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

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

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

$$\begin{array}{lllll}
\pm 0.2500, & \pm 0.5000, & \pm 1.0000, & \pm 2.0000, & \pm 4.0000, \\
\pm 0.3125, & \pm 0.6250, & \pm 1.2500, & \pm 2.5000, & \pm 5.0000, \\
\pm 0.3750, & \pm 0.7500, & \pm 1.5000, & \pm 3.0000, & \pm 6.0000, \\
\pm 0.4375, & \pm 0.8750, & \pm 1.7500, & \pm 3.5000, & \pm 7.0000
\end{array}$$

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];

% predstavljiva 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 predstavljivih 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 predstavljiva 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 nepredstavljivo. Nadomestimo ga s številom $\mathrm{fl}(x)$, ki je bodisi največje predstavljivo število, manjše od x, bodisi najmanjše predstavljivo število, večje od x. Število $\mathrm{fl}(x)$ navadno določimo z zaokroževanjem x. S tem zagotovimo, da ob pogoju, da $|x|$ leži na intervalu med najmanjšim in največjim pozitivnim predstavljivim številom, velja $\mathrm{fl}(x) = x(1 + \delta)$, 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,

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

Od tod sledi $47 = 101111_2$ (ostanke v levem stolpcu prepišemo od spodaj navzgor) in $0.712 = 0.101101\ldots_2$ (celi del v desnem stolpcu prepišemo od zgoraj navzdol). Torej je

$$x = 0.101111101101\ldots_2\cdot 2^6 \quad \text{in} \quad \text{fl}(x) = 0.101111110_2\cdot 2^6.$$

Ker je

$$\begin{aligned} |\text{fl}(x)-x| &= \left|(0.101111101_2 + 2^{-9}) - (0.101111101_2 + 1.01\ldots_2\cdot 2^{-10})\right|\cdot 2^6\\ &< \left|2^{-9}-2^{-10}\right|\cdot 2^6\\ &= 2^{-4}, \end{aligned}$$

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

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

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

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

in določite 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

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

in s tem dokažemo, da lahko število 0.1 predstavimo v želeni obliki. Iz tega rezultata sledi, da je $0.1 = 0.0\overline{0011}_2$. Ker ima 0.1 v dvojiški bazi neskončen decimalni zapis, fl(0.1) dobimo z zaokroževanjem. Na podlagi

$$0.1 = 0.110011001100110011001100\ldots_2 \cdot 2^{-3}$$

sklepamo, da je

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

oziroma

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

kar pomeni, da število 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 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 $x = 0.0769230...$, je $\overline{x} = \mathrm{fl}(x) = 0.76923 \cdot 10^{-1}$. Neodstranljiva napaka D_n je podana z $D_n = f(x) - f(\overline{x})$. Njeno absolutno vrednost lahko s pomočjo izreka o povprečni vrednosti in ocene za relativno napako predstavitve x z $\overline{x}$ v dani aritmetiki ocenimo z

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

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 $D_m = f(\overline{x}) - g(\overline{x})$, njeno absolutno vrednost pa lahko ocenimo z

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

3. Označimo $g(x) = a_0 + a_1 x + a_2 x^2$. Računanje polinoma g v točki $\overline{x}$ s Hornerjevim postopkom

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

v predpisanem sistemu poteka na sledeč način.

i	a_i	$b_{i+1} \cdot \overline{x}$	$c_i = \mathrm{fl}(b_{i+1} \cdot \overline{x})$	$c_i + a_i$	$b_i = \mathrm{fl}(a_i + c_i)$
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 $g(\overline{x})$ 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}.