

1 Numerično računanje

1.1 Predstavitev števil

Predstavljivo število x v sistemu $P(b, t, L, U)$ je zapisano kot $x = \pm m \cdot b^e$, kjer je b baza, e eksponent v mejah $L \leq e \leq U$ in

$$m = 0.c_1 \dots c_t, \quad 0 \leq c_i \leq b-1, \quad i = 1, \dots, t,$$

mantisa. Pri tem zahtevamo, da je $c_1 \neq 0$, razen kadar je $e = L$. Predstavljiva števila s $c_1 \neq 0$ imenujemo normalizirana, ostala so denormalizirana.

Naloga 1.1. Zapišite vsa normalizirana števila v sistemu $P(2, 3, -1, 3)$. Katera ležijo na intervalu $(0, 1)$? Koliko je denormaliziranih števil?

Rešitev. Normalizirana števila v sistemu $P(2, 3, -1, 3)$ so

$$\pm 0.100_2 \cdot 2^e, \quad \pm 0.101_2 \cdot 2^e, \quad \pm 0.110_2 \cdot 2^e, \quad \pm 0.111_2 \cdot 2^e$$

za $e \in \{-1, 0, 1, 2, 3\}$ oziroma

$\pm 0.2500,$	$\pm 0.5000,$	$\pm 1.0000,$	$\pm 2.0000,$	$\pm 4.0000,$
$\pm 0.3125,$	$\pm 0.6250,$	$\pm 1.2500,$	$\pm 2.5000,$	$\pm 5.0000,$
$\pm 0.3750,$	$\pm 0.7500,$	$\pm 1.5000,$	$\pm 3.0000,$	$\pm 6.0000,$
$\pm 0.4375,$	$\pm 0.8750,$	$\pm 1.7500,$	$\pm 3.5000,$	± 7.0000

v desetiškem zapisu. Na intervalu $(0, 1)$ ležijo števila s pozitivnim predznakom pri $e = -1$ in $e = 0$. Denormalizirana števila so določena z mantisami 0.001, 0.010 in 0.011 ter najmanjšim eksponentom $e = -1$. Torej jih je vsega skupaj šest (tri pozitivna in tri negativna).

Naloga 1.2. V Matlabu generirajte vsa predstavljiva števila iz množice $P(5, 4, -5, 5)$ in jih uredite po velikosti od najmanjšega do največjega. Nato poiščite odgovore na spodnja vprašanja.

1. Kakšen je delež denormaliziranih števil?
2. Koliko normaliziranih števil je manjših od π ?
3. Kakšen je povprečni razmik med zaporednimi predstavljivimi števili, ki se od π absolutno razlikujejo za manj kot 1?

Rešitev. Najprej sestavimo program, ki izračuna seznam predstavljivih (**X**), normaliziranih (**Xn**) in denormaliziranih (**Xdn**) števil v danem sistemu.

```
% sistem
b = 5; t = 4; L = -5; U = 5;

% mantise
c = 0:b-1;
M = zeros(b^t, 1);
```

```

i = 1;
for c1 = c
    for c2 = c
        for c3 = c
            for c4 = c
                M(i) = (b.^(1:t))*[c1; c2; c3; c4];
                i = i+1;
            end
        end
    end
end

% normalizirana števila
d = U-L+1;
bm = b^(t-1);
Xpn = zeros((b-1)*bm, d);
for i = 0:d-1
    Xpn(:,i+1) = M(bm+1:end) * b^(L+i);
end
Xpn = Xpn(:);
Xn = [-Xpn(end:-1:1); Xpn];

% denormalizirana števila
Xpdn = M(2:b^(t-1)) * b^L;
Xdn = [-Xpdn(end:-1:1); Xpdn];

% predstavljljiva števila (brez 0, Inf, -Inf in NaN)
X = [Xn(1:end/2); Xdn(1:end/2); Xpdn; Xpn];

```

Za vajo poskusite začetni del zgornjega programa nadgraditi tako, da izračuna vse mantise splošne dolžine t .

1. Delež denormaliziranih števil izračunamo tako, da njihovo število delimo s številom vseh predstavljljivih števil. Rezultat je približno 2.2%.
2. Število normaliziranih števil manjših od π lahko preštejemo z ukazom `sum(Xn<pi)`. Dobimo 8768.
3. Da dobimo predstavljljiva števila, ki se od π razlikujejo za manj kot ena, uporabimo ukaz `S = X(abs(X-pi)<1)`. Nato uporabimo vgrajeni funkciji `mean` in `diff`, da izračunamo povprečni razmik `mean(diff(S))`, ki je enak 0.008.

Število x , ki ni vsebovano v danem sistemu $P(b, t, L, U)$, je nepredstavljljivo. Nadomestimo ga s številom $\text{fl}(x)$, ki je bodisi največje predstavljljivo število, manjše od x , bodisi najmanjše predstavljljivo število, večje od x . Število $\text{fl}(x)$ navadno določimo z zaokroževanjem x . S tem zagotovimo, da ob pogoju, da $|x|$ leži na intervalu med najmanjšim in največjim pozitivnim predstavljljivim številom, velja $\text{fl}(x) = x(1 + \delta)$, kjer je δ število z lastnostjo, da je $|\delta|$ manjša od osnovne zaokrožitvene napake $u = b^{1-t}/2$.

Naloga 1.3. Katero je največje število v množici $P(5, 4, -5, 5)$, ki je manjše od π , in katero je najmanjše število, ki je večje od π ? Katero izmed teh dveh števil je $\text{fl}(\pi)$?

Rešitev. S pomočjo programa iz naloge 1.2 lahko odgovore na vprašanja poiščemo s spodnjimi ukazi.

```
x = pi; % 3.1416

xl = X(find(X<pi,1,'last')); % 3.1360
xf = X(find(X>pi,1,'first')); % 3.1440

if xf-x <= x-xl
    flx = xl;
else
    flx = xf; % 3.1440
end
```

Naloga 1.4. Predstavite število $x = 47.712$ v dvojiškem zapisu in z zaokroževanjem poiščite njegovo najbližje predstavljivo število $\text{fl}(x)$ v sistemu $P(2, 9, -10, 10)$. Preverite, da je relativna napaka $|\text{fl}(x) - x| / |x|$ manjša od osnovne zaokrožitvene napake.

Rešitev. Dvojiški zapis celega oziroma decimalnega dela x dobimo z deljenjem oziroma z množenjem z 2,

$$\begin{array}{ll} 47 = 23 \cdot 2 + 1, & 0.712 \cdot 2 = 0.424 + 1, \\ 23 = 11 \cdot 2 + 1, & 0.424 \cdot 2 = 0.848 + 0, \\ 11 = 5 \cdot 2 + 1, & 0.848 \cdot 2 = 0.696 + 1, \\ 5 = 2 \cdot 2 + 1, & 0.696 \cdot 2 = 0.392 + 1, \\ 2 = 1 \cdot 2 + 0, & 0.392 \cdot 2 = 0.784 + 0, \\ 1 = 0 \cdot 2 + 1, & 0.784 \cdot 2 = 0.568 + 1, \dots \end{array}$$

Od tod sledi $47 = 101111_2$ (ostanke v levem stolpcu prepisemo od spodaj navzgor) in $0.712 = 0.101101\dots_2$ (celi del v desnem stolpcu prepisemo od zgoraj navzdol). Torej je

$$x = 0.101111101101\dots_2 \cdot 2^6 \quad \text{in} \quad \text{fl}(x) = 0.101111110_2 \cdot 2^6.$$

Ker je

$$\begin{aligned} |\text{fl}(x) - x| &= \left| (0.101111110_2 + 2^{-9}) - (0.101111101_2 + 1.01\dots_2 \cdot 2^{-10}) \right| \cdot 2^6 \\ &< \left| 2^{-9} - 2^{-10} \right| \cdot 2^6 \\ &= 2^{-4}, \end{aligned}$$

je relativna napaka $|\text{fl}(x) - x| / |x|$ manjša od 0.0014, kar je manj od osnovne zaokrožitvene napake $2^{1-9}/2 \approx 0.0020$.

Pravimo, da je število x zapisano v enojni natančnosti, če je predstavljeno s številom $\text{fl}(x)$ iz množice $P(2, 24, -125, 128)$. V računalniškem spominu je tako število shranjeno v 32 bitih. Če je normalizirano, je podano v obliki

$$\text{fl}(x) = (-1)^s (1 + f) \cdot 2^{\tilde{e}-127},$$

kjer $s \in \{0, 1\}$ določa predznak (en bit), $\tilde{e} \in \{1, 2, \dots, 2^8 - 1\}$ eksponent (osem bitov) in $f = 0.c_2 c_3 \dots c_{24}$ del mantise (23 bitov). Na podoben način so s 64 biti opisana števila iz $P(2, 53, -1021, 1024)$, ki določajo dvojno natančnost.

Naloga 1.5. Dokažite, da je

$$0.1 = \sum_{i=1}^{\infty} (2^{-4i} + 2^{-4i-1})$$

in določite $\text{fl}(0.1)$ za 0.1 v enojni natančnosti. Kako je to število v tem formatu predstavljeno v računalniku?

Rešitev. Vrsto izračunamo s prevedbo na geometrijsko vrsto

$$\sum_{i=1}^{\infty} (2^{-4i} + 2^{-4i-1}) = \left(1 + \frac{1}{2}\right) \sum_{i=1}^{\infty} (2^{-4})^i = \frac{3}{2} \cdot \frac{2^{-4}}{1 - 2^{-4}} = \frac{1}{10}$$

in s tem dokažemo, da lahko število 0.1 predstavimo v želeni obliki. Iz tega rezultata sledi, da je $0.1 = 0.0001\overline{1}_2$. Ker ima 0.1 v dvojiški bazi neskončen decimalni zapis, $\text{fl}(0.1)$ dobimo z zaokroževanjem. Na podlagi

$$0.1 = 0.1100110011001100110011001 \dots_2 \cdot 2^{-3}$$

sklepamo, da je

$$\text{fl}(0.1) = 0.110011001100110011001101_2 \cdot 2^{-3}$$

oziroma

$$\text{fl}(0.1) = (-1)^0 (1 + 0.10011001100110011001101_2) \cdot 2^{123-127},$$

kar pomeni, da število $\text{fl}(0.1)$ opišemo z biti 0, 01111011 in 10011001100110011001101, ki po vrsti določajo s , \tilde{e} in f . Rezultat v Matlabu preverimo s pomočjo ukaza `single(0.1)`, ki vrne $\text{fl}(0.1)$ za enojno natančnost.

```
x = 0.1;
flx = (repmat([1 1 0 0],1,6)+[zeros(1,23) 1])*2.^(4:27)';
single(x)-x      % 1.4901161e-09
single(x)-flx    % 0
```

1.2 Napake pri računanju

Pri numerični matematiki smo soočeni z napakami v različnih fazah računanja.

1. Navadno pride do napake že pri pripravi vhodnih podatkov na začetku računanja. Napaka, ki je razlika med izvedbo računa s pravimi in dejanskimi podatki, se imenuje *neodstranljiva napaka*.
2. Pri reševanju problema smo se zaradi njegove težavnosti ali računske zahtevnosti pogosto primorani sprijazniti z njegovim približnim reševanjem. Tako namesto originalnega problema rešimo njegov bližnji problem in napaka, ki pri tem nastane, se imenuje *napaka metode*.

3. Nazadnje moramo v zakup vzeti še *zaokrožitveno napako*, ki je posledica zaokroževanja na vsakem računskem koraku izvedbe metode, saj rezultat vsake računske operacije zaokrožujemo na najbližje predstavljivo število.

Seštevek vseh treh napak je celotna napaka izračuna.

Naloga 1.6. Funkcija f je podana s predpisom $f(x) = \sqrt{1+x}$. Izračunajte vrednost $f(x)$ za $x = 1/13$ v sistemu $P(10, 5, -10, 10)$.

1. Ocenite neodstranljivo napako, ki nastane pri predstavitvi x .
2. Namesto funkcije f uporabite Taylorjev polinom funkcije f stopnje 2, ki ga dobite z razvojem okoli točke 0. Ocenite napako metode.
3. Vrednost Taylorjevega polinoma izračunajte s Hornerjevim postopkom. S pomočjo izračuna vrednosti v dvojni natančnosti ocenite zaokrožitveno napako, ki nastane zaradi računanja v dani aritmetiki.

Rešitev. Ocenimo vsako izmed napak, ki se pojavi pri izvedbi postopka.

1. Najprej ocenimo neodstranljivo napako, ki nastane zaradi predstavitve x v predpisanem sistemu. Ker je $x = 0.0769230\dots$, je $\bar{x} = \text{fl}(x) = 0.76923 \cdot 10^{-1}$. Neodstranljiva napaka D_n je podana z $D_n = f(x) - f(\bar{x})$. Njeno absolutno vrednost lahko s pomočjo izreka o povprečni vrednosti in ocene za relativno napako predstavitve x z \bar{x} v dani aritmetiki ocenimo z

$$|D_n| = |f(x) - f(\bar{x})| \leq \max_{\xi \in [0,1]} |f'(\xi)| |x - \bar{x}| < 0.5 \cdot 10^{1-5}/2 = 0.25 \cdot 10^{-4}.$$

2. Napaka metode nastane, ker namesto s funkcijo f računamo s približkom, ki ga dobimo s pomočjo razvoja f v Taylorjevo vrsto. Konkretno, funkcijo f zamenjamo s polinomom $g(x) = 1 + x/2 - x^2/8$. Napaka metode je podana z $D_m = f(\bar{x}) - g(\bar{x})$, njeno absolutno vrednost pa lahko ocenimo z

$$|D_m| = |f(\bar{x}) - g(\bar{x})| \leq \frac{1}{3!} \max_{\xi \in [0,1]} |f'''(\xi)| \bar{x}^3 < \bar{x}^3/16 < 0.29 \cdot 10^{-4}.$$

3. Označimo $g(x) = a_0 + a_1x + a_2x^2$. Računanje polinoma g v točki \bar{x} s Hornerjevim postopkom

$$b_2 = a_2, \quad b_i = b_{i+1}\bar{x} + a_i, \quad i = 1, 0, \quad g(\bar{x}) = b_0,$$

v predpisanem sistemu poteka na sledeč način.

i	a_i	$b_{i+1} \cdot \bar{x}$	$c_i = \text{fl}(b_{i+1} \cdot \bar{x})$	$c_i + a_i$	$b_i = \text{fl}(a_i + c_i)$
2	-0.125				$-0.12500 \cdot 10^0$
1	0.5	-0.009615375	$0.96154 \cdot 10^{-2}$	0.4903846	$0.49038 \cdot 10^0$
0	1	0.03772150074	$0.37722 \cdot 10^{-1}$	1.037722	$0.10377 \cdot 10^1$

Z računanjem v dvojni natančnosti dobimo, da je $g(\bar{x})$ približno 1.0377219, torej je zaokrožitvena napaka D_z po absolutni vrednosti manjša od $0.22 \cdot 10^{-4}$.

Iz obravnave napak sledi, da je celotna napaka manjša od 10^{-4} .

Naloga 1.7. Dani sta diferenčni enačbi

$$\begin{aligned} a_n &= \frac{5}{2}a_{n-1} - a_{n-2}, & n &= 2, 3, \dots, & a_0 &= 1, \quad a_1 = \frac{1}{2}, \\ b_n &= \frac{10}{3}b_{n-1} - b_{n-2}, & n &= 2, 3, \dots, & b_0 &= 1, \quad b_1 = \frac{1}{3}. \end{aligned}$$

1. Z nastavkoma $a_n = \lambda^n$, $\lambda \in \mathbb{R}$, in $b_n = \mu^n$, $\mu \in \mathbb{R}$, poiščite točni rešitvi diferenčnih enačb.
2. V Matlabu generirajte seznama $\mathbf{a} = (a_0, a_1, \dots, a_{50})$ in $\mathbf{b} = (b_0, b_1, \dots, b_{50})$ ter z ukazom `scatter` narišite točke (n, a_n) in (n, b_n) , $n = 0, 1, \dots, 50$. Ali se elementi seznamov ujemajo s točnimi vrednostmi? Pojasnite, zakaj da oziroma ne.
3. Omilite napake, ki nastanejo pri izračunu elementov v seznamu \mathbf{b} tako, da elemente generirate v obratnem vrstnem redu pri začetnih podatkih $b_{50} = 0$ in $b_{49} = 1$ ter jih na koncu skalirate s konstanto, ki zagotovi, da bo $b_0 = 1$. Primerjajte dobljene vrednosti s točnimi.

Rešitev.

1. Uporabimo nastavka za $a_n = \lambda^n$ in $b_n = \mu^n$. Ker sta enačbi dvočlenski, dobimo v obeh primerih kvadratni enačbi z rešitvama $\lambda_1 = 1/2$, $\lambda_2 = 2$ in $\mu_1 = 1/3$, $\mu_2 = 3$. Od tod sledi, da sta splošni rešitvi oblike

$$a_n = A \left(\frac{1}{2}\right)^n + B 2^n, \quad b_n = C \left(\frac{1}{3}\right)^n + D 3^n,$$

kjer so A, B, C, D konstante, ki jih določimo iz začetnih pogojev. Dobimo $a_n = 1/2^n$ in $b_n = 1/3^n$. Vrednosti a_n in b_n z naraščajočim n torej padata proti 0.

2. Elemente seznamov izračunamo na podlagi rekurzivnih formul, ki določata diferenčno enačbo.

```
% seznam a
a = [1 0.5 zeros(1,49)];
for n = 3:51
    a(n) = 5*a(n-1)/2 - a(n-2);
end

% seznam b
b = [1 1/3 zeros(1,49)];
for n = 3:51
    b(n) = 10*b(n-1)/3 - b(n-2);
end
```

Iz diagramov na slikah 1a in 1b je razvidno, da vrednosti a_n padajo proti 0, kot je pričakovano z obzirom na točno rešitev diferenčne enačbe. Po drugi strani pa so izračunane vrednosti b_n pri večjih n povsem napačne (rastejo v pozitivno ali negativno smer, odvisno od vrstnega reda operacij pri generiranju seznama \mathbf{b}). Razlog, da v prvem primeru dobimo točne rezultate, v drugem pa napačne, se skriva v tem, da je v prvem primeru rezultat vsake računske operacije predstavljlivo število v dvojni

natančnosti, medtem ko v drugem primeru operiramo z nepredstavljivimi števili, ki jih vseskozi zaokrožujemo na predstavljiva. Tako že za začetni podatek namesto $b_1 = 1/3$ uporabimo $\tilde{b}_1 = \text{fl}(b_1) = b_1(1+\delta)$, kjer je δ sicer po absolutni vrednosti majhno število, a točna rešitev \tilde{b}_n diferenčne enačbe z začetnima podatkom b_0 in \tilde{b}_1 je

$$\tilde{b}_n = \left(1 - \frac{\delta}{8}\right) \left(\frac{1}{3}\right)^n + \frac{\delta}{8} 3^n$$

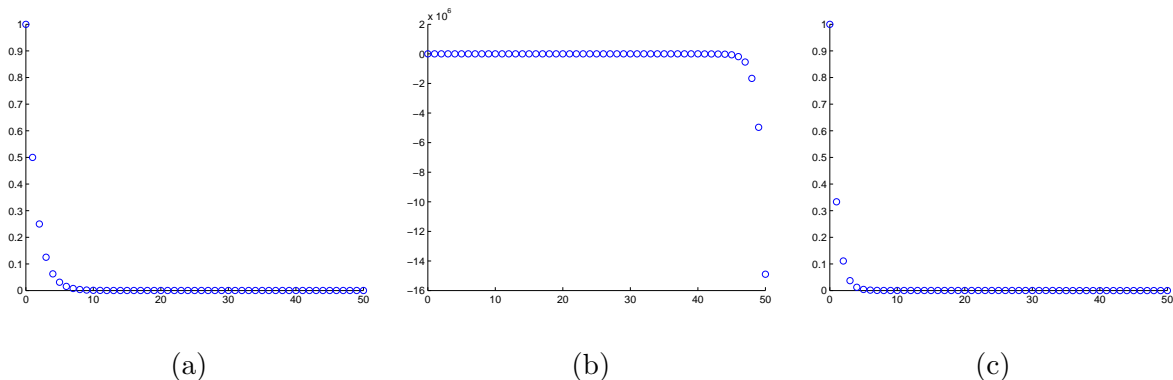
in vpliv faktorja 3^n se pri večjih vrednostih n močno pozna.

3. Iz rekurzivne zveze za vrednosti b_n izrazimo b_{n-2} in z zamikom indeksov dobimo

$$b_n = -b_{n+2} + \frac{10}{3}b_{n+1}.$$

Vzamemo $b_{50} = 0$ in $b_{49} = 1$ ter elemente seznama \mathbf{b} generiramo v obratnem vrstnem redu. Na koncu vse elemente seznama delimo z b_0 in s tem zagotovimo, da je v rezultatu $b_0 = 1$. Vrednosti tega seznama se zelo dobro ujemajo s točnimi, kar potrjuje diagram na sliki 1c.

```
% seznama b generiran v obratnem vrstnem redu
rb = [zeros(1,49) 1 0];
for n = 49:-1:1
    rb(n) = -rb(n+2) + 10*rb(n+1)/3;
end
b = rb/rb(1);
```



Slika 1: Prikaz rezultatov pri rekurzivnem računanju števil v nalogi 1.7.

1.3 Stabilnost izračunov

Pri numeričnem računanju stabilnost obravnavamo v različnih kontekstih. V osnovi nas zanima razlika med točno vrednostjo in izračunanim približkom: to je *direktna napaka*. Če je ta pri vseh začetnih podatkih majhna, pravimo, da je metoda izračuna direktno stabilna. Ocenjevanje direktne napake je ponavadi težavno, zato si pri analizi pomagamo z *obratno napako*: to je razlika med dejanskimi vhodnimi podatki in vhodnimi podatki, pri katerih bi izračunani približek pri dejanskih podatkih predstavljal točno vrednost. Če je obratna napaka majhna pri vseh vhodnih podatkih, je metoda izračuna obratno stabilna. Z obratno stabilnostjo lahko dokažemo direktno stabilnost metode, kadar je problem neobčutljiv.

Naloga 1.8. Naj bosta x in y predstavljam realni števili. Vrednost $z = x^2 - y^2$ izračunajte po standardu IEEE na dva načina:

1. $z = x^2 - y^2$,
2. $z = (x - y)(x + y)$.

V obeh primerih ocenite relativno napako ter obravnavajte direktno in obratno stabilnost izračuna. Predpostavite, da pri računanju ne pride do prekoračitev.

Rešitev. V prvem primeru lahko izračun interpretiramo kot računanje skalarnega produkta vektorjev (x, y) in $(x, -y)$, zato direktna stabilnost ni zagotovljena. V drugem primeru imamo produkt dveh predstavljaljivih števil in z obzirom na to pričakujemo, da je izračun direktno stabilen. Utemeljimo obe opazki bolj natančno.

1. Pri kvadriranju členov x in y dobimo $a_1 = x^2(1 + \alpha_1)$ in $a_2 = y^2(1 + \alpha_2)$, pri čemer je $|\alpha_1| \leq u$ in $|\alpha_2| \leq u$ ter u označuje osnovno zaokrožitveno napako. Ko člena odštejemo, namesto vrednosti z dobimo

$$\bar{z} = (a_1 - a_2)(1 + \beta) = x^2(1 + \alpha_1)(1 + \beta) - y^2(1 + \alpha_2)(1 + \beta)$$

za $|\beta| \leq u$ oziroma

$$\bar{z} = x^2(1 + \delta_1) - y^2(1 + \delta_2)$$

za neki konstanti δ_1 in δ_2 z lastnostjo

$$(1 - u)^2 \leq 1 + \delta_i \leq (1 + u)^2, \quad i = 1, 2.$$

Ker je vrednost u majhna, lahko na podlagi tega ocenimo, da je $|\delta_i| \leq 2u$. Vrednost $\bar{z} = (x\sqrt{1 + \delta_1})^2 - (y\sqrt{1 + \delta_2})^2$ torej ustreza vrednosti z pri malo zmotenih podatkih, zato je izračun obratno stabilen. Po drugi strani pa direktna stabilnost ni zagotovljena, saj je

$$\frac{|\bar{z} - z|}{|z|} = \frac{|\delta_1 x^2 - \delta_2 y^2|}{|x^2 - y^2|}$$

in v primeru, ko sta δ_1 in δ_2 nasprotno predznačeni, velja

$$\frac{|\bar{z} - z|}{|z|} \geq \min\{|\delta_1|, |\delta_2|\} \frac{x^2 + y^2}{|x^2 - y^2|}.$$

Torej je lahko relativna napaka za $x \approx y$ velika.

2. Namesto vsote oziroma razlike števil x in y izračunamo $(x + y)(1 + \alpha_1)$ in $(x - y)(1 + \alpha_2)$, pri čemer sta vrednosti $|\alpha_1|$ in $|\alpha_2|$ manjši ali enaki osnovni zaokrožitveni napaki u . Z množenjem teh dveh členov namesto z dobimo

$$\bar{z} = (x + y)(x - y)(1 + \alpha_1)(1 + \alpha_2)(1 + \beta)$$

za $|\beta| \leq u$. Sledi, da je $\bar{z} = (x\sqrt{1 + \delta})^2 - (y\sqrt{1 + \delta})^2$ za neko število δ z lastnostjo $|\delta| \leq 3u$, kar pomeni, da je izračun obratno stabilen. Poleg tega pa je tudi direktno stabilen, saj je

$$\frac{|\bar{z} - z|}{|z|} \leq \frac{|(x^2 - y^2)\delta|}{|x^2 - y^2|} = |\delta| \leq 3u.$$

Naloga 1.9. Dan je polinom

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0 = a_n (x - x_n)(x - x_{n-1}) \dots (x - x_1),$$

kjer so koeficienti $a_n, a_{n-1}, \dots, a_1, a_0$ in ničle x_n, x_{n-1}, \dots, x_1 predstavljava realna števila. Obravnavajte relativni napaki pri računanju vrednosti polinoma p v točki x s Hornerjevim postopkom ter z množenjem faktorjev, ki jih določajo ničle polinoma. Ali je kateri izmed postopkov direktno stabilen?

Rešitev. Pri Hornerjevem postopku izračun vrednosti polinoma p poteka v zaporedju

$$p(x) = ((\dots((a_n x + a_{n-1})x + a_{n-2})x + \dots)x + a_1)x + a_0.$$

Pri vsaki operaciji seštevanja in množenja pride do zaokrožitvene napake, ki je relativno manjša od osnovne zaokrožitvene napake u . Zato namesto $y = p(x)$ izračunamo

$$\hat{y} = a_n x^n (1 + \delta_n) + a_{n-1} x^{n-1} (1 + \delta_{n-1}) + \dots + a_1 x (1 + \delta_1) + a_0 (1 + \delta_0),$$

kjer za števila δ_i , $i = 0, 1, \dots, n$, ocenjujemo, da so po absolutni vrednosti manjša od $2nu$ (pri vrednostih δ_i , $i = 0, 1, \dots, n-1$, smo lahko natančnejši, velja $|\delta_i| \leq (2i+1)u$). To pomeni, da za relativno napako izračuna velja

$$\frac{|\hat{y} - y|}{|y|} \leq \frac{2nu(|a_n||x^n| + |a_{n-1}||x^{n-1}| + \dots + |a_1||x| + |a_0|)}{|a_n x^n + a_{n-1} x^{n-1} + \dots + a_1 x + a_0|}.$$

Čeprav smo napako omejili navzgor in je lahko ta ocena pregroba, vseeno jasno odraža, da je za točke x blizu ničle polinoma relativna napaka lahko velika, če so koeficienti polinoma različno predznačeni. Postopek izračuna torej ni direktno stabilen.

Če obstaja faktorizacija polinoma na linearne faktorje, lahko izračun vrednosti opravimo stabilneje z množenjem faktorjev. Najprej izračunamo $x - x_i$, $i = 1, 2, \dots, n$, pri čemer zaradi zaokroževanja dobimo $(x - x_i)(1 + \alpha_i)$ za neka števila α_i , ki so po absolutni vrednosti manjša od u . Nato vsako izmed n množenj botruje še k relativni napaki, manjši od u . Torej namesto $y = p(x)$ izračunamo

$$\tilde{y} = a_n (x - x_n) \dots (x - x_1) (1 + \delta) = y(1 + \delta),$$

pri čemer je $|\delta| \leq 2nu$. Torej je relativna napaka

$$\frac{|\tilde{y} - y|}{|y|} = \frac{|y\delta|}{|y|} = |\delta| \leq 2nu$$

in izračun je direktno stabilen.