna in metoda bo absolutno direktno stabilna. Podobno velja za relativno stabilnost in relativne napake. Več ste povedali na predavanjih.

Naloga 2.1 *Vrednost $z = x^2 - y^2$ računamo na dva načina:*

(i). $z = x^2 - y^2$

(ii). $z = (x - y)(x + y)$

Analiziraj algoritma. Oceni relativno napako $\frac{|\widehat{z} - z|}{|z|}$. Ali je kateri od obeh algoritmov direktno/obratno stabilen?

Rešitev.

(i). Imamo $z = x * x - y * y$, $\widehat{a} = x * x(1 + \alpha)$, $\widehat{b} = y * y(1 + \beta)$, $\widehat{z} = (\widehat{a} - \widehat{b})(1 + \gamma)$, kjer je $|\alpha|, |\beta|, |\gamma| \leq u$, u je relativna natančnost. Sledi, da je

$$\widehat{z} = x^2 \overbrace{(1 + \alpha)(1 + \gamma)}^{(1+\delta_1)} - y^2 \overbrace{(1 + \beta)(1 + \gamma)}^{(1+\delta_2)}.$$

Iz ocene $1 - 2u + u^2 = (1 - u)^2 \leq (1 + \delta_1) \leq (1 + u)^2 = 1 + 2u + u^2$, pri majhnem u dobimo $\delta_1, \delta_2 \leq 2 * u$. Ocenimo izraz

$$|\widehat{z} - z| = |x^2\delta_1 - y^2\delta_2| \leq x^2|\delta_1| + y^2|\delta_2| \leq 2u(x^2 + y^2).$$

Torej velja ocena:

$$\frac{|\widehat{z}-z|}{|z|} \leq 2u\frac{x^2+y^2}{x^2-y^2}.$$

Ocena je smiselna. Najlažje to vidimo tako, da izberemo $\delta_1 = u$, $\delta_2 = -u$. Iz tega vidimo, da ta algoritem ni direktno stabilen, saj lahko pri x in y, ki sta si blizu, dobimo veliko napako. Algoritem je obratno stabilen. Če definiramo $\widehat{x} = x\sqrt{(1+\delta_1)}$ in $\widehat{y} = y\sqrt{(1+\delta_2)}$, ki sta blizu x in y, ter predpostavimo, da je računanje točno, dobimo zmoten izraz $\widehat{z} = \widehat{x}^2 - \widehat{y}^2$.

(ii). Oglejmo si še $z = (x-y)(x+y)$. Definirajmo $\widehat{a} = (x-y)(1+\alpha)$, $\widehat{b} = (x+y)(1+\beta)$ ter $\widehat{z} = \widehat{a}\widehat{b}(1+\gamma)$, kjer je $|\alpha|, |\beta|, |\gamma| \leq u$. Izraz je enak

$$\widehat{z} = (x-y)(x+y)\overbrace{(1+\alpha)(1+\beta)(1+\gamma)}^{(1+\delta)},$$

kjer s podobno oceno kot v prejšnji točki dobimo $|\delta| \leq 3u$. Torej velja ocena $|\widehat{z}-z| \leq 3u|z|$. Na koncu dobimo $\frac{|\widehat{z}-z|}{|z|} \leq 3u$. Algoritem je direktno stabilen. Prav tako je obratno stabilen, iskana zmotena x in y sta recimo: $\widehat{x} = x\sqrt{(1+\delta)}$, $\widehat{y} = y\sqrt{(1+\delta)}$.

■

Naloga 2.2 *Dan je polinom*

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n,$$

radi bi izračunali njegovo vrednost v točki x po Hornerjevem algoritmu. Predpostavka je, da so koeficienti $a_0, \ldots, a_n$ in argument x predstavljiva števila. Analiziraj stabilnost izračuna.

Rešitev.

Algoritem 1: Eksaktni algoritem	**Algoritem 2:** Dejanski algoritem
$p_0 = a_0$;	$\widehat{p}_0 = a_0$;
for $i = 1, \ldots, n$ **do**	**for** $i = 1, \ldots, n$ **do**
$\quad p_i = p_{i-1} * x + a_i$;	$\quad \widehat{p}_i = (\widehat{p}_{i-1}(1+\alpha_i) + a_i)(1+\beta_i)$;
$\quad p = p_n$;	$\quad p = p_n$;
end	**end**

Z analizo zaokrožitvenih napak dobimo

$$\widehat{p} = a_0x^n(1+\gamma_0) + a_1x^{n-1}(1+\gamma_1) + \cdots + a_n(1+\gamma_n),$$

kjer je

$$\begin{aligned} 1+\gamma_0 &= (1+\alpha_1)\cdots(1+\alpha_n)(1+\beta_1)\cdots(1+\beta_n), \\ 1+\gamma_i &= (1+\alpha_{i+1})\cdots(1+\alpha_n)(1+\beta_i)\cdots(1+\beta_n), \quad i = 1, \cdots, n-1, \\ 1+\gamma_n &= 1+\alpha_n. \end{aligned}$$

Napako γ_0 obravnavamo posebej, saj pri izračunu $p_0 = a_0$, ne zagrešimo napake. Naredimo samo dva koraka. Dokaz koraka indukcije je preprost in prepuščen bralcu.

$$\widehat{p}_1 = (\widehat{p}_0(1+\alpha_1) + a_1)(1+\beta_1) = a_0x(1+\alpha_1)(1+\beta_1) + a_1(1+\beta_1),$$

$$\widehat{p}_2 = (\widehat{p}_1(1+\alpha_2) + a_2)(1+\beta_1) = a_0x^2(1+\alpha_1)(1+\alpha_2)(1+\beta_1)(1+\beta_2) + a_1x(1+\alpha_2)(1+\beta_1)(1+\beta_2) + a_2(1+\beta_2).$$

Narediti je treba še korak indukcje po stopnji polinoma.

Ocenimo lahko $|\gamma_0| \leq 2nu$ in $|\gamma_i| \leq (2(n-i)+1)\,u$ za $i = 1, \ldots, n$. Tukaj smo upoštevali, da se relativne napaka seštejejo, člene višjega reda zanemarimo.

Računanje vrednosti polinoma je obratno stabilno, saj se izračunane vrednosti ujemajo z eksaktno vrednostjo bližnjega polinoma, ki ima koeficiente $a_i(1+\gamma_i)$ namesto a_i, pri nespremenjenem argumentu x.

Iz absolutne napake $\widehat{p} - p = a_0 x^n \gamma_0 + a_1 x^{n-1}\gamma_1 + \cdots a_n\gamma_n$, sledi ocena

$$|\widehat{p} - p| \leq 2nu \left(|a_0||x^n| + |a_1||x^{n-1}| + \cdots + |a_n| \right),$$

od tod pa

$$\frac{|\widehat{p} - p|}{|p|} \leq \frac{2nu\left(|a_0||x^n| + |a_1||x^{n-1} + \cdots + |a_n|\right)}{|a_0x^n + \cdots + a_n|}.$$

Računanje vrednosti polinoma po Hornerjevem algoritmu ni direktno stabilno. Težave pa, podobno kot pri računanju skalarnega produkta, lahko pričakujemo, če je vsota blizu 0, členi v vsoti pa niso enako predznačeni.

Modelni primer je Wilkinsov zgled za polinom $(x-2)^{19}$. Oglejte si demonstracijski zgled `horner_wilkinson.m` v Matlabu. ∎

Naloga 2.3 *Podani sta dve približno enaki števili $x = 76.54320$ in $y = 76.54311110$ v sistemu $P(10, 7, ...)$. Izračunaj relativne napake $\frac{x-\mathrm{fl}(x)}{x}$, $\frac{y-\mathrm{fl}(y)}{y}$ in $\frac{z-\mathrm{fl}(z)}{z}$, kjer je $z = x - y$.*

Rešitev. Števila so že podana v desetiškem sistemu, zato moramo poskrbeti le, da bo mantisa dolžine 7 in število zaokroženo.

$$\mathrm{fl}(76.54320) = 0.7654320 * 10^2, \qquad \mathrm{fl}(76.54311110) = 0.7654311 * 10^2.$$

$$\frac{x - \mathrm{fl}(x)}{x} = 0, \frac{y - \mathrm{fl}(y)}{y} = 0.000000014.$$

Izračunajmo še $z = x - y = 0.0000889$ in $\mathrm{fl}(z) = \mathrm{fl}(\mathrm{fl}(x) - \mathrm{fl}(y)) = \mathrm{fl}(0.7654320 * 10^2 - 0.7654311 * 10^2) = 0.0000009 * 10^2$. Torej velja $\frac{z-\mathrm{fl}(z)}{z} = \frac{0.0000889-0.00009}{0.0000889} = -0.012373453$. Relativna napaka je velika.

∎

Naloga 2.4 *Pretvori naslednje izraze v stabilno obliko:*

(i). $\sqrt{1+x} - 1$ *za majhne* x,

(ii). $\sqrt{x^2+x} - x$ *za velike* x,

(iii). $\tan(x) - \sin(x)$ *za majhne* x.

Rešitev.

(i). Znebiti se moremo odštevanja dveh približno enakih števil,

$$\sqrt{1+x}-1=\frac{(\sqrt{1+x}-1)(\sqrt{1+x}+1)}{\sqrt{1+x}+1}=\frac{x}{1+\sqrt{1+x}}.$$

Zadnji izraz je v stabilni obliki.

(ii). Probleme nam lahko povzroča overflow, ko hočemo izračunati x^2.

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

(iii). Odštevamo dve približno enaki števili, saj velja $\sin(x)=\tan(x)\approx x$.

$$\tan(x)-\sin(x)=\frac{\sin(x)}{\cos(x)}-\sin(x)=\frac{\sin(x)(1-\cos(x))}{\cos(x)}=\frac{2\sin^2(x/2)\sin(x)}{\cos(x)}.$$

Upoštevali smo $\cos(2(x/2))=\cos^2(x/2)-\sin^2(x/2)$, $1=\cos^2(x/2)+\sin^2(x/2)$ in $\sin(x)=2\sin(x/2)\cos(x/2)$.

■

Primer 2.1 *Stabilno računanje ničel kvadratnega polinoma ax^2+bx+c. Rešitev kvadratne enačbe ax^2+bx+c je enaka*

$$x_{1,2}=\frac{-b\pm\sqrt{b^2-4ac}}{2a}=\frac{2c}{-b\mp\sqrt{b^2-4ac}}.$$

Drugi izraz uporabimo, ko gre $a\to 0$, saj potem velja $x_1\to-\frac{c}{b}$ in $x_2\to\infty$. Eno ničlo lahko zmeraj izračunamo stabilno, drugo pa računamo preko Vietovih formul $x_1x_2=\frac{c}{a}$, $x_1+x_2=-\frac{b}{q}$. Naslednji izračun je stabilen:

$$q=-\frac{1}{2}\left(b+\operatorname{sign}(b)\sqrt{b^2-4ac}\right),$$

$$x_1=\frac{q}{a}\qquad in\qquad x_2=\frac{c}{ax_1}.$$