Jan Grošelj

UVOD V NUMERIČNE METODE

Zbirka nalog z rešitvami

Predgovor

Pričujoče gradivo je delovna verzija zbirke nalog iz začetnih poglavij numerične matematike. Nastaja med izvajanjem vaj pri predmetu Uvod v numerične metode na univerzitetnem študijskem programu matematike na Fakulteti za matematiko in fiziko v Ljubljani.

Zbirka vsebinsko sledi izbranim poglavjem knjige [1]. Čeprav so med naloge vključeni povzetki obravnavanih tem, je za celovito razumevanje snovi neobhodno vzporedno branje knjige ali zapiskov s predavanj. Naloge, ki sestavljajo zbirko, prihajajo iz različnih virov, a v večini izhajajo iz omenjene knjige, kjer so (brez rešitev) zbrane ob koncu poglavij. K tem je dodanih nekaj izpitnih nalog in nalog, prirejenih po tuji literaturi. Kar nekaj nalog poleg razmisleka zahteva tudi računalniško implementacijo, saj numerične metode zaživijo šele na računsko zapletenih primerih. Računalniška koda, ki spremlja rešitve, je zapisana v programskem jeziku programa Matlab, osnove katerega so povzete v [2].

Ker je zbirka v nastajanju, gotovo vsebuje napake. Zato je bralec pozvan h kritičnemu branju. Poročila o napakah ali nejasnostih so zelo dobrodošla; posredujete jih lahko na elektronski naslov `jan.groselj@fmf.uni-lj.si`. Aktualna verzija zbirke je objavljena na spletnem naslovu

`https://www.fmf.uni-lj.si/~groseljj/gradiva/unm_zbirka_nalog.pdf`,

delni angleški prevod pa na spletnem naslovu

`https://www.fmf.uni-lj.si/~groseljj/gradiva/unm_zbirka_nalog_eng.pdf`.

Zadnja sprememba: 3. januar 2024

Kazalo

1. Numerično računanje **7**
1.1. Predstavitev števil . . . 7
1.2. Napake pri računanju . . . 14
1.3. Stabilnost izračunov . . . 17

2. Nelinearne enačbe **23**
2.1. Bisekcija . . . 23
2.2. Navadna iteracija . . . 25
2.3. Izpeljava in lastnosti metod . . . 32

3. Sistemi linearnih enačb **47**
3.1. Matrične norme in občutljivost matrik . . . 47
3.2. LU razcep . . . 56
3.3. Reševanje sistemov posebne oblike . . . 67

4. Sistemi nelinearnih enačb **75**
4.1. Jacobijeva iteracija . . . 75
4.2. Newtonova metoda . . . 79

5. Predoločeni sistemi enačb **85**
5.1. Normalni sistem . . . 85
5.2. QR razcep . . . 89

6. Lastne vrednosti matrik **101**
6.1. Schurova forma . . . 101
6.2. Potenčna metoda . . . 104
6.3. QR iteracija . . . 111

7. Polinomska interpolacija **115**
7.1. Lagrangeeva oblika . . . 115
7.2. Newtonova oblika . . . 119

8. Odvajanje in integriranje **127**
8.1. Pravila za odvajanje 127
8.2. Pravila za integriranje 129

9. Diferencialne enačbe **139**
9.1. Runge–Kutta metode 139
9.2. Veččlenske metode 144

Literatura **145**

1. Numerično računanje

V numerični matematiki se ukvarjamo z izvedbo matematičnih postopkov iz analize in algebre na računskih strojih. Ti prinašajo omejitve pri predstavitvi števil, saj smo z njimi omejeni na končno množico števil, s katerimi lahko operiramo. Te omejitve je treba upoštevati tudi v sami izvedbi postopkov in jih prilagoditi na način, da je negativen vpliv končne predstavitve kar se da majhen.

1.1. Predstavitev števil

Predstavljivo število x iz množice $P(b, t, L, U)$ je oblike $x = \pm m \cdot b^e$, kjer je $b \in \mathbb{N}$ baza, $e \in \mathbb{Z}$ eksponent v mejah $L \leq e \leq U$ za $L, U \in \mathbb{Z}$ in

$$m = 0.c_1 c_2 \ldots c_t, \qquad 0 \leq c_i \leq b-1, \quad i = 1, 2, \ldots, t,$$

mantisa dolžine $t \in \mathbb{N}$. Pri tem zahtevamo $c_1 \neq 0$, razen kadar je $e = L$. Predstavljiva števila s $c_1 \neq 0$ imenujemo normalizirana, ostala so denormalizirana.

Naloga 1.1. Zapišite vsa normalizirana števila iz množice $P(2, 3, -1, 3)$. Katera ležijo na intervalu $(0, 1)$? Koliko denormaliziranih števil vsebuje ta množica?

Rešitev. Normalizirana števila iz množice $P(2, 3, -1, 3)$ 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

± 0.2500,	± 0.5000,	± 1.0000,	± 2.0000,	± 4.0000,
± 0.3125,	± 0.6250,	± 1.2500,	± 2.5000,	± 5.0000,
± 0.3750,	± 0.7500,	± 1.5000,	± 3.0000,	± 6.0000,
± 0.4375,	± 0.8750,	± 1.7500,	± 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_2, 0.010_2 in 0.011_2 ter najmanjšim eksponentom $e = -1$. Torej jih je vsega skupaj šest (tri pozitivna in tri negativna): ± 0.0625, ± 0.1250 in ± 0.1875.

Naloga 1.2. Naj bo $t \in \mathbb{N}$ dolžina mantise in naj bosta $L, U \in \mathbb{Z}$, $L < 0 < U$, meji za eksponente, s katerimi je opisana množica predstavljivih števil $P(\cdot, t, L, U)$. Dokažite, da velja

$$P(2, t, L, U) \subseteq P(4, t, L, U) \subseteq P(2, 2t, 2L, 2U).$$

Ali pri kateri izmed relacij vsebovanosti velja enakost?

Rešitev. Število $x \in P(2, t, L, U)$ lahko predstavimo v obliki

$$x = \pm \left(c_1 \cdot 2^{-1} + c_2 \cdot 2^{-2} + \ldots + c_t \cdot 2^{-t}\right) \cdot 2^e$$

za $c_i \in \{0, 1\}$, $i = 1, 2, \ldots, t$, in $e \in \{L, L+1, \ldots, U\}$. Naj bosta t_1 in t_0 taki števili, da je

$$t = 2t_1 + t_0, \quad t_1 \leq \tfrac{1}{2}t, \quad t_0 \in \{0, 1\},$$

ter e_1 in e_0 taki števili, da je

$$e = 2e_1 + e_0, \quad |e_1| \leq \tfrac{1}{2}|e|, \quad |e_0| \in \{0, 1\}.$$

Potem je $t_1 < t$ in $L < e_1 < U$. Število x lahko zapišemo kot

$$x = \pm \left((2c_1 + c_2) \cdot 4^{-1} + \ldots + (2c_{2t_1-1} + c_{2t_1}) \cdot 4^{-t_1} + 2t_0 c_t \cdot 4^{-(t_1+1)}\right) \cdot 4^{e_1} \cdot 2^{e_0}.$$

Če je $e_0 = 0$, je iz tega že razvidno, da je $x \in P(4, t, L, U)$. Če je $e_0 \in \{-1, 1\}$, pa lahko zapis x preoblikujemo v

$$x = \pm \left(c_1 \cdot 4^{-1} + (2c_2 + c_3) \cdot 4^{-2} + \ldots + (2c_{2t_1} + t_0 c_t) \cdot 4^{-(t_1+1)}\right) \cdot 4^{e_1 + \frac{1}{2}(e_0+1)},$$

kar dokazuje, da je tudi v teh dveh primerih $x \in P(4, t, L, U)$.

Število $x \in P(4, t, L, U)$ lahko zapišemo v obliki

$$x = \pm \left(c_1 \cdot 4^{-1} + c_2 \cdot 4^{-2} + \ldots + c_t \cdot 4^{-t}\right) \cdot 4^e$$

za $c_i \in \{0, 1, 2, 3\}$, $i = 1, 2, \ldots, t$, in $e \in \{L, L+1, \ldots, U\}$. Za vsak i naj bosta $c_{i,1}, c_{i,0} \in \{0, 1\}$ taki števili, da je $c_i = 2c_{i,1} + c_{i,0}$. Potem je

$$x = \pm \left(c_{1,1} \cdot 2^{-1} + c_{1,0} \cdot 2^{-2} + \ldots + c_{t,1} \cdot 2^{-2t+1} + c_{t,0} \cdot 2^{-2t}\right) \cdot 2^{2e},$$

kar dokazuje, da je $x \in P(2, 2t, 2L, 2U)$.

Število $4^{-t} \cdot 4^L$ je vsebovano v $P(4, t, L, U)$, ni pa vsebovano v $P(2, t, L, U)$. Število $\sum_{i=1}^{2t+1} 2^{-i}$ je vsebovano v $P(2, 2t, 2L, 2U)$, ni pa vsebovano v $P(4, t, L, U)$. Torej pri nobeni izmed relacij vsebovanosti ne velja enakost.

Naloga 1.3. S pomočjo Matlaba generirajte vsa predstavljiva števila iz množice $P(5, 4, -5, 5)$ in jih uredite po velikosti od najmanjšega do največjega. Nato poiščite odgovore na naslednja vprašanja.

1. Kakšen je delež denormaliziranih števil?

2. Koliko normaliziranih števil je manjših od π?
3. Kakšen je povprečni razmik med zaporednima predstavljivima številoma, ki se od π absolutno razlikujeta za manj kot 1?

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

```
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 j = 1:d
    Xpn(:,j) = M(bm+1:end) * b^(L+j-1);
end
Xpn = Xpn(:);
Xn = [-Xpn(end:-1:1); Xpn];

% denormalizirana števila
Xpdn = M(2:bm) * 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];
```

1. Delež denormaliziranih števil izračunamo tako, da njihovo število delimo s številom vseh predstavljivih števil. Rezultat je približno 2.2%.
2. Normalizirana števila, manjša 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.

Poskusite začetni del programa nadgraditi tako, da izračuna vse mantise splošne dolžine `t`. Nato poiščite odgovore na zastavljena vprašanja še pri kaki drugi izbiri za `t`.

Število x, ki ni vsebovano v izbrani množici $P(b, t, L, U)$, 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. Pri zaokroževanju za $\mathrm{fl}(x)$ vzamemo tisto izmed teh dveh števil, ki je bližje x. S tem zagotovimo, da ob pogoju, da $|x|$ leži na intervalu med najmanjšim in največjim pozitivnim predstavljivim številom, velja $\mathrm{fl}(x) = x(1 + \delta)$, pri čemer je δ tako število, da je absolutno vrednost $|\delta|$ manjša od osnovne zaokrožitvene napake $u = b^{1-t}/2$.

Naloga 1.4. Katero je največje število v množici $P(5, 4, -5, 5)$, ki je manjše od π, in katero je najmanjše število, večje od π? Katero izmed teh dveh števil je $\mathrm{fl}(\pi)$?

Rešitev. S pomočjo programa iz naloge 1.3 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
```

Pri tem je treba poudariti, da izbira $\mathrm{fl}(x)$ po zgornjem postopku za splošen x ni korektna. V primeru, da je x enako oddaljen od obeh najbližjih predstavljivih števil, se za $\mathrm{fl}(x)$ vzame tistega, ki ima sodo zadnjo števko v mantisi.

Naloga 1.5. Predstavite število $x = 47.712$ v dvojiškem zapisu in z zaokroževanjem poiščite njegovo najbližje predstavljivo število $\mathrm{fl}(x)$ iz množice $P(2, 9, -10, 10)$. Preverite, da je relativna napaka $|\mathrm{fl}(x) - x| \,/\, |x|$ manjša od osnovne zaokrožitvene napake.

Rešitev. Dvojiški zapis celega oziroma decimalnega dela x dobimo z deljenjem ozi-

roma 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

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

Ker je

$$\mathrm{fl}(x) - x = \big((0.101111101_2 + 2^{-9}) - (0.101111101_2 + 1.01\ldots_2 \cdot 2^{-10})\big) \cdot 2^6,$$

velja ocena

$$|\mathrm{fl}(x) - x| < \left|2^{-9} - 2^{-10}\right| \cdot 2^6 = 2^{-4}.$$

To pomeni, da je relativna napaka $|\mathrm{fl}(x) - x| \,/\, |x|$ 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 $\mathrm{fl}(x)$ iz množice $P(2, 24, -125, 128)$. V računalniškem spominu je tako število shranjeno v 32 bitih. Če je normalizirano, je podano v obliki

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

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

Naloga 1.6. Dokažite, da je

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

in določite $\mathrm{fl}(0.1)$ 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, $\mathrm{fl}(0.1)$ dobimo z zaokroževanjem. Na podlagi

$$0.1 = 0.11001100110011001100110011001\ldots_2 \cdot 2^{-3}$$

sklepamo, da je

$$\mathrm{fl}(0.1) = 0.11001100110011001100110011101_2 \cdot 2^{-3}$$

oziroma

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

Število $\mathrm{fl}(0.1)$ torej 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 $\mathrm{fl}(0.1)$ za enojno natančnost.

```
x = 0.1;
flx = [repmat([1 1 0 0],1,5) [1 1 0 1]] * 2.^-(4:27)';
double(single(x))-x      % 1.4901161e-09
double(single(x))-flx    % 0
```

Naloga 1.7. V Matlabu implementirajte funkcijo, ki dano, v dvojni natančnosti predstavljivo število $x \in \mathbb{R}$ zaokroži na najbližje predstavljivo število iz množice $P(2, t, -1021, 1024)$, kjer je dolžina mantiste $t \in \{1, 2, \ldots, 53\}$ vhodni podatek. Funkcijo uporabite za izračun predstavljivih števil iz nalog 1.5 in 1.6.

Rešitev. Neničelno število x lahko zapišemo v obliki $x = d \cdot 2^e$, kjer je $|d| \in [1/2, 1)$ in e neko celo število. Pri iskanju $\mathrm{fl}(x)$ si lahko zato pomagamo z dvojiškim logaritmom. Ker je $\log_2(|x|) = \log_2(|d|) + e$ in velja $-1 \leq \log_2(|d|) < 0$, po zaokroževanju vrednosti $\log_2(|x|)$ na najbližje celo število navzdol dobimo $e - 1$. V Matlabu lahko torej števili e in d določimo z ukazoma `e = floor(log2(abs(x))) + 1` in `d = x/2^e`. Hitreje in zanesljiveje do teh dveh vrednosti pridemo z ukazom `[d,e] = log2(x)`.

Za izračun $\mathrm{fl}(x)$ je treba vrednost d, ki si jo predstavljamo kot

$$d = \pm\left(d_1 \cdot 2^{-1} + d_2 \cdot 2^{-2} + \ldots\right), \quad d_i \in \{0, 1\}, \; i = 1, 2, \ldots,$$

ustrezno zaokrožiti. Ker je

$$d \cdot 2^t = d_1 \cdot 2^{t-1} + d_2 \cdot 2^{t-2} + \ldots + d_t \cdot 2^0 + d_{t+1} \cdot 2^{-1} + \ldots,$$

po zaokroževanju števila $d \cdot 2^t$ in deljenju z 2^t dobimo število m, ki je najbližje d med vsemi števili s t decimalkami v dvojiškem zapisu. Dodatno pazljivi moramo biti v primeru, ko d leži na sredini med dvema predstavljivima številoma. Takrat je $d \cdot 2^t$ oblike $c/2$ za $c \in \mathbb{Z}$ in po standardu IEEE ga zaokrožimo na najbližje sodo število. Za tako določeno vrednost m velja $\mathrm{fl}(x) = m \cdot 2^e$. V Matlabu lahko m izračunamo z ukazom `m = round(d*2^t,'TieBreaker','even') / 2^t`, število $\mathrm{fl}(x)$ pa nato z ukazom `flx = m*2^e`. Tudi tu se lahko poslužimo hitrejšega in zanesljivejšega načina, in sicer z ukazom `flx = pow2(round(pow2(d,t),'TieBreaker','even'),e-t)`.

```
function flx = fl(x,t)
% Vrne najbližje predstavljivo število flx za x v
% dvojiški bazi in mantiso dolžine t.

[d,e] = log2(x);
flx = pow2(round(pow2(d,t),'TieBreaker','even'),e-t);

end
```

Z uporabo funkcije `fl` bi radi izračunali najbližje predstavljivo število za 47.712 pri mantisi dolžine 9 in za 0.1 pri mantisi dolžine 24. Rezultati spodnjih ukazov potrdijo, da s funkcijo `fl` dobimo enaki števili, kot sta izpeljani v nalogah 1.5 in 1.6.

```
flx = [1 0 1 1 1 1 1 1 0] * 2.^(5:-1:-3)';
flx - fl(47.712,9)    % 0

flx = double(single(0.1));
flx - fl(0.1,24)      % 0
```

Naloga 1.8. Naj bo 1_- največje število v dvojni natančnosti, ki je manjše od 1, in 1_+ najmanjše število v dvojni natančnosti, ki je večje od 1. Katero število v dvojni natančnosti je najbližje številu $1_- \cdot 1_+$?

Rešitev. Števila v dvojni natančnosti so predstavljena v dvojiški bazi z mantiso dolžine 53. Največje število, ki je manjše od 1, je

$$1_- = 0.\underbrace{11\ldots1}_{53}{}_2 \cdot 2^0 = 1 - 2^{-53},$$

najmanjše število, ki je večje od 1, pa

$$1_+ = 0.1\underbrace{00\ldots0}_{51}1_2 \cdot 2^1 = 1 + 2^{-52}.$$

Produkt teh dveh števil je

$$1_- \cdot 1_+ = (1 + 2^{-52}) - (2^{-53} + 2^{-105}) = 1 + (2^{-53} - 2^{-105}),$$

iz česar je razvidno, da je $1_- \cdot 1_+$ število med 1 in 1_+ ter da je bližje 1 kot 1_+.

Naloga 1.9. Dokažite, da za zaporedni pozitivni normalizirani števili x in y iz množice predstavljivih števil $P(b, t, L, U)$ velja $b^{-t}x \leq |x - y| \leq b^{1-t}x$.

Rešitev. Pozitivno število x iz množice $P(b, t, L, U)$ je oblike

$$x = \left(c_1 b^{-1} + c_2 b^{-2} + \ldots + c_t b^{-t}\right) \cdot b^e$$

za $c_i \in \{0, 1, \ldots, b-1\}$, $i = 1, 2, \ldots, t$, in $e \in \{L, L+1, \ldots, U\}$. Predpostavljamo, da je x normalizirano število, zato velja $c_1 \neq 0$. Zapišemo ga kot $x = d \cdot b^{e-t}$, kjer za

$$d = c_1 b^{t-1} + c_2 b^{t-2} + \ldots + c_t$$

velja $b^{t-1} \leq d < b^t$.

Če x ni največje normalizirano število, je najmanjše normalizirano število y, ki je večje od x, enako $y = x + b^{e-t}$. Torej je

$$|x - y| = b^{e-t} = \frac{|x|}{d}$$

in oceni za d potrjujeta dokazovano spodnjo in zgornjo mejo za $|x - y|$.

Obravnavajmo še največje pozitivno normalizirano število y, ki je manjše od x. Če je $c_1 > 1$ ali če obstaja $j \in \{2, 3, \ldots, t\}$, da je $c_j \neq 0$, je $y = x - b^{e-t}$, zato tako kot prej velja $|x - y| = b^{e-t}$. Sicer je $x = b^{t-1} \cdot b^{e-t} = b^{e-1}$ in če x ni najmanjše pozitivno normalizirano število ($e > L$), velja

$$y = \big((b-1)b^{t-1} + (b-1)b^{t-2} + \ldots + (b-1)\big) \cdot b^{e-t-1} = x - b^{e-t-1}.$$

Torej je $b^{-t}x = b^{-t} \cdot b^{e-1} = b^{e-t-1} = |x - y|$, kar potrjuje spodnjo in zgornjo mejo za $|x - y|$.

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 1.10. Funkcija f je podana s predpisom $f(x) = \sqrt{1+x}$. Z računanjem v množici $P(10, 5, -10, 10)$ določite vrednost za $f(x)$ pri argumentu $x = 1/13$.

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 vrednosti polinoma g v točki $\overline{x}$ s Hornerjevim postopkom poteka na sledeč način:
$$b_2 = a_2, \qquad b_1 = b_2\overline{x} + a_1, \qquad b_0 = b_1\overline{x} + a_0.$$
Izhodni podatek postopka je b_0, ki ustreza vrednosti $g(\overline{x})$. V predpisani aritmetiki izvedbo postopka podaja naslednja tabela.

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	$0.10377 \cdot 10^1$

Rezultat, ki predstavlja približek za $f(x)$, označimo z y. Z izvedbo Hornerjevega postopka v dvojni natančnosti, rezultat katerega privzamemo za točno vrednost $g(\overline{x})$, dobimo približno 1.0377219, torej je zaokrožitvena napaka D_z po absolutni vrednosti manjša od $0.22 \cdot 10^{-4}$.

Ker je
$$|f(x) - y| = |f(x) - f(\overline{x}) + f(\overline{x}) - g(\overline{x}) + g(\overline{x}) - y| \leq |D_n| + |D_m| + |D_z|\,,$$
iz obravnave posameznih napak sledi, da je celotna napaka manjša od 10^{-4}.

Naloga 1.11. Dani sta diferenčni enačbi

$$a_n = \tfrac{5}{2}a_{n-1} - a_{n-2}, \qquad n = 2, 3, \ldots, \qquad a_0 = 1,\ a_1 = \tfrac{1}{2},$$

$$b_n = \tfrac{10}{3}b_{n-1} - b_{n-2}, \qquad n = 2, 3, \ldots, \qquad b_0 = 1,\ b_1 = \tfrac{1}{3}.$$

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 $\boldsymbol{a} = (a_0, a_1, \ldots, a_{50})$ in $\boldsymbol{b} = (b_0, b_1, \ldots, b_{50})$ ter z ukazom `scatter` narišite točke (n, a_n) in (n, b_n), $n = 0, 1, \ldots, 50$. Ali se elementi seznamov ujemajo s točnimi vrednostmi? Pojasnite, zakaj da oziroma ne.

3. Omilite napake, ki nastanejo pri izračunu elementov v seznamu $\boldsymbol{b}$ 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 je $b_0 = 1$. 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

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

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 padata proti 0.

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

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

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

Iz grafa na sliki 1.1a, ki ga narišemo z ukazom `scatter(0:50, a)`, je razvidno, da vrednosti a_n padajo proti 0, ko n raste, kar je glede na točno rešitev diferenčne enačbe pričakovano pričakovano. Graf je narisan v logaritemski skali, kar dosežemo z ukazom `set(gca, 'YScale', 'log')`. Graf na sliki 1.1b, narisan z

ukazom `scatter(0:50, b)`, po drugi strani kaže, da izračunane vrednosti b_n z naraščanjem n najprej padajo, nato pa začnejo rasti in so glede na točno rešitev diferenčne enačbe povsem napačne.

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 predstavljivo število v dvojni natančnosti, medtem ko v drugem primeru operiramo z nepredstavljivimi števili. Za začetni podatek namesto $b_1 = 1/3$ uporabimo $\tilde{b}_1 = \mathrm{fl}(b_1) = b_1(1+\delta)$. Pri tem je δ sicer po absolutni vrednosti majhno število, a točna rešitev $\tilde{b}_n$ diferenčne enačbe z začetnima podatkoma b_0 in $\tilde{b}_1$ je

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

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

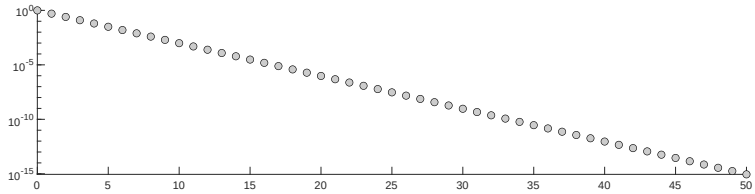
$$b_n = \tfrac{10}{3}b_{n+1} - b_{n+2}.$$

Vzamemo $b_{50} = 0$ in $b_{49} = 1$ ter elemente seznama $\boldsymbol{b}$ 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 graf na sliki 1.1c.

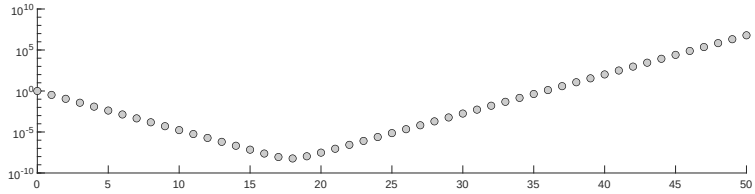
```
% seznam b, generiran v obrantem vrstnem redu
rb = [zeros(1,49) 1 0];
for n = 49:-1:1
    rb(n) = 10/3*rb(n+1) - rb(n+2);
end
b = rb/rb(1);
```

1.3. Stabilnost izračunov

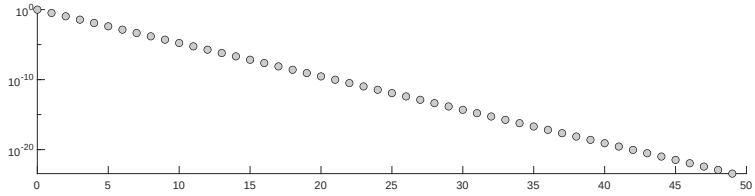
Pri numeričnem računanju stabilnost obravnavamo v različnih kontekstih. V osnovi nas zanima razlika med točno vrednostjo in izračunanim približkom: to je *direktna napaka*. Če je ta pri vseh začetnih podatkih majhna, pravimo, da je metoda izračuna direktno stabilna. Ocenjevanje direktne napake je ponavadi težavno, zato si pri analizi pomagamo z *obratno napako*: to je razlika med točnimi vhodnimi podatki in vhodnimi podatki, ki so prirejeni tako, da se izračunani približek ujema s točno vrednostjo na prirejenih podatkih. 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.



(A) *Elementi seznama* ***a****.*



(B) *Elementi seznama* ***b****.*



(C) *Elementi seznama* ***b*** *(generiran v obratnem vrstnem redu)*

SLIKA 1.1: *Prikaz rezultatov pri rekurzivnem računanju števil v nalogi 1.11.*

Naloga 1.12. Računski stroj uporablja dvojiško aritmetiko z mantiso sode dolžine $t \geq 6$. Naj bo $x = 2^{-1}+2^{-k}+2^{-t}$ in $y = 2^{-1}+2^{-k}$, pri čemer je $k = t/2+1$. Vrednost $x^2 - y^2$ izračunamo z izrazom `x*x - y*y`. Z obravnavo relativne napake dokažite, da izračun ni direktno stabilen. Ugotovitve podprite s preizkusom v Matlabu na podlagi računanja v enojni natačnosti.

Rešitev. Za števili x in y velja

$$x^2 = 2^{-2} + 2^{-k} + 2^{-t} + 2^{-2k} + 2^{1-k-t} + 2^{-2t}, \quad y^2 = 2^{-2} + 2^{-k} + 2^{-2k}.$$

Ker je $2k = t + 2$ in $k \geq 4$, ob zaokroževanju dobimo

$$\mathrm{fl}(x^2) = 2^{-2} + 2^{-k} + 2^{-t} + 2^{-t-1}, \quad \mathrm{fl}(y^2) = 2^{-2} + 2^{-k}.$$

Pri tem upoštevamo, da je y^2 ravno na sredini med najbližjima predstavljivima številoma $2^{-2} + 2^{-k}$ in $2^{-2} + 2^{-k} + 2^{-t-1}$, in za $\mathrm{fl}(y^2)$ vzamemo prvega, ker ima sodo zadnjo števko. Vrednost $x^2 - y^2$, izračunana z izrazom `x*x - y*y`, je potem

$$z = \mathrm{fl}(\mathrm{fl}(x^2) - \mathrm{fl}(y^2)) = 2^{-t} + 2^{-t-1}$$

in za relativno napako izračuna velja

$$\frac{\left|z-(x^2-y^2)\right|}{|x^2-y^2|} = \frac{2^{-t-1}-2^{1-k-t}-2^{-2t}}{2^{-t}+2^{1-k-t}+2^{-2t}} > \frac{2^{-t-2}}{2^{-t+1}} = \frac{1}{8}.$$

Torej je relativna napaka bistveno večja od osnovne zaokrožitvene napake, ki je enaka 2^{-t}, zato izračun ni direktno stabilen.

Za preizkus v Matlabu izberemo $t = 24$, saj to ustreza dolžini mantise pri enojni natančnosti. Izračunana napaka je večja od ocene za relativno napako.

```
t = 24; k = 13;

dx = 2^-1 + 2^-k + 2^-t;
dy = 2^-1 + 2^-k;
dz = dx*dx - dy*dy;

x = single(dx);
y = single(dy);
z = x*x - y*y;

abs(double(z)-dz) / abs(dz)     % 0.4996
```

Na stroju, ki deluje v skladu s standardom IEEE, je vsaka osnovna računska operacija $\oplus$ (seštevanje, odštevanje, množenje, deljenje) implementirana tako, da za vsaki predstavljivi števili x in y ob predpostavki, da ne pride do prekoračitve, velja

$$\mathrm{fl}(x \oplus y) = (x \oplus y)(1+\delta),$$

kjer je δ število z absolutno vrednostjo $|\delta|$, ki ne presega velikosti osnovne zaokrožitvene napake uporabljene aritmetike.

Naloga 1.13. Naj bosta x in y predstavljivi realni števili. Na stroju, ki ustreza standardu IEEE, vrednost $x^2 - y^2$ izračunamo z izrazom `(x-y) * (x+y)`, pri čemer ne pride do prekoračitve. Dokažite, da je izračun obratno in direktno stabilen. V Matlabu na primeru iz naloge 1.12 preverite, da je pri takem izračunu v enojni natančnosti relativna napaka res pričakovano majhna.

Rešitev. Ker računanje poteka po standardu IEEE, namesto vsote oziroma razlike števil x in y izračunamo $(x+y)(1+\alpha_1)$ in $(x-y)(1+\alpha_2)$, kjer sta α_1 in α_2 števili z absolutnima vrednostma $|\alpha_1|$ in $|\alpha_2|$, ki ne presegata osnovne zaokrožitvene napake u. Z množenjem teh dveh vrednosti dobimo

$$z = (x+y)(x-y)(1+\alpha_1)(1+\alpha_2)(1+\beta),$$

kjer je β število z lastnostjo $|\beta| \leq u$. Torej je

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

kjer je δ število, za katero velja

$$(1-u)^3 \leq 1+\delta \leq (1+u)^3.$$

Ker je vrednost u majhna, na podlagi tega ocenimo $|\delta| \lesssim 3u$.

Iz zgornjih ugotovitev sledi, da je $z = (x\sqrt{1+\delta})^2 - (y\sqrt{1+\delta})^2$, torej je z točna rešitev pri malo zmotenih podatkih in izračun je obratno stabilen. Poleg tega za relativno napako velja

$$\frac{\left|z-(x^2-y^2)\right|}{|x^2-y^2|} \leq \frac{\left|(x^2-y^2)\delta\right|}{|x^2-y^2|} = |\delta| \lesssim 3u,$$

kar dokazuje direktno stabilnost izračuna. To potrdi tudi izračun v enojni natančnosti na primeru iz naloge 1.12. Izračunana relativna napaka je celo manjša od osnovne zaokrožitvene napake.

```
t = 24; k = 13;

dx = 2^-1 + 2^-k + 2^-t;
dy = 2^-1 + 2^-k;
dz = dx*dx - dy*dy;

x = single(dx);
y = single(dy);
z = (x-y) * (x+y);

abs(double(z)-dz) / abs(dz)     % 5.9590e-08
```

Naloga 1.14. Dan je polinom

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

kjer so koeficienti $a_n, a_{n-1}, \ldots, a_1, a_0$ in ničle $x_n, x_{n-1}, \ldots, x_1$ predstavljiva 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 pri računanju po standardu IEEE kateri izmed postopkov direktno stabilen? Demonstrirajte postopka v Matlabu z izračuni v enojni natančnosti pri podatkih $n = 9$, $a_9 = 1$, $x_1 = x_2 = \ldots = x_9 = 2$ in $x = 2 + 2^{-5}$.

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

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

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

$$y_1 = a_n x^n(1+\delta_n) + a_{n-1}x^{n-1}(1+\delta_{n-1}) + \ldots + a_1 x(1+\delta_1) + a_0(1+\delta_0),$$

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

$$\frac{|y - y_1|}{|y|} \lesssim \frac{2nu\left(|a_n|\,|x^n| + |a_{n-1}|\,\left|x^{n-1}\right| + \ldots + |a_1|\,|x| + |a_0|\right)}{|a_n x^n + a_{n-1}x^{n-1} + \ldots + a_1 x + a_0|}.$$

Čeprav smo napako (približno) 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.

V primeru, da obstaja faktorizacija polinoma na linearne faktorje, lahko izračun vrednosti opravimo stabilneje z množenjem faktorjev. Najprej izračunamo $x - x_i$, $i = 1, 2, \ldots, n$, 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

$$y_2 = a_n(x - x_n)\ldots(x - x_1)(1 + \delta) = y(1 + \delta),$$

pri čemer je $|\delta| \lesssim 2nu$. To pomeni, da za relativno napako velja

$$\frac{|y - y_2|}{|y|} = \frac{|y\delta|}{|y|} = |\delta| \lesssim 2nu$$

in izračun je direktno stabilen.

Zgornje ugotovitve potrjuje tudi konkreten primer izvedbe izračunov v Matlabu v enojni natančnosti. Točna vrednost pri podanih podatkih je $y = (2^{-5})^9 = 2^{-45}$. Pri računanju s Hornerjevim postopkom (funkcija `polyval`) dobimo zelo veliko relativno napako, medtem ko je napaka pri izračunu na podlagi množenja faktorjev v tem primeru celo enaka 0, saj so rezultati vseh operacij predstavljiva števila v enojni natančnosti.

```
X = single(2*ones(1,9));
A = poly(X);

x = single(2 + 2^-5);
y = 2^-45;

y1 = double(polyval(A,x));
abs(y-y1) / abs(y)     % 1.0737e+09

y2 = double(prod(x-X));
abs(y-y2) / abs(y)     % 0
```

2. Nelinearne enačbe

Če nelinearno enačbo z neznanko x predstavimo v obliki $f(x) = 0$, lahko njeno reševanje interpretiramo kot iskanje ničle funkcije f. Numerične metode za reševanje tega problema so iterativne, kar pomeni, da približek za rešitev iščemo v več korakih s ponavljanjem izračunov, ki v trenutnem koraku izboljšajo približek iz prejšnjega koraka.

2.1. Bisekcija

Bisekcija je robustna metoda za iskanje ničle zvezne funkcije f na intervalu $[a, b]$, za katero predpostavljamo, da je v robovih intervala nasprotno predznačena (velja torej $f(a)f(b) < 0$). Temelji na dejstvu, da ima taka funkcija vsaj eno ničlo na intervalu $[a, b]$. Kot namiguje ime, v vsakem koraku metode razpolovimo trenutni interval ter naslednji korak nadaljujemo na desnem ali levem podintervalu, v odvisnosti od tega, na katerem izmed njiju je f nasprotno predznačena.

Naloga 2.1. Implementirajte bisekcijo v Matlabu. Pozorni bodite na ekonomičnost in numerično stabilnost izvedbe postopka. Izvedite ga na intervalu $[1, 2]$ za funkcijo f, podano s predpisom $f(x) = x + 4 - e^{x^2}$. Napravite toliko korakov metode, da je napaka izračunanega približka manjša od 10^{-12}.

Rešitev. Pripravimo funkcijo `bisekcija`, ki kot vhodne podatke sprejme funkcijo `f`, števili `a` in `b`, ki določata robova intervala, ter število korakov bisekcije `N`. Pri implementaciji metode pazimo, da funkcijsko vrednost v določeni točki izračunamo le enkrat. S spremenljivko `s`, ki predstavlja dolžino trenutnega intervala, se izognemo morebitnim numeričnim težavam, ki bi se lahko pojavile pri izračunu središča intervala z izrazom `(a+b)/2`.

```
function x = bisekcija(f,a,b,N)
% funkcija
%  x = bisekcija(f,a,b,N)
% izvede bisekcijo in določi približek x za ničlo
% funkcije f na intervalu [a,b]
%
```

```
% vhodni podatki:
%  f    funkcija,
%  a    začetek intervala,
%  b    konec intervala,
%  N    število korakov bisekcije
%
% izhodni podatek:
%  x    približek za ničlo funkcije f

fa = f(a);
if sign(fa) == sign(f(b))
   error('f v točkah a in b ni nasprotno predznačena');
end

k = 0;
s = b-a;
while k < N
   s = s/2;
   x = a+s;
   fx = f(x);
   if sign(fa) == sign(fx)
      a = x;
      fa = fx;
   end
   k = k+1;
end

end
```

Za izračun približka za ničlo podane funkcije na intervalu $[1, 2]$, ki se od točne vrednosti razlikuje za manj kot 10^{-12}, zadošča $\lceil \log_2(10^{12}) \rceil = 40$ korakov bisekcije. S pomočjo vgrajene funkcije `fzero` se prepričamo, da v tem primeru dovolj natančen približek dobimo že z 39 koraki.

```
f = @(x) x+4-exp(x^2);
bisekcija(f,1,2,39)  % 1.290718421716520
bisekcija(f,1,2,40)  % 1.290718421715610
fzero(f,[1 2])       % 1.290718421715963
```

Naloga 2.2. Analizirajte, koliko ničel na intervalu $[0, 1]$ ima funkcija f, ki je podana s predpisom

$$f(x) = \left(x - \frac{1}{2}\right)^2 \left(x - \frac{3}{4}\right) - \frac{1}{2^n}$$

za izbrano število $n \in \mathbb{N}$. Kakšen je izid bisekcije (glede na implementacijo iz naloge 2.1), če računamo v enojni natančnosti?

Rešitev. Funkcijo f lahko analiziramo na podlagi odvoda

$$f'(x) = 3\left(x - \frac{1}{2}\right)\left(x - \frac{2}{3}\right).$$

Razvidno je, da je funkcija f na intervalu $[0, 1/2]$ naraščajoča, na intervalu $[0, 2/3]$ padajoča, na intervalu $[2/3, 1]$ pa ponovno naraščajoča. Ker je

$$f(0) = -\frac{3}{16} - \frac{1}{2^n}, \quad f(\tfrac{1}{2}) = -\frac{1}{2^n}, \quad f(1) = \frac{1}{16} - \frac{1}{2^n}$$

sklepamo, da je f na intervalu $[0, 2/3]$ negativna in zato tam nima ničle, na intervalu $[2/3, 1]$ pa ima eno ničlo, če je $n \geq 4$ (pri $n = 4$ je ničla ravno 1).

Pri izvedbi bisekcije za funkcijo f v enojni natančnosti vse poteka po pričakovanjih, razen če pride do podkoračitve in je $\mathrm{fl}(2^{-n}) = 0$. Najmanjše pozitivno predstavljivo število v enojni natačnosti je $2^{-24} \cdot 2^{-125}$, torej se to zgodi, če je $n > 149$. Pri izvedbi bisekcije v tem primeru dobimo $\mathrm{fl}(f(1/2)) = 0$ in (glede na implementacijo v nalogi 2.1, saj je `sign(0) = 0`) iskanje ničle nadaljujemo na intervalu $[0, 1/2]$ ter nato na intervalih $[1/4, 1/2], [3/8, 1/2], [7/16, 1/2], \ldots$. Končamo v bližini $1/2$, ki pa ni dejanska ničla funkcije f.

```
a = single(0);
b = single(1);

f149 = @(x) (x-0.5)^2*(x-0.75)-2^-149;
bisekcija(f149,a,b,50)    % 0.7500
fzero(f149,[0 1])         % 0.7500

f150 = @(x) (x-0.5)^2*(x-0.75)-2^-150;
bisekcija(f150,a,b,50)    % 0.5000
fzero(f150,[0 1])         % 0.7500
```

2.2. Navadna iteracija

Eden izmed splošnih pristopov za iskanje ničle funkcije f oziroma reševanje nelinearne enačbe $f(x) = 0$ je prepis enačbe v ekvivalentno obliko $x = g(x)$. Funkcijo g imenujemo iteracijska funkcija, saj rešitev enačbe iščemo z iteracijo

$$x_{r+1} = g(x_r), \quad r = 0, 1, \ldots.$$

Če je g na določenem intervalu, ki vsebuje x_0, skrčitev, zaporedje $(x_r)_r$ po Banachovem skrčitvenem načelu konvergira k negibni točki g, ki ustreza ničli funkcije f.

Naloga 2.3. Funkcija f je podana s predpisom $f(x) = x^5 - 10x + 1$. Njeno ničlo iščemo z iteracijsko funkcijo $g(x) = (x^5 + 1)/10$.

1. Utemeljite, da ima funkcija f na intervalu $[0, 0.2]$ natanko eno ničlo.
2. Dokažite, da začetni približek $x_0 = 0$ zagotavlja konvergenco iteracijskega zaporedja $x_{r+1} = g(x_r)$ k ničli funkcije f na intervalu $[0, 0.2]$.
3. Ocenite, kako dobro se približek $x_2 = g(g(0))$ ujema z ničlo funkcije f na intervalu $[0, 0.2]$.

Rešitev.

1. Funkcija f je za $x = 0$ enaka 1, za $x = 0.2$ pa $0.2^5 - 1 < 0$. Ker je zvezna, ima na intervalu $[0, 0.2]$ vsaj eno ničlo. Poleg tega je funkcija f na intervalu $[0, 0.2]$ padajoča, saj je $f'(x) < 0$ za vsak x s tega intervala. To potrjuje, da je ničla ena sama.
2. Odvod iteracijske funkcije g je enak $x^4/2$, zato je po absolutni vrednosti manjši od 1 natanko tedaj, ko je $|x| < \sqrt[4]{2}$. To zagotavlja konvergenco za vsak začetni približek z intervala $(-\sqrt[4]{2}, \sqrt[4]{2}) \approx (-1.1892, 1.1892)$.
3. Naj bo $\alpha \in (0, 0.2)$ ničla funkcije f ($f(\alpha) = 0$) oziroma negibna točka funkcije g ($g(\alpha) = \alpha$). Opazimo, da za vsak $x \in [0, \alpha]$ velja
$$g(x) - \alpha \leq \frac{\alpha^5 + 1}{10} - \alpha = \frac{f(\alpha)}{10} = 0,$$
kar pomeni, da se vsi členi iteracijskega zaporedja nahajajo na intervalu $[0, \alpha]$. Zato lahko po izreku o povprečni vrednosti razliko med zaporednima približkoma x_r in x_{r+1}, $r \in \mathbb{N}$, ocenimo z
$$|x_{r+1} - x_r| = |g(x_r) - g(x_{r-1})| < g'(0.2)\,|x_r - x_{r-1}| = 8 \cdot 10^{-4}\,|x_r - x_{r-1}|\,.$$
Ker začnemo iteracijo z $x_0 = 0$, v naslednjih dveh korakih dobimo $x_1 = 1/10$ in $x_2 = 1/10 + 1/10^6$. Ocenimo
$$|x_2 - \alpha| \leq |x_2 - x_3| + |x_3 - x_4| + \ldots < \left(8 \cdot 10^{-4} + \left(8 \cdot 10^{-4}\right)^2 + \ldots\right)|x_1 - x_2|$$
in od tod sklepamo
$$|x_2 - \alpha| < \frac{8 \cdot 10^{-4}}{1 - 8 \cdot 10^{-4}} 10^{-6} \approx 8 \cdot 10^{-10}.$$
Torej že drugi približek iteracije predstavlja dober približek za ničlo funkcije f.

Naloga 2.4. Iteracijska funkcija g je podana s predpisom $g(x) = -x^2 + 8x - 12$.

1. Določite negibne točke iteracije $g(x) = x$ in opredelite, katere so privlačne in katere odbojne.

2. Za katere začetne približke v okolici negibnih točk lahko na podlagi odvoda funkcije g zagotovimo konvergenco navadne iteracije?

3. Določite, za katere začetne približke je navadna iteracija konvergentna. Kam konvergirajo zaporedja?

Rešitev.

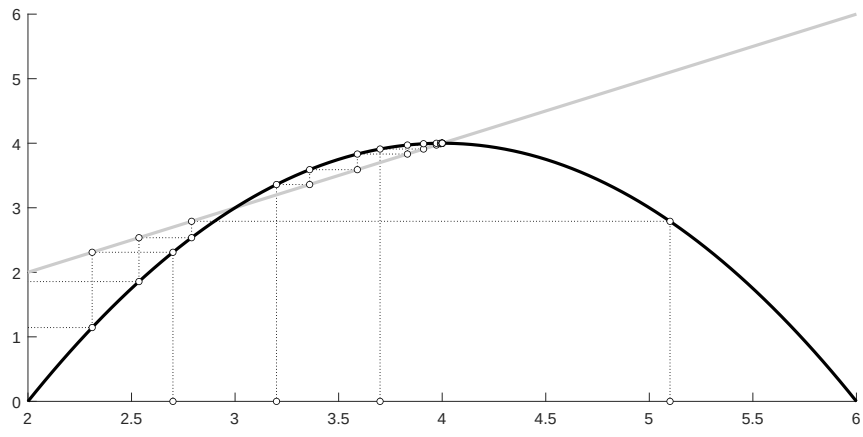
1. Negibne točke dobimo z reševanjem kvadratne enačbe $g(x) = x$. Negibni točki sta torej dve, prva je 3, druga pa 4. Z odvajanjem iteracijske funkcije dobimo, da je $g'(3) = 2$ in $g'(4) = 0$, kar pomeni, da je 3 odbojna, 4 pa privlačna točka.

2. Konvergenco navadne iteracije v okolici negibne točke $x = 4$ lahko na podlagi odvoda funkcije g zagotovimo za začetne približke x_0, za katere velja $|g'(x_0)| < 1$. Ta pogoj je izpolnjen natanko tedaj, ko je $|-2x_0 + 8| < 1$ oziroma $x_0 \in (3.5, 4.5)$.

3. Na podlagi grafa iteracijske funkcije g in simetrale lihih kvadrantov (slika 2.1 prikazuje navadno iteracijo pri začetnih približkih 2.7, 3.2, 3.7 in 5.1) domnevamo, da iteracija konvergira k 4 pri začetnih približkih z intervala $(3, 5)$, pri vseh drugih začetnih približki pa divergira ali obstane v 3. Opazimo, da za približek x_r na r-tem koraku iteracije velja

$$x_r - 4 = g(x_{r-1}) - 4 = -(x_{r-1} - 4)^2 = \ldots = -(x_0 - 4)^{2^r},$$

kar potrjuje, da za vsak $x_0 \in (3, 5)$ velja $\lim_{r\to\infty} x_r = 4$. Od tod sledi tudi, da se za vsak $x_0 \in (-\infty, 3) \cup (5, \infty)$ približki x_r zmanjšujejo brez meje, pri začetnih približkih $x_0 = 3$ in $x_0 = 5$ pa velja $x_r = 3$ za vsak $r \in \mathbb{N}$.



Slika 2.1: *Grafični prikaz navadne iteracije iz naloge 2.4.*

Naloga 2.5. V Matlabu sestavite funkcijo, ki izvede navadno iteracijo.

1. Vhodni podatki naj bodo iteracijska funkcija, začetni približek, toleranca in maksimalno število korakov.
2. Izhodni podatki naj bodo izračunan približek po končani iteraciji, seznam vseh približkov tekom iteracije in število opravljenih korakov.
3. Funkcija naj izvaja iteracijo, dokler se zadnja dva približka absolutno ne razlikujeta za manj, kot je predpisana toleranca, ali število korakov ne preseže danega maksimalnega števila korakov.

Za test implementacije uporabite iteracijsko funkcijo iz naloge 2.4.

Rešitev. Funkcijo, ki izvede navadno iteracijo, poimenujemo `iteracija`.

```
function [x,X,k] = iteracija(g,x0,tol,N)
% funkcija
%  [x,X,k] = iteracija(g,x0,tol,N)
% izvede navadno iteracijo z dano iteracijsko funkcijo
% in začetnim približkom
%
% vhodni podatki:
%  g     iteracijska funkcija,
%  x0    začetni približek,
%  tol   toleranca absolutnega ujemanja dveh zaporednih
%        približkov,
%  N     maksimalno število korakov iteracije
%
% izhodni podatki:
%  x     zadnji približek izračunan z navadno iteracijo,
%  X     seznam vseh izračunanih približkov,
%  k     število opravljenih korakov iteracije

X = x0;
k = 0;
while k < N
    k = k+1;
    X(k+1) = g(X(k));
    if abs(X(k+1)-X(k)) < tol
        break;
    end
end
x = X(k+1);

end
```

Test kaže, da se računski rezultati ujemajo s teoretičnimi izpeljavami iz naloge 2.4.

```
g = @(x)-x^2+8*x-12; tol = 1e-10; N = 1e3;
[x,~,k] = iteracija(g,3.5,tol,N)     % x = 4, k = 7
[x,~,k] = iteracija(g,4.5,tol,N)     % x = 4, k = 7
[x,~,k] = iteracija(g,3,tol,N)       % x = 3, k = 1
[x,~,k] = iteracija(g,2,tol,N)       % x = -Inf, k = 1000
```

V praksi je pomembno, kako hitro iteracijsko zaporedje $(x_r)_r$ konvergira k negibni točki α. Pravimo, da je red konvergence enak p, če obstaja taka konstanta $C > 0$, da je

$$\lim_{r\to\infty} \frac{|x_{r+1} - \alpha|}{|x_r - \alpha|^p} = C.$$

Če je funkcija g dovoljkrat zvezno odvedljiva, lahko red p enostavno določimo z odvajanjem: veljati mora $g^{(k)}(\alpha) = 0$, $k = 1, 2, \ldots, p-1$, in $g^{(p)}(\alpha) \neq 0$ ob dodatni predpostavki, da je $|g'(\alpha)| < 1$, če je $p = 1$.

Naloga 2.6. Za iskanje ničle funkcije $f(x) = x^2 - x - 2$ uporabite štiri različne iteracijske funkcije:

1. $g_1(x) = x^2 - 2$,
2. $g_2(x) = \sqrt{x+2}$,
3. $g_3(x) = 1 + 2/x$,
4. $g_4(x) = (x^2 + 2)/(2x - 1)$.

Za vsako funkcijo analizirajte konvergenco v okolici ničel -1 in 2 ter določite njen red. Ugotovitve preverite v Matlabu s pomočjo funkcije, implementirane v nalogi 2.5. Narišite grafe števila korakov iteracije v odvisnosti od začetnih približkov na intervalu $[-2, 4]$.

Rešitev. Premislimo, kako se iteracijske funkcije obnašajo v okolici ničel funkcije f.

1. Ker je $g_1'(x) = 2x$, sta tako -1 kot 2 odbojni negibni točki iteracijske funkcije g_1. Ničlo funkcije f torej dobimo le v posebnih primerih, ko za začetni približek izberemo -2, -1, 0, 1 ali 2.

2. Iz $g_2'(x) = 1/\sqrt{x+2}$ sledi, da je 2 privlačna negibna točka, v okolici katere je red konvergence enak 1. Enostavno je dokazati, da iteracijsko zaporedje konvergira k 2 za vsak začetni približek, ki je večji ali enak -2. Na drugi strani -1 ni negibna točka funkcije g_2, zato te ničle funkcije f z g_2 ne moremo poiskati.

3. Najprej opazimo, da sta negibni točki funkcije g_3 tako -1 kot 2, vendar iz $g_3'(-1) = -2$ in $g_3'(2) = -1/2$ sklepamo, da je le 2 privlačna negibna točka. Za približke v okolici -1 torej ni pričakovati konvergence k -1, v -1 z iteracijo

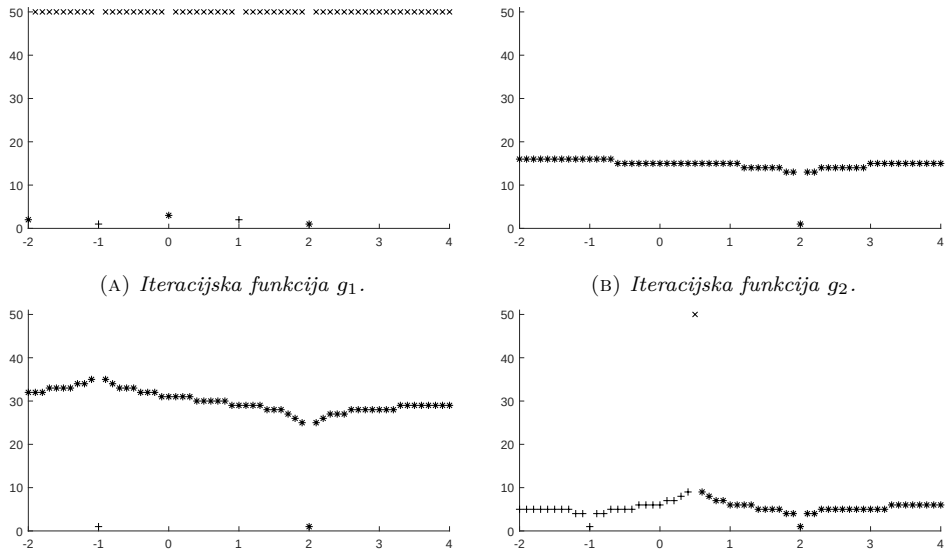
končamo le, če v -1 z njo začnemo. Po drugi strani vrednost odvoda g_3 v točki 2 zagotavlja konvergenco k -2 za začetne približke v okolici -2, a ponovno le reda 1. S podrobnejšo analizo iteracijske funkcije lahko dokažemo, da iteracijsko zaporedje konvergira k -2 za vse začetne približke, razen -1. Pri začetnih približkih 0 in -2 v prvem oziroma drugem koraku iteracije delimo z 0.

4. Funkcija g_4 ima negibni točki -1 in 2, v obeh pa je vrednost odvoda g_4 enaka 0. Red konvergence v okolici -1 in 2 je torej vsaj reda 2 in za začetne približke blizu negibnih točk se lahko nadejamo hitre konvergence k eni ali drugi ničli funkcije f. Iz analize vrednost $g_4(x) - x$ sledi, da iteracijsko zaporedje konvergira k 2 za začetni približek večji od $1/2$ in k -1 za začetni približek manjši od $1/2$.

V Matlabu preverimo, kakšne rezultate dobimo z uporabo iteracijskih funkcij pri začetnih približkih $-1/2$ in 3. Iteracijo izvedemo s pomočjo funkcije `iteracija` iz naloge 2.5 pri toleranci `tol = 1e-10` in maksimalnem številu korakov `N = 100`.

```
g1 = @(x) x.^2-2;
g2 = @(x) sqrt(x+2);
g3 = @(x) 1+2./x;
g4 = @(x) (x.^2+2)./(2*x-1);

tol = 1e-10; N = 100;

[x,~,k] = iteracija(g1,-0.5,tol,N) % x = 0.5914,  k = 100
[x,~,k] = iteracija(g1,3,tol,N)    % x = Inf,     k = 100
[x,~,k] = iteracija(g2,-0.5,tol,N) % x = 2,       k = 19
[x,~,k] = iteracija(g2,3,tol,N)    % x = 2,       k = 18
[x,~,k] = iteracija(g3,-0.5,tol,N) % x = 2,       k = 39
[x,~,k] = iteracija(g3,3,tol,N)    % x = 2,       k = 35
[x,~,k] = iteracija(g4,-0.5,tol,N) % x = -1,      k = 5
[x,~,k] = iteracija(g4,3,tol,N)    % x = 2,       k = 6
```

Na sliki 2.2 so prikazani grafi števila korakov iteracij pri začetnih približkih iz seznama `linspace(-2,4,61)`. Oznaka `*` predstavlja konvergenco k ničli 2, oznaka `+` pa konvergenco k ničli -1. Oznaka `x` pomeni, da se je iteracija končala brez konvergence po maksimalnem številu korakov (v tem primeru 50). Slika 2.2a prikazuje rezultate za iteracijsko funkcijo g_1, kjer v eni izmed ničel končamo le pri začetnih približkih -2, -1, 0, 1 in 2. Na sliki 2.2b je prikazano število korakov iteracij pri funkcji g_2, ki ima za negibno točko ničlo 2. Število korakov, s katerimi dosežemo natančnost približka v okviru predpisane tolerance, je manjše kot pri iteracijski funkciji g_3, kot kaže slika 2.2c. Iz te slike je tudi jasno razvidno, da je 1 odbojna točka iteracije z g_3, saj iteracijsko zaporedje le pri začetnem približku -1 konvergira k -1. Iteracijska zaporedja najhitreje konvergirajo pri iteracijski funkciji g_4. Slika 2.2d kaže, da je limita za začetne približke manjše od $1/2$ enaka -1, za začetne približke večje od $1/2$ pa 2. Pri začetnem približku $1/2$ iteracijsko zaporedje ne konvergira k nobeni izmed ničel.



(A) *Iteracijska funkcija* g_1.

(B) *Iteracijska funkcija* g_2.

(C) *Iteracijska funkcija* g_3.

(D) *Iteracijska funkcija* g_4.

SLIKA 2.2: *Število korakov v odvisnosti od začetnih približkov pri iteracijah iz naloge 2.6.*

Naloga 2.7. Dokažite, da lahko kvadratni koren pozitivnega števila a izračunamo z iteracijo

$$x_{r+1} = x_r \frac{x_r^2 + 3a}{3x_r^2 + a}, \quad r = 0, 1, \ldots,$$

pri poljubnem začetnem približku $x_0 > 0$ ter določite red konvergence iteracijskega zaporedja v bližini $\sqrt{a}$.

Rešitev. Enostavno je preveriti, da je funkcija $g(x) = x(x^2+3a)/(3x^2+a)$ iteracijska funkcija za enačbo $x(x^2 - a) = 0$. Negibne točke iteracije so $-\sqrt{a}$, 0 in $\sqrt{a}$.

Radi bi dokazali, da iteracija za $x_0 > 0$ konvergira k negibni točki $\sqrt{a}$. Najprej opazimo, da za vsak približek x_r, $r \in \mathbb{N}$, velja

$$x_r - \sqrt{a} = g(x_{r-1}) - \sqrt{a} = \frac{(x_{r-1} - \sqrt{a})^3}{3x_{r-1}^2 + a}.$$

Od tod po indukciji sledi, da je za začetni približek $x_0 < \sqrt{a}$ vsak približek x_r manjši od $\sqrt{a}$. Podobno, če je $x_0 > \sqrt{a}$, je tudi vsak približek x_r večji od $\sqrt{a}$. Nadalje, ker je

$$x_{r+1} - x_r = g(x_r) - x_r = \frac{2x_r(a - x_r^2)}{3x_r^2 + a},$$

za $x_r < \sqrt{a}$ velja $x_{r+1} > x_r$, za $x_r > \sqrt{a}$ pa $x_{r+1} < x_r$. To dokazuje, da je zaporedje približkov $(x_r)_r$ pri začetnem približku $x_0 > 0$ konvergentno, saj je bodisi strogo

naraščajoče in navzgor omejeno s $\sqrt{a}$ bodisi strogo padajoče in navzdol omejeno s $\sqrt{a}$. Naj bo limita zaporedja označena z α. Zanjo velja

$$\alpha = \lim_{r\to\infty} x_r = \lim_{r\to\infty} x_{r+1} = \lim_{r\to\infty} g(x_r) = g(\lim_{r\to\infty} x_r) = g(\alpha),$$

zato je negibna točka funkcije g. Ker je $\sqrt{a}$ edina neigbna točka g na intervalu $(0,\infty)$, je $\alpha = \sqrt{a}$.

Red konvergence določimo z odvajanjem. Ker je

$$g'(x) = \frac{3(x^2-a)^2}{(3x^2+a)^2}, \qquad g''(x) = \frac{48xa}{(3x^2+a)^3}(x^2-a),$$

je $g'(\sqrt{a}) = g''(\sqrt{a}) = 0$ in red konvergence je vsaj kubičen. Red v resnici je kubičen, saj je $g'''(\sqrt{a}) \neq 0$, kot sledi iz

$$g'''(x) = \left(\frac{48xa}{(3x^2+a)^3}\right)'(x^2-a) + \frac{48xa}{(3x^2+a)^3}2x.$$

Naloga 2.8. Določite parametre α, β in γ v iterativni formuli

$$x_{r+1} = \alpha x_r + \beta\frac{a}{x_r^2} + \gamma\frac{a^2}{x_r^5}, \quad r = 0,1,\ldots,$$

za računanje kubičnega korena neničelnega števila a tako, da bo konvergenca iteracijskega zaporedja za začetni približek v okolici $\sqrt[3]{a}$ kubična.

Rešitev. Obravnavajmo iteracijsko funkcijo $g(x) = \alpha x + \beta a/x^2 + \gamma a^2/x^5$. Če želimo, da je $\sqrt[3]{a}$ negibna točka g, mora veljati $\alpha+\beta+\gamma = 1$. Da bo red konvergence vsaj kubičen, mora biti $g'(\sqrt[3]{a}) = 0$ in $g''(\sqrt[3]{a}) = 0$, kar je ekvivalentno zahtevama $\alpha - 2\beta - 5\gamma = 0$ in $6\beta + 30\gamma = 0$. Z reševanjem sistema enačb za dane parametre ugotovimo, da mora biti $\alpha = 5/9$, $\beta = 5/9$ in $\gamma = -1/9$. Ker je pri teh parametrih $g'''(\sqrt[3]{a}) = 10/\sqrt[3]{a^2} \neq 0$, je red konvergence kubičen.

2.3. Izpeljava in lastnosti metod

Za iskanje ničel funkcije so na voljo številni recepti, po katerih lahko pripravimo ustrezno iteracijsko funkcijo za izvedbo navadne iteracije. Eden takih je tangentna (ali Newtonova oziroma Newton–Raphsonova) metoda. Približek x_{r+1} za ničlo odvedljive funkcije f dobimo iz približka x_r z zvezo

$$x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)}.$$

Geometrijsko x_{r+1} predstavlja presečišče abscisne osi in tangente funkcije f v točki x_r.

Naloga 2.9. Babilonska metoda za računanje kvadratnega korena $\sqrt{a}$ pozitivnega števila a temelji na iteraciji

$$x_{r+1} = \frac{1}{2}\left(x_r + \frac{a}{x_r}\right), \quad r = 0, 1, \ldots.$$

Za izvedbo enega koraka iteracije so potrebne le tri osnovne računske operacije.

1. Preverite, da iteracija ustreza tangentni metodi za funkcijo $f(x) = x^2 - a$.
2. Kakšen je red konvergence iteracijskega zaporedja v okolici negibne točke $\sqrt{a}$?
3. Dokažite, da zaporedje $(x_r)_r$ konvergira k $\sqrt{a}$ za vsak začetni približek $x_0 > 0$.

Rešitev.

1. V formulo, ki določa približek po tangentni metodi, vstavimo $f(x) = x^2 - a$. Z nekaj preurejanja dobimo babilonsko metodo.
2. Red konvergence babilonske metode v okolici $\sqrt{a}$ lahko določimo z odvajanjem iteracijske funkcije $g(x) = (x + a/x)/2$. Izračunamo
$$g'(x) = \frac{1}{2}\left(1 - \frac{a}{x^2}\right), \qquad g''(x) = \frac{a}{x^3}$$
in iz $g'(\sqrt{a}) = 0$ in $g''(\sqrt{a}) = 1/\sqrt{a} \neq 0$ sklepamo, da je red konvergence v okolici $\sqrt{a}$ kvadratičen.
3. Odvod funkcije g je po absolutni vrednosti manjši od 1 za vsak $x \in (\sqrt{a/3}, \infty)$, zato zaporedje $(x_r)_r$ konvergira k $\sqrt{a}$ za vsak začetni približek x_0 s tega intervala. Treba je dokazati še, da to velja tudi v primeru, ko je $x_0 \in (0, \sqrt{a/3}]$. Opazimo, da je
$$g(x_0) - \sqrt{a} = \frac{x_0}{2} + \frac{a}{2x_0} - \sqrt{a} = \frac{1}{2x_0}\left(x_0 - \sqrt{a}\right)^2,$$
iz česar sledi, da za vsak začetni približek $x_0 > 0$ približek $x_1 = g(x_0)$ leži na intervalu $[\sqrt{a}, \infty)$; tu pa je g po prejšnjem razmisleku skrčitev, zato je limita iteracijskega zaporedja negibna točka $\sqrt{a}$.

Naloga 2.10. Naj bo $f : \mathbb{R} \to \mathbb{R}$ vsaj dvakrat odvedljiva funkcija z ničlo α. Pravimo, da je α ničla kratnosti m, $m \in \mathbb{N}$, če lahko f zapišemo kot $f(x) = (x - \alpha)^m h(x)$ za funkcijo $h : \mathbb{R} \to \mathbb{R}$, ki nima ničle v α.

1. Naj bo α enostavna ničla ($m = 1$). Preverite, da je red konvergence tangentne metode v njeni bližini vsaj kvadratičen.
2. Naj bo $m > 1$. Dokažite, da za iteracijsko funkcijo $g(x) = x - f(x)/f'(x)$ tangentne metode velja
$$\lim_{x \to \alpha} g'(x) = 1 - \frac{1}{m}.$$
Kaj lahko na podlagi tega sklepate o hitrosti konvergence tangentne metode v bližini α?

3. Kako bi prilagodili tangentno metodo, da bi bil red konvergence v okolici ničle kratnosti $m > 1$ vsaj kvadratičen?

Rešitev.

1. Za odvod iteracijske funkcije g velja

$$g'(x) = \left(x - \frac{f(x)}{f'(x)}\right)' = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2},$$

torej je $g'(\alpha) = 0$ in red konvergence v okolici enostavne ničle α je vsaj kvadratičen.

2. Glede na definicijo kratnosti ničle lahko prvi odvod funkcije f izrazimo kot

$$f'(x) = m(x-\alpha)^{m-1}h(x) + (x-\alpha)^m h'(x),$$

drugi odvod pa kot

$$f''(x) = (m-1)m(x-\alpha)^{m-2}h(x) + 2m(x-\alpha)^{m-1}h'(x) + (x-\alpha)^m h''(x).$$

Iz tega sledi, da za odvod iteracijske funkcije g velja

$$g'(x) = \frac{(x-\alpha)^2 h(x)h''(x) + 2m(x-\alpha)h(x)h'(x) + (m-1)mh(x)^2}{(x-\alpha)^2 h'(x)^2 + 2m(x-\alpha)h(x)h'(x) + m^2 h(x)^2},$$

na podlagi česar sklepamo, da je

$$\lim_{x\to\alpha} g'(x) = \frac{(m-1)m}{m^2} = 1 - \frac{1}{m}.$$

To pomeni, da je red konvergence v okolici večkratne ničle linearen in tudi, da je hitrost konvergence tem manjša, čim večja je kratnost ničle.

3. Iteracijsko funkcijo g bi radi zamenjali s sorodno funkcijo G, za katero velja $\lim_{x\to\alpha} G'(x) = 0$. Ker je $m \lim_{x\to\alpha} g'(x) = m - 1$, poskusimo s funkcijo

$$G(x) = x - m\frac{f(x)}{f'(x)}.$$

Njen odvod je

$$G'(x) = 1 - m\frac{f'(x)^2 + f(x)f''(x)}{f'(x)^2} = 1 - m + mg'(x)$$

in je v limiti, ko gre x proti α, enak 0. Seveda lahko metodo z iteracijsko funkcijo h izkoristimo le, če vnaprej poznamo kratnost ničle funkcije f.

Naloga 2.11. Naj bo $f : \mathbb{R} \to \mathbb{R}$ vsaj trikrat odvedljiva funkcija in α ničla f kratnosti več kot 1.

1. Dokažite, da je α enostavna ničla funkcije $F(x) = f(x)/f'(x)$.
2. Izpeljite metodo za iskanje ničle funkcije f, ki ustreza tangentni metodi za funkcijo F. Utemeljite, da je red konvergence izpeljane metode v okolici α vsaj kvadratičen.
3. Naj g označuje iteracijsko funkcijo metode iz druge točke. Dokažite, da je vsaka ničla funkcije f negibna točka funkcije g in da je vsaka negibna točka funkcije g, ki ni ničla f, odbojna.

Rešitev.

1. Naj $m \in \mathbb{N}$ označuje kratnost ničle α. Torej je $f(x) = (x-\alpha)^m h(x)$, kjer je $h : \mathbb{R} \to \mathbb{R}$ funkcija, za katero velja $h(\alpha) \neq 0$. Funkcijo F lahko potem zapišemo v obliki $F(x) = (x-\alpha)H(x)$ za funkcijo $H : \mathbb{R} \to \mathbb{R}$, ki je podana s predpisom
$$H(x) = \frac{h(x)}{mh(x) + (x-\alpha)h'(x)}.$$
Ker je $H(\alpha) = \frac{1}{m} \neq 0$, je α enostavna ničla funkcije F.
2. Iteracijska funkcija je podana z
$$g(x) = x - \frac{F(x)}{F'(x)}$$
in ker je red konvergence tangentne metode v okolici enostavne ničle kvadratičen, je ta lastnost za funkcijo g v okolici α zagotovljena po prvi točki. Z upoštevanjem predpisa za F iteracijsko funkcijo izrazimo glede na f kot
$$g(x) = x - \frac{f(x)f'(x)}{f'(x)^2 - f(x)f''(x)}.$$
3. Iz predpisa za funkcijo g je razvidno, da je $\beta \in \mathbb{R}$ njena negibna točka natanko tedaj, ko je $f(\beta) = 0$ ali $f'(\beta) = 0$. Če je torej β negibna točka g in ni ničla funkcije f, je $f'(\beta) = 0$ in iz predpisa za odvod
$$\begin{aligned} g'(x) = 1 &- \frac{f'(x)^2 + f(x)f''(x)}{f'(x)^2 - f(x)f''(x)} \\ &- f(x)f'(x)\frac{2f'(x)f''(x) - f'(x)f''(x) - f(x)f'''(x)}{\left(f'(x)^2 - f(x)f''(x)\right)^2} \end{aligned}$$
sledi, da je $g'(\beta) = 2$, kar pomeni, da je β odbojna negibna točka.

Naloga 2.12. V odvisnosti od začetnih približkov analizirajte konvergenco tangentne metode za funkcijo $f(x) = x^3 - x$.

1. Razvijte f v Taylorjevo vrsto okoli trenutnega približka in jo izvrednotite v 1. S pomočjo dobljene formule dokažite, da za vsak začetni približek iz $(1/\sqrt{3}, \infty)$ metoda konvergira k 1. S podobnim sklepom utemeljite, da za začetni približek iz $(-\infty, -1/\sqrt{3})$ metoda konvergira k -1.

2. Opišite, do kakšnih težav pride pri začetnih približkih $\pm 1/\sqrt{3}$ in $\pm 1/\sqrt{5}$.

3. Kako se obnaša zaporedje približkov pri začetnih približkih iz $(-1/\sqrt{5}, 1/\sqrt{5})$. Ali metoda konvergira k 0?

4. Ugotovite, kam konvergira zaporedje približkov pri začetnih približkih z intervalov $(-1/\sqrt{3}, -1/\sqrt{5})$ in $(1/\sqrt{5}, 1/\sqrt{3})$.

Rešitev. Iteracijska funkcija je podana s predpisom

$$g(x) = x - \frac{x^3 - x}{3x^2 - 1} = \frac{2x^3}{3x^2 - 1}.$$

Vemo, da so negibne točke iteracije ničle funkcije f, to so -1, 0 in 1. Ker so enostavne ničle, vemo tudi, da je konvergenca v neki njihovi okolici vsaj kvadratična. Oglejmo si natančneje, kaj se dogaja z iteracijskimi zaporedji pri različnih začetnih približkih.

1. Iz razvoja f v Taylorjevo vrsto okoli x_{r-1}, izvrednotenega v ničli 1, sledi, da je

$$x_r - 1 = \frac{f''(\xi_{r-1})}{2f'(x_{r-1})}(x_{r-1} - 1)^2, \qquad r \in \mathbb{N},$$

   za nek ξ_{r-1} med 1 in x_{r-1}. Ker je f na intervalu $(1/\sqrt{3}, \infty)$ strogo naraščajoča in strogo konveksna, za vsak x_{r-1} s tega intervala velja $x_r > 1$ za vsak $r \in \mathbb{N}$. V posebnem to pomeni, da za začetni približek $x_0 \in (1/\sqrt{3}, \infty)$ velja $f(x_1) > 0$, zato iz definicije tangentne metode sledi $1 < x_2 < x_1$. Po indukciji lahko ta sklep nadaljujemo do ugotovitve, da je zaporedje $(x_r)_r$ strogo padajoče in navzdol omejeno z 1. Ker je 1 edina ničla funkcije f na intervalu $(1/\sqrt{3}, \infty)$, smo s tem dokazali, da za vsak začetni približek s tega intervala zaporedje $(x_r)_r$ konvergira k 1. Podoben razmislek pripelje do zaključka, da zaporedje $(x_r)_r$ konvergira k -1 za vsak začetni približek $x_0 \in (-\infty, -1/\sqrt{3})$.

2. Tangentna metoda pri začetnih približkih $x_0 = \pm 1/\sqrt{3}$ propade, saj sta $\pm 1/\sqrt{3}$ pola iteracijske funkcije g. Težave lahko pričakujemo tudi pri začetnih približkih $\pm 1/\sqrt{5}$, saj sta rešitvi enačbe

$$g(x_0) = \frac{2x_0^3}{3x_0^2 - 1} = -x_0.$$

   To pomeni, da za prva dva približka velja $x_2 = -x_1 = x_0$, iz česar sklepamo, da začetna približka $\pm 1/\sqrt{5}$ povzročita cikel reda 2.

3. Obravnavajmo začetne približke z intervala $(-1/\sqrt{5}, 1/\sqrt{5})$. Prepričajmo se, da zagotavljajo konvergenco k 0. Najprej iz padanja iteracijske funkcije g na obravnavanem intervalu sklepamo, da je $g(x) \in (0, 1/\sqrt{5})$ za vsak $x \in (-1/\sqrt{5}, 0)$ in $g(x) \in (-1/\sqrt{5}, 0)$ za vsak $x \in (0, 1/\sqrt{5})$. Nato opazimo še, da je $g(x) + x < 0$ za vsak $x \in (-1/\sqrt{5}, 0)$ in $g(x) + x > 0$ za vsak $x \in (0, 1/\sqrt{5})$. Od tod sledi, da členi zaporedja alternirajoče menjavajo predznak. Če je $x_0 > 0$, zaporedje sodih členov pada proti 0, zaporedje lihih členov pa raste proti 0. Za $x_0 < 0$ je situacija ravno obratna, ne glede na to pa celotno zaporedje približkov konvergira k 0.

4. Za začetne približke z intervala $(-1/\sqrt{3}, -1/\sqrt{5})$ lahko s pomočjo grafa iteracijske funkcije g in simetrale lihih kvadrantov razberemo, da zaporedje približkov konvergira bodisi k -1 bodisi k 1, ki je najbolj oddaljena ničla. Pri začetnem približku $x_0 = -1/2$ je na primer $x_1 = 1$, zato je zaradi zveznosti g pri začetnih približkih blizu $-1/2$ približek x_1 blizu 1, kar implicira konvergenco k 1. Konvergenco k -1 v resnici dobimo le za začetne približke x_0 na majhnem odseku, kjer je $g(x_0) \in (1/\sqrt{5}, 1/\sqrt{3})$. Po simetriji podobno velja za začetne približke z intervala $(1/\sqrt{5}, 1/\sqrt{3})$. Zaključimo, da je tangentna metoda za določene začetne približke lahko zelo nepredvidljiva.

Naloga 2.13. Podajte primer funkcije f, za katero tangentna metoda propade ne glede na izbiro začetnega približka. Natančneje, poiščite tako funkcijo f, da za poljuben začetni približek x_0 za vsak približek x_r, $r \in \mathbb{N}$, velja $x_r = -x_{r-1}$.

Rešitev. Poskušajmo določiti tako funkcijo f, da pri tangentni metodi za vsak začetni približek x_0 velja $x_r = -x_{r-1}$, $r \in \mathbb{N}$. To pomeni, da se tekom iteracije izmenjujeta približka x_0 in $-x_0$. Ker je $x_r = x_{r-1} - f(x_{r-1})/f'(x_{r-1})$, mora biti funkcija f rešitev navadne diferencialne enačbe

$$\frac{f'(x)}{f(x)} = \frac{1}{2x}$$

in zato zadošča $|f(x)| = C\sqrt{|x|}$ za neko konstanto C. Vzemimo na primer

$$f(x) = \sqrt{|x|}.$$

Ker je $f'(x) = x/(2\,|x|\,\sqrt{|x|})$, res velja

$$x_r = x_{r-1} - \frac{f(x_{r-1})}{f'(x_{r-1})} = x_{r-1} - \frac{\sqrt{|x_{r-1}|} \cdot 2\,|x_{r-1}|\,\sqrt{|x_{r-1}|}}{x_{r-1}} = -x_{r-1}$$

za vsak $r \in \mathbb{N}$.

Naloga 2.14. Implementirajte tangentno metodo v Matlabu in z njo poiščite ničlo funkcije f, ki je podana s predpisom $f(x) = x + 4 - e^{x^2}$. Koliko korakov iteracije je treba izvesti pri začetnem približku $x_0 = 1$, da se zadnji izračunani približek v 10 decimalkah ujema z vrednostjo približka, ki ga dobite z vgrajeno funkcijo `fzero`?

Rešitev. Pri implementaciji uporabimo funkcijo `iteracija` iz naloge 2.5.

```
function [x,X,k] = tangentna(f,df,x0,tol,N)
% funkcija
%  [x,X,k] = tangentna(f,df,x0,tol,N)
% izvede tangentno metodo za iskanje ničle funkcije f
%
% vhodna podatka:
%  f    funkcija, ničlo katere iščemo,
%  df   odvod funkcije f.
%
% ostali vhodni in izhodni podatki so enaki kot pri
% funkciji 'iteracija'

g = @(x) x - f(x)/df(x);
[x,X,k] = iteracija(g,x0,tol,N);

end
```

V tabeli 2.1 je podanih prvih osem približkov tangentne metode, ki jih dobimo z izvedbo ukaza `[x,X,k] = tangentna(f,df,1,1e-15,100)` za primerno definirani funkciji `f` in `df`. Primerjamo jih z vrednostjo 1.29071842171596, ki je rezultat ukaza `fzero(f,0)`. Ugotovimo, da je približek x_6 na šestem koraku prvi, ki se z rezultatom vgrajene metode ujema v več kot desetih decimalkah. V zadnjem stolpcu tabele je podano skupno število funkcijskih izračunov funkcij f in f', uporabljenih za izračun posameznega približka.

korak	približek	napaka	funkcijski izračuni
1	1.51429853102254	$2.2358 \cdot 10^{-1}$	2
2	1.36287475280475	$7.2156 \cdot 10^{-2}$	4
3	1.29944386111342	$8.7254 \cdot 10^{-3}$	6
4	1.29085495577572	$1.3653 \cdot 10^{-4}$	8
5	1.29071845546466	$3.3749 \cdot 10^{-8}$	10
6	1.29071842171597	$1.9984 \cdot 10^{-15}$	12
7	1.29071842171596	0	14
8	1.29071842171596	0	16

Tabela 2.1: *Izvedba tangentne metode v nalogi 2.14.*

Slaba lastnost tangentne metode je, da za njeno izvedbo poleg vrednosti funkcije potrebujemo tudi vrednosti njenega odvoda. Temu se lahko izognemo z uporabo sekantne metode, ki jo geometrijsko interpretiramo na podoben način kot tangentno metodo: novi približek x_{r+1} določimo kot presečišče abscisne osi s sekanto funkcije f skozi točki x_r in x_{r-1}, kar da

$$x_{r+1} = x_r - \frac{f(x_r)(x_r - x_{r-1})}{f(x_r) - f(x_{r-1})}.$$

Sekantna metoda ni sestavljena po receptu navadne iteracije, saj za izračun novega približka namesto enega uporabimo dva prejšnja približka. Zato tudi pri začetku izvajanja iteracije potrebujemo dva začetna približka (x_0 in x_1).

Naloga 2.15. Dokažite, da iterativna formula

$$x_{r+1} = \frac{x_{r-1}x_r + a}{x_{r-1} + x_r}, \quad r = 1, 2, \ldots,$$

za računanje kvadratnega korena pozitivnega števila a ustreza sekantni metodi za funkcijo $f(x) = x^2 - a$ in da je zaporedje približkov $(x_r)_r$ konvergentno za poljubna začetna približka x_0 in x_1, večja od $\sqrt{a}$.

Rešitev. Po definiciji sekantne metode za funkcijo $f(x) = x^2 - a$ izračunamo

$$x_{r+1} = x_r - \frac{(x_r^2 - a)(x_r - x_{r-1})}{(x_r^2 - a) - (x_{r-1}^2 - a)} = \frac{x_r(x_r^2 - x_{r-1}^2) - (x_r^2 - a)(x_r - x_{r-1})}{x_r^2 - x_{r-1}^2},$$

kar se s krajšanjem izraza $x_r - x_{r-1}$ poenostavi v iskano formulo. Za dokaz konvergence zaporedja približkov najprej opazimo, da so pri pozitivnih začetnih približkih vsi členi zaporedja pozitivni. Ker je

$$x_{r+1} - x_r = \frac{x_{r-1}x_r + a}{x_{r-1} + x_r} - x_r = \frac{a - x_r^2}{x_{r-1} + x_r},$$

je približek x_{r+1} manjši od x_r, če je $x_r > \sqrt{a}$. Poleg tega je

$$x_{r+1} - \sqrt{a} = \frac{x_{r-1}x_r + a}{x_{r-1} + x_r} - \sqrt{a} = \frac{(x_{r-1} - \sqrt{a})(x_r - \sqrt{a})}{x_{r-1} + x_r},$$

iz česar lahko sklepamo, da je za poljubna začetna približka x_0 in x_1, ki sta večja od $\sqrt{a}$, zaporedje padajoče in navzdol omejeno s $\sqrt{a}$. Torej ima limito $\alpha \geq \sqrt{a}$, ki zadošča enačbi

$$\alpha = \frac{\alpha^2 + a}{\alpha + \alpha},$$

od kjer sledi $\alpha = \sqrt{a}$.

Naloga 2.16. Implementirajte sekantno metodo v Matlabu in jo preizkusite s funkcijo f iz naloge 2.14 pri začetnih približkih $x_0 = 1$ in $x_1 = 1.1$. Dobljene rezultate primerjajte z rezultati tangentne metode.

Rešitev. Funkcijo `sekantna`, ki izvede sekantno metodo, implementiramo na zelo podoben način kot funkcijo `iteracija` iz naloge 2.5. Pri tem pazimo, da funkcijsko vrednost za vsak približek izračunamo le enkrat, saj izvrednotenje funkcije predstavlja največji računski zalogaj pri izvedbi metode.

```
function [x,X,k] = sekantna(f,x0,x1,tol,N)
% funkcija
%  [x,X,k] = sekantna(f,x0,x1,tol,N)
% izvede sekantno metodo za iskanje ničle funkcije f
%
% vhodni podatki:
%  f         funkcija, ničlo katere iščemo,
%  x0, x1    začetna približka metode
%
% ostali vhodni in izhodni podatki so enaki kot pri
% funkciji 'iteracija'

X = [x0 x1];
k = 0;
fxk = f(X(1));
while k < N
    k = k+1;
    fxkn = f(X(k+1));
    X(k+2) = X(k+1) - fxkn*(X(k+1)-X(k))/(fxkn-fxk);
    fxk = fxkn;
    if abs(X(k+2)-X(k+1)) < tol
        break;
    end
end
x = X(k+2);

end
```

Približki, izračunani z izvedbo ukaza `[x,X,k] = sekantna(f,1,1.1,1e-15,100)`, so prikazani v tabeli 2.2. Približek x_8, dobljen na sedmem koraku, se z vrednostjo `fzero(f,0)` ujema v enajstih decimalkah, približek x_9 pa v vseh. Torej je število korakov pri sekantni metodi v tem primeru le za ena večje od števila korakov pri tangentni metodi. Pravzaprav je sekantna metoda učinkovitejša, saj smo na vsakem koraku opravili le en izračun funkcijske vrednosti namesto dveh pri tangentni metodi.

korak	približek	napaka	funkcijski izračuni
1	1.42632775255262	$1.3561 \cdot 10^{-1}$	3
2	1.24363560557646	$4.7083 \cdot 10^{-2}$	4
3	1.27978689949541	$1.0932 \cdot 10^{-2}$	5
4	1.29168083871343	$9.6242 \cdot 10^{-4}$	6
5	1.29069926729793	$1.9154 \cdot 10^{-5}$	7
6	1.29071838835441	$3.3362 \cdot 10^{-8}$	8
7	1.29071842171712	$1.1571 \cdot 10^{-12}$	9
8	1.29071842171596	0	10
9	1.29071842171596	0	11

TABELA 2.2: *Izvedba sekantne metode v nalogi 2.16.*

Ena izmed posplošitev tangentne metode je metoda (f, f', f''), pri kateri poleg prvega uporabimo tudi drugi odvod funkcije f. Približek x_{r+1} izračunamo na podlagi približka x_r po formuli

$$x_{r+1} = x_r - \frac{f(x_r)}{f'(x_r)} - \frac{f''(x_r)f(x_r)^2}{2f'(x_r)^3}.$$

V literaturi se ta iteracija pogosto navaja kot Schröderjeva metoda drugega reda. Metode višjega reda vključujejo še višje odvode funkcije f.

Naloga 2.17. Naj bo f analitična funkcija z enostavno ničlo α. Z uporabo razvoja inverzne funkcije f v Taylorjevo vrsto izpeljite metodo (f, f', f''). Z odvajanjem iteracijske funkcije izračunajte red konvergence iteracijskega zaporedja v okolici α.

Rešitev. Ker je α enostavna ničla funkcije f, je $f'(\alpha) \neq 0$ in po izreku o inverzni funkciji obstaja $\delta > 0$, da je f na $(\alpha - \delta, \alpha + \delta)$ obrnljiva. Natančneje, obstaja $\varepsilon > 0$ in taka funkcija $F : (-\varepsilon, \varepsilon) \to (\alpha - \delta, \alpha + \delta)$, da je $F(f(x)) = x$ za vsak $x \in (\alpha - \delta, \alpha + \delta)$. Funkcijo F razvijemo v Taylorjevo vrsto okoli $y \in (-\varepsilon, \varepsilon)$:

$$F(z) = F(y) + F'(y)(z - y) + \frac{1}{2}F''(y)(z - y)^2 + \dots.$$

Vzemimo $z = 0$ in $y = f(x)$ za $x \in (\alpha - \delta, \alpha + \delta)$. Iz

$$F(f(x)) = x, \qquad F'(f(x))f'(x) = 1, \qquad F''(f(x))f'(x)^2 + F'(f(x))f''(x) = 0$$

izrazimo $F'(f(x)) = 1/f'(x)$ in $F''(f(x)) = -f''(x)/f'(x)^3$. Če v Taylorjevi vrsti zanemarimo člene s stopnjo, višjo od 2, dobimo

$$\alpha \approx x - \frac{f(x)}{f'(x)} - \frac{f''(x)f(x)^2}{2f'(x)^3},$$

kar določa iteracijsko funkcijo

$$g(x) = x - \frac{f(x)}{f'(x)} - \frac{f''(x)f(x)^2}{2f'(x)^3}.$$

Red konvergence metode določimo z odvajanjem g. Izračunamo

$$g'(x) = \frac{3f''(x)^2 - f'(x)f'''(x)}{2f'(x)^4} f(x)^2.$$

Od tod je razvidno, da je $g'(\alpha) = 0$ in $g''(\alpha) = 0$, medtem ko je vrednost $g'''(\alpha)$ v splošnem različna od nič, zato je red konvergence kubičen.

Še ena metoda, ki poseže po drugem odvodu pri računanju približka za ničlo funkcije f, je Halleyjeva metoda. Približek x_{r+1} izračunamo iz približka x_r po formuli

$$x_{r+1} = x_r - \frac{2f(x_r)f'(x_r)}{2f'(x_r)^2 - f(x_r)f''(x_r)}.$$

Halleyjeva metoda spada v razred Householderjevih metod

$$x_{r+1} = x_r + d\frac{(1/f)^{(d-1)}(x_r)}{(1/f)^{(d)}(x_r)}.$$

Dobimo jo pri $d = 2$, medtem ko tangentna metoda ustreza izbiri $d = 1$.

Naloga 2.18. Naj bo f vsaj dvakrat odvedljiva funkcija. Dokažite, da Halleyjeva metoda ustreza tangentni metodi za funkcijo $F(x) = f(x)/\sqrt{|f'(x)|}$. Kakšen je red konvergence metode v okolici enostavne ničle funkcije f ob predpostavki, da je f vsaj petkrat odvedljiva?

Rešitev. Izračunamo

$$F'(x) = \frac{2f'(x)^2 - f(x)f''(x)}{2f'(x)\sqrt{|f'(x)|}}$$

in izpeljemo iteracijsko funkcijo

$$g(x) = x - \frac{F(x)}{F'(x)} = x - \frac{2f(x)f'(x)}{2f'(x)^2 - f(x)f''(x)}.$$

Z odvajanjem iteracijske funkcije določimo še red metode. Z nekaj računanja dobimo, da je

$$g'(x) = \frac{3f''(x)^2 - 2f'(x)f'''(x)}{(f(x)f''(x) - 2f'(x)^2)^2} f(x)^2,$$

iz česar sledi, da je red konvergence iteracijskega zaporedja v okolici ničle funkcije f v splošnem kubičen.

Naloga 2.19. Poenostavite Halleyjevo metodo za funkcijo f, ki je podana s predpisom $f(x) = x^2 - a$, $a > 0$. Rezultat iteracije ob ustreznem začetnem približku tedaj predstavlja približek za $\sqrt{a}$.

Rešitev. Prva odvoda funkcije $f(x) = x^2 - a$ sta $f'(x) = 2x$ in $f''(x) = 2$. Z nekaj računanja iteracijsko funkcijo g poenostavimo v

$$g(x) = x\frac{x^2 + 3a}{3x^2 + a}.$$

Zanjo smo v nalogi 2.7 že dokazali, da določa iteracijsko zaporedje, ki konvergira k $\sqrt{a}$ za vsak začetni približek $x_0 > 0$.

Naloga 2.20. V Matlabu primerjajte tangetno metodo, sekantno metodo, metodo (f, f', f'') in Halleyjevo metodo za iskanje ničle funkcije f iz naloge 2.14. Preizkusite vse metode pri začetnih približkih x_0 iz seznama `1:0.1:10` (ter $x_1 = x_0 + 0.1$ za sekantno metodo). Pri vsaki izvedbi metode poiščite najmanjše tako število k, da se približek x_k absolutno razlikuje od rezultata `fzero(f,1)` za manj kot 10^{-10}. Nato za vsako metodo narišite graf števil k v odvisnosti od začetnih približkov ter graf števila funkcijskih izračunov v odvisnosti od začetnih približkov. Komentirajte rezultate.

Rešitev. Implementacija tangentne metode (funkcija `tangentna`) je opisana v nalogi 2.14, implementacija sekantne metode (funkcija `sekantna`) pa v nalogi 2.16. Funkciji `fdfddf` in `halley` za izvedbo metode (f, f', f'') in Halleyjeve metode implementiramo na podoben način, pomagamo si lahko s funkcijo `iteracija` iz naloge 2.5. Rezultati izvedbe ukazov `[x,X,k] = fdfddf(f,df,ddf,1,1e-15,100)` in `[x,X,k] = halley(f,df,ddf,1,1e-15,100)` ob primerno definiranih funkcijah `f`, `df`, `ddf`, ki določajo f, f', f'', so prikazani v tabelah 2.3 in 2.4.

korak	približek	napaka	funkcijski izračuni
1	1.02811573481699	$2.6260 \cdot 10^{-1}$	3
2	1.11691838286784	$1.7380 \cdot 10^{-1}$	6
3	1.25367570678920	$3.7043 \cdot 10^{-2}$	9
4	1.29047854328615	$2.3988 \cdot 10^{-4}$	12
5	1.29071842165685	$5.9109 \cdot 10^{-11}$	15
6	1.29071842171596	0	18
7	1.29071842171596	0	21

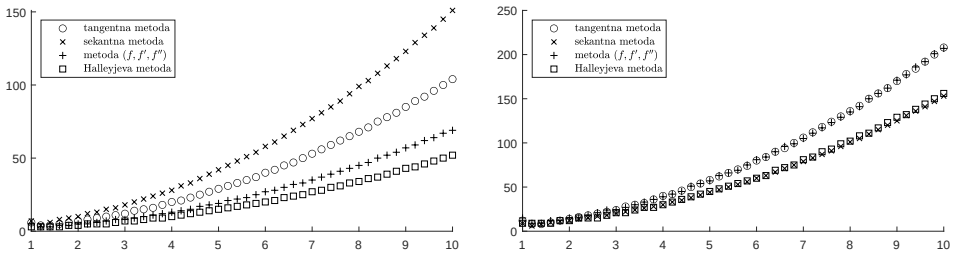
TABELA 2.3: *Izvedba metode (f, f', f'') v nalogi 2.20.*

korak	približek	napaka	funkcijski izračuni
1	1.26437572777572	$2.6343 \cdot 10^{-2}$	3
2	1.29070002729498	$1.8394 \cdot 10^{-5}$	6
3	1.29071842171596	$6.2172 \cdot 10^{-15}$	9
4	1.29071842171596	0	12
5	1.29071842171596	0	15

TABELA 2.4: *Izvedba Halleyjeve metode v nalogi 2.20.*

Za vsako metodo in vsak začetni približek x_0 iz seznama `1:0.1:10` poiščemo najmanjši k, da se približek x_k od vrednosti `x`, dobljene z vgrajeno funkcijo, absolutno razlikuje za manj kot `e = 1e-10`. Če je `X` seznam vseh izračunanih približkov (za katerega predpostavimo, da vsebuje približek x_k, ki zadošča kriteriju), lahko indeks k poiščemo z ukazom `find(abs(X - x) < e, 1)`. Ta namreč vrne indeks prvega elementa v seznamu `X`, ki se od `x` absolutno razlikuje za manj kot `e`. Ker je prvi element seznama `X` začetni približek x_0, moramo od rezultata odšteti 1 (oziroma 2 pri sekantni metodi).

```
f = @(x) x + 4 - exp(x^2);
df = @(x) 1 - 2*x*exp(x^2);
ddf = @(x) - 2*exp(x^2) - 4*x^2*exp(x^2);

x = fzero(f,1);
x0 = 1:0.1:10; tol = 1e-15; N = 200; e = 1e-10;

[k1,k2,k3,k4] = deal(zeros(size(x0)));
for i = 1:length(x0)
    [~,X1] = tangentna(f,df,x0(i),tol,N);
    k1(i) = find(abs(X1-x) < e, 1) - 1;

    [~,X2] = sekantna(f,x0(i),x0(i)+0.1,tol,N);
    k2(i) = find(abs(X2-x) < e, 1) - 2;

    [~,X3] = fdfddf(f,df,ddf,x0(i),tol,N);
    k3(i) = find(abs(X3-x) < e, 1) - 1;

    [~,X4] = halley(f,df,ddf,x0(i),tol,N);
    k4(i) = find(abs(X4-x) < e, 1) - 1;
end
```



(A) *Število korakov* (B) *Število funkcijskih izračunov*

Slika 2.3: *Primerjava metod za reševanje nelinearnih enačb na primeru iz naloge 2.20.*

Grafi, s katerimi primerjamo učinkovitost metod, so prikazani na sliki 2.3. Pri pripravi grafov števila funkcijskih izračunov upoštevamo, da pri tangentni metodi

na vsakem koraku izvrednotimo dve funkciji, pri sekantni eno, pri metodi (f, f', f'') in Halleyjevi metodi pa tri. Rezultati kažejo, da za izbrano funkcijo f najhitreje konvergira Halleyjeva metoda, najpočasneje pa sekantna metoda. Metoda (f, f', f'') konvergira hitreje od tangentne. Iz grafa funkcijskih izračunov pa je razvidno, da sta sekantna in Halleyjeva metoda učinkovitejši od tangentne metode in metode (f, f', f'').

3. Sistemi linearnih enačb

Sistem $n \in \mathbb{N}$ linearnih enačb z n neznankami lahko predstavimo v obliki $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, kjer je $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ kvadratna matrika, $\boldsymbol{x} \in \mathbb{R}^n$ vektor neznank in $\boldsymbol{b} \in \mathbb{R}^n$ vektor prostih vrednosti. Rešitev sistema obstaja in je enolična natanko tedaj, ko je matrika $\boldsymbol{A}$ obrnljiva. Pri računanju rešitve $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ se običajno poslužujemo direktnih metod, ki se izognejo računanju inverza matrike. Med metodami izbiramo na podlagi občutljivosti in drugih posebnih lastnosti matrike sistema.

3.1. Matrične norme in občutljivost matrik

Matrična norma $\|\cdot\|$ je preslikava, ki kvadratni matriki priredi realno vrednost. Določena je na podoben način kot vektorska norma, le da jo poleg treh osnovnih lastnosti (pozitivnost, homogenost in trikotniška neenakost) definira še četrta lastnost (submultiplikativnost), ki povezuje normo produkta s produktom norm. Natančneje, za vsaki matriki $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ in vsak skalar $\alpha \in \mathbb{C}$ velja:

- $\|\boldsymbol{A}\| \geq 0$ in $\|\boldsymbol{A}\| = 0$ natanko tedaj, ko je $\boldsymbol{A} = \boldsymbol{0}$ (pozitivnost),
- $\|\alpha\boldsymbol{A}\| = |\alpha|\,\|\boldsymbol{A}\|$ (homogenost),
- $\|\boldsymbol{A} + \boldsymbol{B}\| \leq \|\boldsymbol{A}\| + \|\boldsymbol{B}\|$ (trikotniška neenakost),
- $\|\boldsymbol{A}\boldsymbol{B}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{B}\|$ (submultiplikativnost).

Naloga 3.1. Naj bo $\|\cdot\|$ matrična norma. Dokažite, da norma $\|\boldsymbol{I}\|$ identitete $\boldsymbol{I}$ ni manjša od 1 in sklepajte, da za vsako obrnljivo matriko $\boldsymbol{A}$ velja $\|\boldsymbol{A}^{-1}\| \geq \|\boldsymbol{A}\|^{-1}$.

Rešitev. Iz submultiplikativnosti norme sledi $\|\boldsymbol{I}\| \leq \|\boldsymbol{I}\|^2$ in ker je $\|\boldsymbol{I}\| > 0$, to dokazuje $\|\boldsymbol{I}\| \geq 1$. Za obrnljivo matriko $\boldsymbol{A}$ po submultiplikativnosti norme velja

$$\|\boldsymbol{I}\| = \left\|\boldsymbol{A}\boldsymbol{A}^{-1}\right\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{A}^{-1}\|,$$

kar potrjuje zvezo $\|\boldsymbol{A}^{-1}\| \geq \|\boldsymbol{A}\|^{-1}$.

Naloga 3.2. Naj bo $\|\cdot\|$ matrična norma. Dokažite, da za vsako matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ velja $\rho(\boldsymbol{A}) \leq \|\boldsymbol{A}\|$, kjer $\rho(\boldsymbol{A})$ označuje spektralni radij matrike $\boldsymbol{A}$.

Rešitev. Spektralni radij $\rho(\boldsymbol{A})$ ustreza absolutni vrednosti dominantne lastne vrednosti λ matrike $\boldsymbol{A}$. Naj bo $\boldsymbol{x} \in \mathbb{C}^n$ pripadajoči lastni vektor in $\boldsymbol{X} \in \mathbb{C}^{n\times n}$ matrika, v kateri je vsak stolpec enak $\boldsymbol{x}$. Iz homogenosti in submultiplikativnosti matrične norme sledi

$$|\lambda|\,\|\boldsymbol{X}\| = \|\lambda\boldsymbol{X}\| = \|\boldsymbol{A}\boldsymbol{X}\| \leq \|\boldsymbol{A}\|\,\|\boldsymbol{X}\|\,.$$

Ker je $\boldsymbol{x} \neq \boldsymbol{0}$, je $\boldsymbol{X} \neq \boldsymbol{0}$ in zato $\|\boldsymbol{X}\| > 0$. Sledi $|\lambda| \leq \|\boldsymbol{A}\|$.

Matrične norme pogosto definiramo s pomočjo vektorskih. Preprosta ideja je, da elemente $a_{i,j}$, $i,j = 1,2,\ldots,n$, matrike $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ zložimo v vektor

$$\mathrm{vec}(\boldsymbol{A}) = (a_{1,1}, a_{1,2}, \ldots, a_{1,n}, a_{2,1}, a_{2,2}, \ldots, a_{2,n}, \ldots, a_{n,1}, a_{n,2}, \ldots, a_{n,n})$$

in na njem uporabimo vektorsko p-normo $\|\cdot\|_p$, $p \geq 1$. Dobimo preslikavo

$$N_p(\boldsymbol{A}) = \|\mathrm{vec}(\boldsymbol{A})\|_p\,,$$

ki ustreza prvim trem lastnostim matrične norme.

Naloga 3.3. Naj $a_{i,j}$, $i,j = 0,1,\ldots,n$, označujejo elemente matrike $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ in naj bo

$$N_\infty(\boldsymbol{A}) = \max_{i,j} |a_{i,j}|.$$

1. S primerom pokažite, da N_∞ ni matrična norma.
2. Utemeljite, da $\|\boldsymbol{A}\| = nN_\infty(\boldsymbol{A})$ je matrična norma.

Rešitev.

1. Preslikava N_∞ ni matrična norma, saj submultiplikativnost ni zagotovljena; za

$$\boldsymbol{A} = \boldsymbol{B} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

   na primer velja $N_\infty(\boldsymbol{A}\boldsymbol{B}) = 2 > 1 = N_\infty(\boldsymbol{A})N_\infty(\boldsymbol{B})$.

2. Preverimo, da preslikava $\|\cdot\|$ izpolnjuje vse lastnosti iz definicije matrične norme. Naj bosta $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{C}^{n\times n}$ poljubni matriki z elementi $a_{i,j}$, $b_{i,j}$ ter naj bo $\alpha \in \mathbb{C}$ poljuben skalar.

   - Pozitivnost: Po definiciji absolutne vrednost je

$$\|\boldsymbol{A}\| = n \max_{i,j} |a_{i,j}| \geq 0.$$

     Prav tako velja

$$\|\boldsymbol{A}\| = n \max_{i,j} |a_{i,j}| = 0 \quad \Leftrightarrow \quad a_{i,j} = 0 \;\; \forall\, i,j \quad \Leftrightarrow \quad \boldsymbol{A} = \boldsymbol{0}.$$

- Homogenost: Po pravilu za absolutno vrednost produkta je

$$\|\alpha \boldsymbol{A}\| = n \max_{i,j} |\alpha\, a_{i,j}| = n\ |\alpha| \max_{i,j} |a_{i,j}| = |\alpha|\ \|\boldsymbol{A}\|\,.$$

- Trikotniška neenakost: Iz trikotniške neenakosti za absolutno vrednost sledi

$$\|\boldsymbol{A}+\boldsymbol{B}\| = n \max_{i,j} |a_{i,j}+b_{i,j}| \le n \left(\max_{i,j} |a_{i,j}| + \max_{i,j} |b_{i,j}| \right) = \|\boldsymbol{A}\| + \|\boldsymbol{B}\|\,.$$

- Submultiplikativnost: Po trikotniški neenakosti in pravilu za absolutno vrednost produkta velja

$$\begin{aligned}\|\boldsymbol{AB}\| = n \max_{i,j} \left| \sum_{k=1}^{n} a_{i,k} b_{k,j} \right| &\le n \max_{i,k} |a_{i,k}|\ \max_{j} \sum_{k=1}^{n} |b_{k,j}| \\ &\le n \max_{i,k} |a_{i,k}|\ n \max_{k,j} |b_{k,j}| = \|\boldsymbol{A}\|\ \|\boldsymbol{B}\|\,.\end{aligned}$$

Naloga 3.4. Frobeniusova norma je izpeljana iz evklidske vektorske norme. Za matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ z elementi $a_{i,j}$, $i,j = 0,1,\ldots,n$, je podana s predpisom

$$\|\boldsymbol{A}\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}.$$

1. Prepričajte se, da je $\|\cdot\|_F$ matrična norma.
2. Dokažite, da je $\|\boldsymbol{A}\|_F^2 = \operatorname{sled}(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$.

Rešitev.

1. Pozitivnost, homogenost in trikotniška neenakost sledijo iz lastnosti vektorske 2-norme. Preverimo submultiplikativnost. Ker za matriki $\boldsymbol{A},\boldsymbol{B} \in \mathbb{C}^{n\times n}$ z elementi $a_{i,j}$, $b_{i,j}$ po Cauchy–Schwarzevi neenakosti velja

$$\sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{i,k} b_{k,j} \right|^2 \le \sum_{i,j=1}^{n} \left(\sum_{k=1}^{n} |a_{i,k}|^2 \sum_{k=1}^{n} |b_{k,j}|^2 \right),$$

je

$$\|\boldsymbol{AB}\|_F = \sqrt{\sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{i,k} b_{k,j} \right|^2} \le \sqrt{\sum_{i,k=1}^{n} |a_{i,k}|^2} \sqrt{\sum_{k,j=1}^{n} |b_{k,j}|^2} = \|\boldsymbol{A}\|_F\ \|\boldsymbol{B}\|_F\,.$$

2. Element matrike $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ na mestu (i,j) je enak $\sum_{k=1}^{n} \overline{a_{k,i}} a_{k,j}$, zato je

$$\operatorname{sled}(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}) = \sum_{i=1}^{n} \sum_{k=1}^{n} \overline{a_{k,i}} a_{k,i} = \sum_{i,k=1}^{n} |a_{k,i}|^2 = \|\boldsymbol{A}\|_F^2\,.$$

Pomemben razred matričnih norm so operatorske norme, ki jih na podlagi izbrane vektorske norme $\|\cdot\|$ za matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ definiramo s predpisom

$$\|\boldsymbol{A}\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Definicijo lahko poenostavimo, velja $\|\boldsymbol{A}\| = \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{A}\boldsymbol{x}\|$.

Naloga 3.5. Naj bo $\|\cdot\|$ operatorska norma. Dokažite, da za obrnljivo matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ velja

$$\left\|\boldsymbol{A}^{-1}\right\|^{-1} = \min_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}.$$

Rešitev. Ker je matrika $\boldsymbol{A}$ obrnljiva, za vsak $\boldsymbol{x} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$ obstaja $\boldsymbol{y} \in \mathbb{C}^n \setminus \{\boldsymbol{0}\}$, da je $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{y}$. Torej po definiciji operatorske norme velja

$$\left\|\boldsymbol{A}^{-1}\right\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\left\|\boldsymbol{A}^{-1}\boldsymbol{x}\right\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{A}\boldsymbol{y}\neq\boldsymbol{0}} \frac{\|\boldsymbol{y}\|}{\|\boldsymbol{A}\boldsymbol{y}\|}$$

oziroma, ker je množica $\{\boldsymbol{A}\boldsymbol{y};\, \boldsymbol{y} \in \mathbb{C}^n\} \setminus \{\boldsymbol{0}\}$ enaka množici $\mathbb{C}^n \setminus \{\boldsymbol{0}\}$,

$$\left\|\boldsymbol{A}^{-1}\right\| = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{x}\|}{\|\boldsymbol{A}\boldsymbol{x}\|} = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \left(\frac{\|\boldsymbol{A}\boldsymbol{x}\|}{\|\boldsymbol{x}\|}\right)^{-1}.$$

Slednje po zvezi med maksimumom in minimumom dokazuje trditev.

Matrične p-norme $\|\cdot\|_p$, $p \geq 1$, so primer operatorskih norm, ki so inducirane iz vektorskih p-norm. Posebej enostavni sta 1-norma in ∞-norma, za kateri velja

$$\|\boldsymbol{A}\|_1 = \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{i,j}|, \quad \|\boldsymbol{A}\|_\infty = \max_{i=1,\ldots,n} \sum_{j=1}^{n} |a_{i,j}|$$

in ju lahko torej določimo kar s seštevanjem absolutnih vrednosti elementov $a_{i,j}$ matrike $\boldsymbol{A} \in \mathbb{C}^{n\times n}$.

Naloga 3.6. Dokažite, da za poljubno matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ velja $\|\boldsymbol{A}\|_1 = \|\boldsymbol{A}^{\mathsf{H}}\|_\infty$.

Rešitev. Iz lastnosti 1-norme in ∞-norme sledi

$$\left\|\boldsymbol{A}^{\mathsf{H}}\right\|_\infty = \max_{i=1,\ldots,n} \sum_{j=0}^{n} |\overline{a_{j,i}}| = \max_{j=1,\ldots,n} \sum_{i=0}^{n} |a_{i,j}| = \|\boldsymbol{A}\|_1.$$

Operatorska 2-norma je ena izmed najbolj standardnih matričnih norm. Imenujemo jo tudi spektralna norma, saj za matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ velja

$$\|\boldsymbol{A}\|_2 = \max_{i=1,\ldots,n} \sqrt{\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})},$$

pri čemer so z $\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$, $i = 1, 2, \ldots, n$, označene lastne vrednosti matrike $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$. Ker je matrika $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ hermitska in nenegativno definitna, so vse lastne vrednosti nenegativne. Vrednosti $\sqrt{\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})}$ imenujemo singularne vrednosti matrike $\boldsymbol{A}$. Izračun 2-norme je običajno zahteven, a včasih zadostuje že ocena, ki jo dobimo na podlagi lažje izračunljivih norm.

Naloga 3.7. Dokažite, da za poljubno matriko $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ velja $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}^{\mathsf{H}}\|_2$.

Rešitev. Naj bo $i \in \{1, 2, \ldots, n\}$ in naj bo $\boldsymbol{x}_i$ lastni vektor, ki pripada lastni vrednosti $\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$. Potem je

$$\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{x}_i = \boldsymbol{A}\,\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})\,\boldsymbol{x}_i = \lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})\,\boldsymbol{A}\boldsymbol{x}_i,$$

kar dokazuje, da je $\boldsymbol{A}\boldsymbol{x}_i$ lastni vektor, $\lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$ pa lastna vrednost matrike $\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}}$. Podoben sklep velja tudi v obratno smer. Torej imata matriki $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ in $\boldsymbol{A}\boldsymbol{A}^{\mathsf{H}}$ enak nabor lastnih vrednosti.

Naloga 3.8. Dokažite, da za spektralno normo $\|\boldsymbol{A}\|_2$ matrike $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ veljajo naslednje ocene.

1. $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\|_F \le \|\boldsymbol{A}\|_2 \le \|\boldsymbol{A}\|_F$
2. $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\|_\infty \le \|\boldsymbol{A}\|_2 \le \sqrt{n}\,\|\boldsymbol{A}\|_\infty$
3. $\frac{1}{\sqrt{n}}\|\boldsymbol{A}\|_1 \le \|\boldsymbol{A}\|_2 \le \sqrt{n}\,\|\boldsymbol{A}\|_1$
4. $N_\infty(\boldsymbol{A}) \le \|\boldsymbol{A}\|_2 \le nN_\infty(\boldsymbol{A})$
5. $\|\boldsymbol{A}\|_2 \le \sqrt{\|\boldsymbol{A}\|_1\,\|\boldsymbol{A}\|_\infty}$

Rešitev.

1. Upoštevamo ugotovitev iz naloge 3.4, da je $\|\boldsymbol{A}\|_F^2 = \operatorname{sled}(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A})$. Obe neenakosti dokažemo z uporabo dejstev, da je sled matrike enaka vsoti njenih lastnih vrednosti in da so lastne vrednosti matrike $\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}$ nenegativne.

2. Enostavno je preveriti, da za vsak vektor $\boldsymbol{x} \in \mathbb{C}^n$ velja

$$\|\boldsymbol{x}\|_\infty \le \|\boldsymbol{x}\|_2 \le \sqrt{n}\,\|\boldsymbol{x}\|_\infty\,.$$

Od tod sledi

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \le \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\sqrt{n}\,\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \sqrt{n}\,\|\boldsymbol{A}\|_\infty\,,$$

$$\|\boldsymbol{A}\|_2 = \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2} \ge \max_{\boldsymbol{x}\neq\boldsymbol{0}} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_\infty}{\sqrt{n}\,\|\boldsymbol{x}\|_\infty} = \frac{1}{\sqrt{n}}\,\|\boldsymbol{A}\|_\infty\,.$$

3. Oceni lahko dokažemo s podobnim sklepanjem kot v prejšnji točki, saj za vsak vektor $\boldsymbol{x} \in \mathbb{C}^n$ velja
$$\frac{1}{\sqrt{n}} \|\boldsymbol{x}\|_1 \leq \|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_1 .$$
Lahko pa se poslužimo kar zveze, dokazane v prejšnji točki, za matriko $\boldsymbol{A}^{\mathsf{H}}$ in neenakosti utemeljimo na podlagi nalog 3.6 in 3.7.

4. Element $a_{i,j}$ matrike $\boldsymbol{A}$ na mestu (i,j) lahko zapišemo kot $\boldsymbol{e}_i^{\mathsf{T}} \boldsymbol{A} \boldsymbol{e}_j$, kjer sta $\boldsymbol{e}_i$ in $\boldsymbol{e}_j$ standardna enotska vektorja. Po Cauchy–Schwarzevi neenakosti in definiciji $\|\cdot\|_2$ je
$$|a_{i,j}| = \left|\boldsymbol{e}_i^{\mathsf{T}} \boldsymbol{A} \boldsymbol{e}_j\right| \leq \|\boldsymbol{e}_i\|_2 \|\boldsymbol{A}\boldsymbol{e}_j\|_2 \leq \|\boldsymbol{A}\|_2 ,$$
iz česar sledi $N_\infty(\boldsymbol{A}) \leq \|\boldsymbol{A}\|_2$. Neenakost $\|\boldsymbol{A}\|_2 \leq n N_\infty(\boldsymbol{A})$ lahko dokažemo s pomočjo ocene v prvi točki naloge, če sklepamo
$$\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F \leq \sqrt{\sum_{i,j=1}^{n} |a_{i,j}|^2} \leq \sqrt{n^2 \max_{i,j} |a_{i,j}|^2} = n \max_{i,j} |a_{i,j}| = n N_\infty(\boldsymbol{A}).$$

5. Po nalogi 3.2 je spektralni radij matrike omejen z normo matrike, zato je
$$\|\boldsymbol{A}\|_2^2 = \max_{i=1,\ldots,n} \lambda_i(\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}) \leq \left\|\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\right\|_\infty .$$
Zaradi submultiplikativnosti in dejstva, dokazanega v nalogi 3.6, je
$$\left\|\boldsymbol{A}^{\mathsf{H}}\boldsymbol{A}\right\|_\infty \leq \left\|\boldsymbol{A}^{\mathsf{H}}\right\|_\infty \|\boldsymbol{A}\|_\infty = \|\boldsymbol{A}\|_1 \|\boldsymbol{A}\|_\infty$$
in ocena sledi.

Naloga 3.9. S pomočjo rezultatov iz naloge 3.8 poiščite čim boljšo oceno za spektralno normo matrike $\boldsymbol{A}$ z elementi
$$a_{i,j} = \begin{cases} -2i; & i = j \\ n-i; & j = i+1 \\ n-j; & i = j+1 \\ 0; & \text{sicer} \end{cases}, \qquad i,j = 1,2,\ldots,n.$$
Nato v Matlabu s pomočjo vgrajene funkcije `norm` preverite izplejane ocene za primer $n = 1000$ in izračunajte $\|\boldsymbol{A}\|_2$. Primerjajte čase izračunov posameznih norm.

Rešitev. Najprej izračunamo
$$\|\boldsymbol{A}\|_1 = \max_{j=1,\ldots,n} (|-2j| + |n-j+1| + |n-j|) = 2n+1.$$
Ker je $\boldsymbol{A}$ simetrična, je tudi $\|\boldsymbol{A}\|_\infty = 2n+1$. Po absolutni vrednosti največji element matrike $\boldsymbol{A}$ je $-2n$, zato je $N_\infty(\boldsymbol{A}) = 2n$. Nadalje,
$$\|\boldsymbol{A}\|_F^2 = \sum_{i=1}^{n} (-2i)^2 + 2\sum_{i=1}^{n-1} i^2 = 6\sum_{i=1}^{n-1} i^2 + 4n^2,$$

od kjer po formuli za piramidna števila sledi

$$\|\boldsymbol{A}\|_F = \sqrt{(n-1)n(2n-1)+4n^2} = \sqrt{2n^3+n^2+1}.$$

Na podlagi izpeljanih ocen je najboljša spodnja meja za $\|\boldsymbol{A}\|_2$ enaka

$$\max\left\{\sqrt{\frac{2n^3+n^2+1}{n}}, \frac{2n+1}{\sqrt{n}}, 2n\right\},$$

najboljša zgornja meja pa

$$\min\left\{\sqrt{2n^3+n^2+1}, (2n+1)\sqrt{n}, 2n^2, 2n+1\right\}.$$

To poenostavimo v oceno

$$2n \leq \|\boldsymbol{A}\|_2 \leq 2n+1.$$

V Matlabu lahko matriko $\boldsymbol{A}$ zgeneriramo z uporabo ukazov `diag`. Norme matrik računamo s funkcijo `norm`, ki ji kot prvi vhodni podatek podamo matriko, z drugim pa določimo, katero normo računamo.

```
n = 1000;
A = diag(n-1:-1:1,1) - 2*diag(1:n) + diag(n-1:-1:1,-1);
norm(A,1)          % 2001
norm(A,Inf)        % 2001
norm(A,'fro')      % približno 4473.2550
norm(A(:),Inf)     % 2000
norm(A,2)          % približno 2000.9986
```

Z merjenjem časa izvajanja ukazov se lahko prepričamo, da je računanje 2-norme veliko počasnejše od računanja ostalih norm. Če tridiagonalno matriko $\boldsymbol{A}$ podamo v razpršeni obliki (ukaz `spdiags([n-(1:n)' -2*(1:n)' n-(0:n-1)'], [-1 0 1], n, n)`), spektralne norme ne moremo izračunati z ukazom `norm`. Namesto tega lahko uporabimo funkcijo `normest`, a ta vrne le (včasih nezanesljiv) približek za 2-normo.

Naloga 3.10. Naj bo $\boldsymbol{A} \in \mathbb{C}^{n\times n}$ matrika z elementi $a_{i,j}$ in naj $|\boldsymbol{A}| \in \mathbb{R}^{n\times n}$ označuje matriko z elementi $|a_{i,j}|$.

1. Dokažite, da je $\|\boldsymbol{A}\|_2 \leq \||\boldsymbol{A}|\|_2$.
2. S primerom dokažite, da v splošnem ne velja $\|\boldsymbol{A}\|_2 = \||\boldsymbol{A}|\|_2$.

Rešitev.

1. Naj bo $\boldsymbol{x} = (x_1, x_2, \ldots, x_n) \in \mathbb{C}^n$ tak enotski vektor, da velja $\|\boldsymbol{A}\|_2 = \|\boldsymbol{A}\boldsymbol{x}\|_2$. Opazimo, da je

$$\|\boldsymbol{A}\boldsymbol{x}\|_2^2 = \sum_{i=1}^n \left|\sum_{j=1}^n a_{i,j}x_j\right|^2 \leq \sum_{i=1}^n \left(\sum_{j=1}^n |a_{i,j}|\,|x_j|\right)^2 = \||\boldsymbol{A}|\,|\boldsymbol{x}|\|_2^2,$$

kjer je $|\boldsymbol{x}| = (|x_1|, |x_2|, \ldots, |x_n|) \in \mathbb{R}^n$ enotski vektor. Zato po definiciji matrične 2-norme velja $\|\boldsymbol{A}\|_2 \leq \||\boldsymbol{A}|\|_2$.

2. Poskusimo poiskati simetrično matriko $\boldsymbol{A} \in \mathbb{R}^{2\times 2}$, za katero velja $\|\boldsymbol{A}\|_2 < \||\boldsymbol{A}|\|_2$. Pišimo
$$\boldsymbol{A} = \begin{bmatrix} a & b \\ b & c \end{bmatrix}, \quad a, b, c \in \mathbb{R},$$
in obravnavajmo normi $\|\boldsymbol{A}\|_2$ in $\||\boldsymbol{A}|\|_2$ na podlagi lastnosti, da 2-norma matrike ustreza največji singularni vrednosti matrike. Ker je
$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \begin{bmatrix} a^2+b^2 & ab+bc \\ ba+cb & b^2+c^2 \end{bmatrix}, \quad |\boldsymbol{A}|^\mathsf{T}|\boldsymbol{A}| = \begin{bmatrix} a^2+b^2 & |ab|+|bc| \\ |ba|+|cb| & b^2+c^2 \end{bmatrix},$$
je matrika $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ pri izbiri $c = -a$ diagonalna in velja $\|\boldsymbol{A}\|_2 = \sqrt{a^2+b^2}$. Po drugi strani v tem primeru $\||\boldsymbol{A}|\|_2$ ustreza korenu večje izmed rešitev λ kvadratne enačbe
$$\left(a^2+b^2-\lambda\right)^2 - (2\,|ab|)^2 = 0,$$
torej
$$\||\boldsymbol{A}|\|_2 = \sqrt{a^2+b^2+2\,|ab|}.$$
Če vzamemo na primer $a = 1$, $b = 1$, $c = -1$, dobimo matriko $\boldsymbol{A}$, za katero velja $\|\boldsymbol{A}\|_2 = \sqrt{2} < 2 = \||\boldsymbol{A}|\|_2$.

Občutljivost ali pogojenostno število $\kappa(\boldsymbol{A}) = \|\boldsymbol{A}^{-1}\|\,\|\boldsymbol{A}\|$ obrnljive matrike $\boldsymbol{A}$ je merilo, ki nam pove, kako občutljiv je problem reševanja sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. Za numerično izračunani približek $\widehat{\boldsymbol{x}}$ točne rešitve $\boldsymbol{x}$ namreč velja
$$\frac{\|\widehat{\boldsymbol{x}} - \boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \mathcal{O}(\kappa(\boldsymbol{A})u),$$
kjer je u osnovna zaokrožitvena napaka. Običajno nas zanima spektralna občutljivost $\kappa_2(\boldsymbol{A})$, ki jo dobimo z izbiro 2-norme v definiciji $\kappa(\boldsymbol{A})$, in ustreza kvocientu med največjo in najmanjšo singularno vrednostjo matrike $\boldsymbol{A}$.

Naloga 3.11. Polinom
$$p(x) = a_n x^n + a_{n-1}x^{n-1} + \ldots + a_0$$
stopnje $n \in \mathbb{N}_0$ z realnimi koeficienti $a_0, a_1, \ldots, a_n$, ki se z izbrano funkcijo f ujema v paroma različnih točkah $x_0, x_1, \ldots, x_n$, je določen z rešitvijo sistema linearnih enačb $\boldsymbol{V} \cdot \boldsymbol{a} = \boldsymbol{f}$, kjer
$$\boldsymbol{V} = \left[x_i^j\right]_{i,j=0}^n = \begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{bmatrix}$$

označuje Vandermondovo matriko, $\boldsymbol{a} = (a_0, a_1, \ldots, a_n)$ predstavlja vektor iskanih koeficientov polinoma p, vektor $\boldsymbol{f} = (f(x_0), f(x_1), \ldots, f(x_n))$ pa je sestavljen iz vrednosti funkcije f v točkah $x_0, x_1, \ldots, x_n$.

1. Komentirajte obstoj polinoma p.

2. V Matlabu izračunajte spektralne občutljivosti Vandermondovih matrik pri izbiri točk $x_k = k/n$, $k = 0, 1, \ldots, n$, za $n \in \{1, 2, \ldots, 10\}$.

3. Naj bo
$$f(x) = x^5 + x^4 + x^3 + x^2 + x + 1.$$
Polinom p stopnje 5, ki se v točkah $k/5$, $k = 0, 1, \ldots, 5$, ujema s funkcijo f, je kar funkcija f sama, saj ustreza polinomu stopnje 5. Preverite, da v Matlabu z reševanjem zgornjega sistema enačb ne dobimo povsem takega rezultata.

Rešitev.

1. Znano je, da je
$$\det(\boldsymbol{V}) = \prod_{0 \le j < k \le n} (x_k - x_j).$$
Ker so točke $x_0, x_1, \ldots, x_n$ paroma različne, je $\det(\boldsymbol{V}) \neq 0$ in sistem, ki določa p, ima enolično rešitev.

2. Vandermondovo matriko lahko v Matlabu zgeneriramo z ukazom `vander`, vendar, pozor, rezultat klica je matrika z elementi $[x_i^{n-j}]_{i,j=0}^n$. Za računanje (spektralne) občutljivosti matrike uporabimo vgrajeno funkcijo `cond`. Ugotovimo, da z večanjem n občutljivost hitro raste. Pri $n = 2$ je približno 2.6, pri $n = 5$ že približno $5.8 \cdot 10^4$, pri $n = 10$ pa kar $4.5 \cdot 10^{12}$.

3. Rešujemo sistem z Vandermondovo matriko velikosti 6×6. Uporabimo vgrajeni operator \. Rezultat bi moral biti vektor samih enic, vendar se v drugi normi od njega razlikuje za približno $2.3386 \cdot 10^{-13}$, kar ni zanemarljiva napaka.

```
X = (0:5)'/5;
V = fliplr(vander(X));
f = polyval(ones(1,6),X);
a = V\f;
norm(ones(6,1)-a)    % približno 2.3386 * 1e-13
```

Naloga 3.12. Dokažite, da je spektralna občutljivost Vandermondove matrike $\boldsymbol{V}$, ki je podana v nalogi 3.11, enaka 1, če točke $x_0, x_1, \ldots, x_n$ ustrezajo korenom enote, to je
$$x_k = e^{2\pi k i/(n+1)}, \quad k = 0, 1, \ldots, n.$$

Rešitev. Označimo $\omega = e^{2\pi i/(n+1)}$. Točka x_k ustreza vrednosti ω^k, zato je

$$\boldsymbol{V} = \begin{bmatrix} 1 & \omega^0 & \omega^0 & \dots & \omega^0 \\ 1 & \omega^1 & \omega^2 & \dots & \omega^n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega^n & \omega^{2n} & \dots & \omega^{n^2} \end{bmatrix}.$$

Oglejmo si produkt $\boldsymbol{V}$ in $\boldsymbol{V}^{\mathsf{H}}$. Diagonalni element $(\boldsymbol{V}\boldsymbol{V}^{\mathsf{H}})_{j,j}$, $j = 0, 1, \dots, n$, je enak

$$\sum_{k=0}^{n} \omega^{jk}\,\overline{\omega^{jk}} = \sum_{k=0}^{n} (\omega\,\overline{\omega})^{jk} = n + 1,$$

saj je $\overline{\omega} = \omega^{-1}$, medtem ko za vsak izvendiagonalni element velja

$$\sum_{k=0}^{n} \omega^k = \frac{1 - \omega^{n+1}}{1 - \omega} = 0,$$

saj je $\omega^{n+1} = 1$. Potemtakem je $\boldsymbol{V}\boldsymbol{V}^{\mathsf{H}} = (n+1)\boldsymbol{I}$ in $\kappa_2(\boldsymbol{V}) = 1$.

3.2. LU razcep

Osnovna numerična metoda za reševanje sistemov linearnih enačb temelji na LU razcepu. Vsako matriko $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ z obrnljivimi vodilnimi podmatrikami lahko enolično predstavimo v obliki $\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$, kjer je $\boldsymbol{L} \in \mathbb{R}^{n\times n}$ spodnje trikotna matrika z enkami na diagonali in $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ zgornje trikotna matrika z neničelnimi elementi na diagonali.

LU razcep izračunamo v $n-1$ korakih tako, da začetno matriko $\boldsymbol{A}^{(0)} = \boldsymbol{A}$ postopno preoblikujemo in jo v koraku $j \in \{1, 2, \dots, n-1\}$ zamenjamo z matriko $\boldsymbol{A}^{(j)} = \boldsymbol{L}_j\boldsymbol{A}^{(j-1)}$, ki jo dobimo s pomočjo elementarne eliminacije $\boldsymbol{L}_j = \boldsymbol{I} - \boldsymbol{l}_j\boldsymbol{e}_j^{\mathsf{T}}$. Pri tem $\boldsymbol{e}_j \in \mathbb{R}^n$ označuje standardni enotski vektor, vektor

$$\boldsymbol{l}_j = \begin{bmatrix} 0 & \dots & 0 & \frac{\alpha_{j+1}}{\alpha_j} & \dots & \frac{\alpha_n}{\alpha_j} \end{bmatrix}^{\mathsf{T}} \in \mathbb{R}^n$$

pa je določen z elementi α_i, $i = 1, 2, \dots, n$, stolpca j v matriki $\boldsymbol{A}^{(j-1)}$. Zato je

$$\boldsymbol{L}_j \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_n \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} \alpha_1 & \alpha_2 & \dots & \alpha_j & 0 & \dots & 0 \end{bmatrix}^{\mathsf{T}}$$

in $\boldsymbol{A}^{(n-1)}$ je zgornje trikotna matrika, ki ustreza matriki $\boldsymbol{U}$ v razcepu. Poleg tega velja $\boldsymbol{L}_j^{-1} = \boldsymbol{I} + \boldsymbol{l}_j\boldsymbol{e}_j^{\mathsf{T}}$ in

$$\boldsymbol{L}_1^{-1}\boldsymbol{L}_2^{-1}\cdots\boldsymbol{L}_{n-1}^{-1} = \boldsymbol{I} + \boldsymbol{l}_1\boldsymbol{e}_1^{\mathsf{T}} + \boldsymbol{l}_2\boldsymbol{e}_2^{\mathsf{T}} + \ldots + \boldsymbol{l}_{n-1}\boldsymbol{e}_{n-1}^{\mathsf{T}},$$

zato je inverz matrike $\boldsymbol{L}_{n-1}\cdots\boldsymbol{L}_2\boldsymbol{L}_1$ spodnja trikotna matrika, ki ustreza matriki $\boldsymbol{L}$. Po tem postopku LU razcep matrike izračunamo z $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ osnovnimi računskimi operacijami.

Naloga 3.13. Dana je matrika

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & 3 & -4 \\ -4 & -1 & -4 & 7 \\ 2 & 3 & 5 & -3 \\ -2 & -2 & -7 & 9 \end{bmatrix}.$$

Izračunajte LU razcep matrike $\boldsymbol{A}$.

Rešitev. Matriko $\boldsymbol{A}$ s pomočjo treh elementarnih eliminacij $\boldsymbol{L}_1$, $\boldsymbol{L}_2$, $\boldsymbol{L}_3$ preoblikujemo v zgornje trikotno matriko $\boldsymbol{U}$. Ker je matrika $\boldsymbol{L} = (\boldsymbol{L}_3\boldsymbol{L}_2\boldsymbol{L}_1)^{-1}$ spodnje trikotna in ima na diagonali enice, lahko vse bistvene elemente matrik $\boldsymbol{L}$ in $\boldsymbol{U}$ hranimo v matriki enake velikosti, kot je $\boldsymbol{A}$. Postopek je povzet v naslednjih korakih.

1. korak:

$$\begin{bmatrix} 2 & 1 & 3 & -4 \\ -4 & -1 & -4 & 7 \\ 2 & 3 & 5 & -3 \\ -2 & -2 & -7 & 9 \end{bmatrix} \overset{\boldsymbol{L}_1}{\sim} \begin{bmatrix} 2 & 1 & 3 & -4 \\ -2 & 1 & 2 & -1 \\ 1 & 2 & 2 & 1 \\ -1 & -1 & -4 & 5 \end{bmatrix}, \quad \boldsymbol{L}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

2. korak:

$$\begin{bmatrix} 2 & 1 & 3 & -4 \\ -2 & 1 & 2 & -1 \\ 1 & 2 & 2 & 1 \\ -1 & -1 & -4 & 5 \end{bmatrix} \overset{\boldsymbol{L}_2}{\sim} \begin{bmatrix} 2 & 1 & 3 & -4 \\ -2 & 1 & 2 & -1 \\ 1 & 2 & -2 & 3 \\ -1 & -1 & -2 & 4 \end{bmatrix}, \quad \boldsymbol{L}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & -2 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}$$

3. korak:

$$\begin{bmatrix} 2 & 1 & 3 & -4 \\ -2 & 1 & 2 & -1 \\ 1 & 2 & -2 & 3 \\ -1 & -1 & -2 & 4 \end{bmatrix} \overset{\boldsymbol{L}_3}{\sim} \begin{bmatrix} 2 & 1 & 3 & -4 \\ -2 & 1 & 2 & -1 \\ 1 & 2 & -2 & 3 \\ -1 & -1 & 1 & 1 \end{bmatrix}, \quad \boldsymbol{L}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}$$

Iskani matriki $\boldsymbol{L}$ in $\boldsymbol{U}$ sta dani z

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & -1 & 1 & 1 \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} 2 & 1 & 3 & -4 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix}.$$

Naloga 3.14. V Matlabu pripravite funkcijo, ki izračuna LU razcep dane matrike. Preizkusite funkcijo na matrikah

$$\boldsymbol{A}_1 = \begin{bmatrix} 2 & -1 & 1 & 4 \\ 4 & -3 & 7 & 14 \\ 0 & -3 & 18 & 19 \\ 6 & -2 & 7 & 14 \end{bmatrix}, \qquad \boldsymbol{A}_2 = \begin{bmatrix} 2 & -2 & -4 & 4 \\ 4 & -3 & -6 & 14 \\ -8 & 3 & 6 & 19 \\ 6 & -2 & 7 & 14 \end{bmatrix}.$$

Rešitev. LU razcep $\boldsymbol{A} = \boldsymbol{LU}$ matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ opravimo v $n-1$ korakih. Pri tem sproti

- dopolnjujemo matriko $\boldsymbol{L}$, ki jo na začetku z ukazom `eye` nastavimo na identično matriko, in
- spreminjamo matriko $\boldsymbol{A}$, ki jo na koncu z ukazom `triu` oklestimo v zgornjo trikotno matriko $\boldsymbol{U}$.

V koraku $j \in \{1, 2, \ldots, n-1\}$ popravimo vrstice matrike $\boldsymbol{A}$ z indeksi i od $j+1$ do n. Element $\boldsymbol{A}(i,j)$ delimo z diagonalnim elementom $\boldsymbol{A}(j,j)$ in dobimo $\boldsymbol{L}(i,j)$. Elemente v vrstici i matrike $\boldsymbol{A}$, ki se nahajajo v stolpcih z indeksi k od $j+1$ do n (to je desno od elementa $\boldsymbol{A}(i,j)$), popravimo tako, da od njih odštejemo produkt $\boldsymbol{L}(i,j)$ in $\boldsymbol{A}(j,k)$.

```
n = size(A,1);
L = eye(n);
for j = 1:n-1
    for i = j+1:n
        L(i,j) = A(i,j)/A(j,j);
        for k = j+1:n
            A(i,k) = A(i,k) - L(i,j)*A(j,k);
        end
    end
end
U = triu(A);
```

Pri izvedbi zgornjega postopka na matriki $\boldsymbol{A}_1$ dobimo LU razcep $\boldsymbol{A}_1 = \boldsymbol{L}_1\boldsymbol{U}_1$, ki je določen z matrikama

$$\boldsymbol{L}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 3 & 1 & 0 \\ 3 & -1 & 3 & 1 \end{bmatrix}, \qquad \boldsymbol{U}_1 = \begin{bmatrix} 2 & -1 & 1 & 4 \\ 0 & -1 & 5 & 6 \\ 0 & 0 & 3 & 1 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

Pri računanju LU razcepa $\boldsymbol{A}_2 = \boldsymbol{L}_2\boldsymbol{U}_2$ matrike $\boldsymbol{A}_2$ pa pride do zapleta, saj dobimo

$$\boldsymbol{L}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ -4 & -5 & 1 & 0 \\ 3 & 4 & \infty & 1 \end{bmatrix}, \qquad \boldsymbol{U}_2 = \begin{bmatrix} 2 & -2 & -4 & 4 \\ 0 & 1 & 2 & 6 \\ 0 & 0 & 0 & 65 \\ 0 & 0 & 0 & -\infty \end{bmatrix}.$$

Težava je, da smo v predzadnjem koraku na diagonali dobili ničlo, po deljenju z njo pa v matriki $\boldsymbol{L}$ element ∞. Očitno torej LU razcep matrike $\boldsymbol{A}_2$ ne obstaja.

Naloga 3.15. Poiščite tako spodnje trikotno matriko $\boldsymbol{L}$ in zgornje trikotno matriko $\boldsymbol{U}$, da za neobrnljivo matriko

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

velja $\boldsymbol{A} = \boldsymbol{LU}$. Ali je ta razcep enoličen? Zakaj to ni v nasprotju s trditvijo o enoličnosti in obstoju LU razcepa matrike?

Rešitev. Po običajnem postopku LU razcepa dobimo

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

V matriki $\boldsymbol{L}$ lahko element na mestu $(3,3)$ spremenimo na poljubno vrednost pa produkt $\boldsymbol{L}$ in $\boldsymbol{U}$ ostane enak $\boldsymbol{A}$. To ni v nasprotju s trditvijo o obstoju in enoličnosti LU razcepa, saj je element matrike $\boldsymbol{U}$ na mestu $(3,3)$ enak 0 in $\boldsymbol{U}$ ne ustreza lastnostim zgornje trikotne matrike iz LU razcepa. Prav tako $\boldsymbol{L}$ ustreza lastnostim spodnje trikotne matrike iz LU razcepa le, če je element na mestu $(3,3)$ enak 1.

Naj bosta $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ matrika in $\boldsymbol{b} \in \mathbb{R}^n$ vektor, ki določata sistem linearnih enačb $\boldsymbol{Ax} = \boldsymbol{b}$. Denimo, da obstaja LU razcep $\boldsymbol{A} = \boldsymbol{LU}$ matrike $\boldsymbol{A}$. Tedaj za sistem velja $\boldsymbol{L}(\boldsymbol{Ux}) = \boldsymbol{b}$ in rešitev $\boldsymbol{x} \in \mathbb{R}^n$ lahko poiščemo z reševanjem sistemov $\boldsymbol{Ly} = \boldsymbol{b}$ in $\boldsymbol{Ux} = \boldsymbol{y}$. Ker v obeh sistemih nastopata trikotni matriki, je računanje njunih rešitev enostavno in zahteva skupno le $2n^2 + n$ računskih operacij.

Naloga 3.16. Naj bo $\boldsymbol{A} \in \mathbb{R}^{4\times 4}$ matrika iz naloge 3.13 in

$$\boldsymbol{b} = \begin{bmatrix} 8 & -14 & 7 & -16 \end{bmatrix}^{\mathsf{T}}.$$

Z uporabo LU razcepa matrike $\boldsymbol{A}$ izračunajte rešitev sistema $\boldsymbol{Ax} = \boldsymbol{b}$.

Rešitev. Sistem $\boldsymbol{Ax} = \boldsymbol{b}$ rešimo z reševanjem sistemov $\boldsymbol{Ly} = \boldsymbol{b}$ in $\boldsymbol{Ux} = \boldsymbol{y}$ za matriki $\boldsymbol{L}$ in $\boldsymbol{U}$, podani v rešitvi naloge 3.13. Prvi sistem rešimo s premo substitucijo (od zgoraj navzdol):

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ -1 & -1 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 8 \\ -14 \\ 7 \\ -16 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 2 \\ -5 \\ -1 \end{bmatrix};$$

drugega pa z obratno substitucijo (od spodaj navzgor):

$$\begin{bmatrix} 2 & 1 & 3 & -4 \\ 0 & 1 & 2 & -1 \\ 0 & 0 & -2 & 3 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 8 \\ 2 \\ -5 \\ -1 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ -1 \\ 1 \\ -1 \end{bmatrix}.$$

Sistem $\boldsymbol{Ax} = \boldsymbol{b}$ ima enolično rešitev natanko tedaj, ko je matrika $\boldsymbol{A}$ obrnljiva. Da poiščemo rešitev sistema tudi v primerih, ko niso vse vodilne podmatrike matrike $\boldsymbol{A}$ obrnljive, med računanjem LU razcepa pivotiramo. Pri delnem pivotiranju pred izvedbo vsake elementarne eliminacije poskrbimo, da se v stolpcu, ki ga eliminiramo, na diagonali nahaja po absolutni vrednosti največji element med tistimi, ki ležijo na diagonali oziroma pod njo. To dosežemo z zamenjavo vrstic, čemur pravimo pivotiranje in kar predstavimo s permutacijsko matriko $\boldsymbol{P}$, ki se odraža v končnem razcepu. LU razcep matrike $\boldsymbol{A}$ z delnim pivotiranjem ima namreč obliko $\boldsymbol{PA} = \boldsymbol{LU}$, kjer sta $\boldsymbol{L}$ in $\boldsymbol{U}$ matriki z enakimi lastnostmi kot pri običajnem LU razcepu.

Naloga 3.17. Naj bo

$$\boldsymbol{A} = \begin{bmatrix} 2 & 1 & -2 & 1 \\ 2 & 1 & -4 & 2 \\ 3 & -2 & 3 & -1 \\ -1 & 3 & -1 & 1 \end{bmatrix}.$$

Poiščite LU razcep matrike $\boldsymbol{A}$ z delnim pivotiranjem in ga uporabite pri izračunu determinante $\boldsymbol{A}$.

Rešitev. Razcep lahko shematično predstavimo podobno kot običajni LU razcep v nalogi 3.13, le da tu po potrebi na vsakem koraku postopka pivotiramo.

1. korak:

$$\begin{bmatrix} 2 & 1 & -2 & 1 \\ 2 & 1 & -4 & 2 \\ 3 & -2 & 3 & -1 \\ -1 & 3 & -1 & 1 \end{bmatrix} \underset{1\leftrightarrow 3}{\overset{\boldsymbol{P}_1}{\sim}} \begin{bmatrix} 3 & -2 & 3 & -1 \\ 2 & 1 & -4 & 2 \\ 2 & 1 & -2 & 1 \\ -1 & 3 & -1 & 1 \end{bmatrix} \overset{\boldsymbol{L}_1}{\sim} \begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ \frac{2}{3} & \frac{7}{3} & -4 & \frac{5}{3} \\ -\frac{1}{3} & \frac{7}{3} & 0 & \frac{2}{3} \end{bmatrix}$$

2. korak:

$$\begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ \frac{2}{3} & \frac{7}{3} & -4 & \frac{5}{3} \\ -\frac{1}{3} & \frac{7}{3} & 0 & \frac{2}{3} \end{bmatrix} \underset{2\leftrightarrow 2}{\overset{\boldsymbol{P}_2}{\sim}} \begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ \frac{2}{3} & \frac{7}{3} & -4 & \frac{5}{3} \\ -\frac{1}{3} & \frac{7}{3} & 0 & \frac{2}{3} \end{bmatrix} \overset{\boldsymbol{L}_2}{\sim} \begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ \frac{2}{3} & 1 & 2 & -1 \\ -\frac{1}{3} & 1 & 6 & -2 \end{bmatrix}$$

3. korak:

$$\begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ \frac{2}{3} & 1 & 2 & -1 \\ -\frac{1}{3} & 1 & 6 & -2 \end{bmatrix} \underset{3\leftrightarrow 4}{\overset{\boldsymbol{P}_3}{\sim}} \begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ -\frac{1}{3} & 1 & 6 & -2 \\ \frac{2}{3} & 1 & 2 & -1 \end{bmatrix} \overset{\boldsymbol{L}_3}{\sim} \begin{bmatrix} 3 & -2 & 3 & -1 \\ \frac{2}{3} & \frac{7}{3} & -6 & \frac{8}{3} \\ -\frac{1}{3} & 1 & 6 & -2 \\ \frac{2}{3} & 1 & \frac{1}{3} & -\frac{1}{3} \end{bmatrix}$$

Na ta način matriko $\boldsymbol{A}$ postopno transformiramo v zgornje trikotno matriko $\boldsymbol{U}$. Velja

$$\boldsymbol{L}_3\boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{P}_2\boldsymbol{L}_1\boldsymbol{P}_1\boldsymbol{A} = \boldsymbol{U}.$$

Če uvedemo matriki $\widetilde{\boldsymbol{L}}_2 = \boldsymbol{P}_3\boldsymbol{L}_2\boldsymbol{P}_3^{-1}$ in $\widetilde{\boldsymbol{L}}_1 = \boldsymbol{P}_3\boldsymbol{P}_2\boldsymbol{L}_1\boldsymbol{P}_2^{-1}\boldsymbol{P}_3^{-1}$, lahko zvezo zapišemo kot

$$\boldsymbol{L}_3\widetilde{\boldsymbol{L}}_2\widetilde{\boldsymbol{L}}_1\boldsymbol{P}_3\boldsymbol{P}_2\boldsymbol{P}_1\boldsymbol{A} = \boldsymbol{U}.$$

Permutacijsko matriko $\boldsymbol{P} = \boldsymbol{P}_3\boldsymbol{P}_2\boldsymbol{P}_1$ dobimo iz identitete z uporabo enakih zamenjav vrstic, kot smo jih izvedli pri pivotiranju:

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \underset{1\leftrightarrow 3}{\overset{\boldsymbol{P}_1}{\sim}} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \underset{3\leftrightarrow 4}{\overset{\boldsymbol{P}_3}{\sim}} \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Matriki $\boldsymbol{L} = (\boldsymbol{L}_3\widetilde{\boldsymbol{L}}_2\widetilde{\boldsymbol{L}}_1)^{-1}$ in $\boldsymbol{U}$ razberemo iz izračunane tabele. Ključno je, da smo v postopku izvedli zamenjavo celotnih vrstic tabele in tako izračunali elemente matrik $\widetilde{\boldsymbol{L}}_1^{-1}$, $\widetilde{\boldsymbol{L}}_2^{-1}$, $\boldsymbol{L}_3^{-1}$, ki se nahajajo pod diagonalo. Razcep $\boldsymbol{PA} = \boldsymbol{LU}$ matrike $\boldsymbol{A}$ je torej določen z matrikami

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \frac{2}{3} & 1 & 0 & 0 \\ -\frac{1}{3} & 1 & 1 & 0 \\ \frac{2}{3} & 1 & \frac{1}{3} & 1 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} 3 & -2 & 3 & -1 \\ 0 & \frac{7}{3} & -6 & \frac{8}{3} \\ 0 & 0 & 6 & -2 \\ 0 & 0 & 0 & -\frac{1}{3} \end{bmatrix}, \quad \boldsymbol{P} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}.$$

Determinanto matrike $\boldsymbol{A}$ izračunamo na podlagi lastnosti

$$\det(\boldsymbol{P})\det(\boldsymbol{A}) = \det(\boldsymbol{PA}) = \det(\boldsymbol{LU}) = \det(\boldsymbol{L})\det(\boldsymbol{U}).$$

Ker je $\det(\boldsymbol{P}) = 1$, $\det(\boldsymbol{L}) = 1$ in $\det(\boldsymbol{U}) = -14$, je $\det(\boldsymbol{A}) = -14$.

Naloga 3.18. V Matlabu lahko LU razcep matrike z delnim pivotiranjem izračunamo z vgrajeno funkcijo `lu`. Preizkusite jo na matrikah $\boldsymbol{A}_1$ in $\boldsymbol{A}_2$ iz naloge 3.14.

Rešitev. Iz dokumentacije funkcije `lu` (`help lu`) je moč razbrati, da jo lahko uporabimo z različnimi vhodnimi in izhodnimi podatki. Izberemo način, ki sprejme matriko, vrne pa seznam treh matrik.

```
[L,U,P] = lu(A);
```

Rezultat klica je LU razcep $\boldsymbol{PA} = \boldsymbol{LU}$ matrike $\boldsymbol{A}$ z delnim pivotiranjem. Matrika $\boldsymbol{L}$ je spodnje trikotna z enicami na diagonali, matrika $\boldsymbol{U}$ je zgornje trikotna, matrika $\boldsymbol{P}$ pa je permutacijska in povzema postopek pivotiranja pri izračunu razcepa. Pri izvedbi LU razcepa matrike $\boldsymbol{A}_1$ dobimo drugačen razcep od izračunanega po metodi brez pivotiranja. Smiseln razcep dobimo tudi pri matriki $\boldsymbol{A}_2$, na kateri je postopek LU razcepa brez pivotiranja odpovedal.

Naloga 3.19. V Matlabu sestavite funkcijo, ki poišče rešitev sistema $\boldsymbol{Ax} = \boldsymbol{b}$ za obrnljivo matriko $\boldsymbol{A}$ in vektor $\boldsymbol{b}$. V funkciji najprej z vgrajenim ukazom izračunajte LU razcep matrike $\boldsymbol{A}$ z delnim pivotiranjem, nato pa $\boldsymbol{x}$ določite s pomočjo premih in obratnih substitucij. Preverite, da se rezultati implementirane funkcije ujemajo z rezultati vgrajene funkcije `linsolve` oziroma \ pri naključno izbranih matrikah in vektorjih, ki jih generirate z ukazom `rand`.

Rešitev. Če LU razcep $\boldsymbol{PA} = \boldsymbol{LU}$ matrike $\boldsymbol{A}$ izračunamo z vgrajenim ukazom `lu`, lahko rešitev sistema $\boldsymbol{Ax} = \boldsymbol{b}$ dobimo z reševanjem dveh sistemov: $\boldsymbol{Ly} = \boldsymbol{Pb}$ in $\boldsymbol{Ux} = \boldsymbol{y}$. V teh dveh sistemih nastopata spodnje in zgornje trikotna matrika, zato ju lahko rešimo s premo in obratno substitucijo.

Pri reševanju prvega sistema je treba določiti vse komponente vektorja $\boldsymbol{y}$, ki jih indeksiramo z indeksi i od 1 do n. Na začetku $\boldsymbol{y}$ definiramo kot produkt permutacijske matrike $\boldsymbol{P}$ in vektorja $\boldsymbol{b}$. Nato vsak element $\boldsymbol{y}(i)$ popravimo tako, da od njega odštejemo produkte $\boldsymbol{L}(i,k)$ in $\boldsymbol{y}(k)$ za vse predhodne indekse k od 1 do $i-1$. Ker je matrika $\boldsymbol{L}$ spodnje trikotna in ima na diagonali enice, je končni vektor $\boldsymbol{y}$ res rešitev sistema $\boldsymbol{Ly} = \boldsymbol{Pb}$.

```
y = P*b;
for i = 2:n
    for k = 1:i-1
        y(i) = y(i) - L(i,k)*y(k);
    end
end
```

Pri reševanju sistema $\boldsymbol{Ux} = \boldsymbol{y}$ uberemo obratni pristop. Na začetku za vektor $\boldsymbol{x}$ vzamemo kar $\boldsymbol{y}$, nato pa komponente vektorja $\boldsymbol{x}$, ki jih indeksiramo z i, popravljamo po vrsti od n proti 1. V koraku i od komponente $\boldsymbol{x}(i)$ odštejemo produkte $\boldsymbol{U}(i,k)$ in $\boldsymbol{x}(k)$ za vse indekse k od $i+1$ do n, na koncu pa vrednost $\boldsymbol{x}(i)$ še delimo z diagonalnim elementom $\boldsymbol{U}(i,i)$, ki tu (v splošnem) ni enak 1. Rezultat tega postopka je rešitev $\boldsymbol{x}$ sistema $\boldsymbol{Ux} = \boldsymbol{y}$, ki predstavlja tudi rešitev prvotnega sistema $\boldsymbol{Ax} = \boldsymbol{y}$.

```
x = y;
for i = n:-1:1
    for k = i+1:n
        x(i) = x(i) - U(i,k)*x(k);
    end
    x(i) = x(i)/U(i,i);
end
```

Naloga 3.20. Naj bo $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $n \in \mathbb{N}$, obrnljiva matrika, za katero poznamo LU razcep z delnim pivotiranjem. Naj bo $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ obrnljiva matrika, ki jo dobimo tako, da v matriki $\boldsymbol{A}$ zamenjamo enega izmed stolpcev z vektorjem $\boldsymbol{b} \in \mathbb{R}^n$. Naj bo $\boldsymbol{c} \in \mathbb{R}^n$. Izpeljite čim bolj učinkovit postopek za izračun rešitve $\boldsymbol{x}$ sistema $\boldsymbol{Bx} = \boldsymbol{c}$. Natančno preštejte število osnovnih računskih operacij, ki jih postopek zahteva.

Rešitev. Opazimo, da lahko matriko $\boldsymbol{B}$ predstavimo kot $\boldsymbol{B} = \boldsymbol{AE}$, kjer matrika $\boldsymbol{E} \in \mathbb{R}^{n\times n}$ ustreza identični matriki v vseh stolpcih, razen v stolpcu, ki ima enak indeks kot stolpec, v katerem se matriki $\boldsymbol{A}$ in $\boldsymbol{B}$ razlikujeta. Naj bo ta stolpec matrike $\boldsymbol{E}$ označen z $\boldsymbol{e} \in \mathbb{R}^n$. Zanj velja $\boldsymbol{Ae} = \boldsymbol{b}$ in ga lahko ob znanem LU razcepu matrike $\boldsymbol{A}$ izračunamo s premo in obratno substitucijo ($2n^2 + n$ osnovnih računskih operacij).

Naj bo s $\boldsymbol{PA} = \boldsymbol{LU}$ predstavljen LU razcep matrike $\boldsymbol{A}$ z delnim pivotiranjem. Sistem $\boldsymbol{Bx} = \boldsymbol{c}$ lahko sedaj rešimo v treh korakih. Najprej poiščemo rešitev $\boldsymbol{z} \in \mathbb{R}^n$ sistema $\boldsymbol{Lz} = \boldsymbol{Pc}$ s premo substitucijo (n^2 osnovnih računskih operacij). Nato poiščemo rešitev $\boldsymbol{y} \in \mathbb{R}^n$ sistema $\boldsymbol{Uy} = \boldsymbol{z}$ z obratno substitucijo ($n^2 + n$ osnovnih računskih operacij). Nazadnje poiščemo še rešitev $\boldsymbol{x} \in \mathbb{R}^n$ sistema $\boldsymbol{Ex} = \boldsymbol{y}$, ki ustreza rešitvi sistema $\boldsymbol{Bx} = \boldsymbol{c}$.

Naj bo $k \in \mathbb{N}$ indeks stolpca $\boldsymbol{e}$ v matriki $\boldsymbol{E}$. Sistem $\boldsymbol{Ex} = \boldsymbol{y}$ lahko rešimo tako, da najprej z eno osnovno računsko operacijo izračunamo komponento $\boldsymbol{x}$ z indeksom k. Nato z obratno substitucijo z $2(k-1)$ osnovnimi računskimi operacijami izračunamo komponente $\boldsymbol{x}$ z indeksi od 1 do $k-1$ in s premo substitucijo z $2(n-k)$ osnovnimi računskimi operacijami še komponente $\boldsymbol{x}$ z indeksi od $k+1$ do n.

Celoten postopek izračuna $\boldsymbol{x}$ torej zahteva $4n^2 + 4n - 1$ osnovnih računskih operacij.

Namesto delnega pivotiranja lahko pri računanju LU razcepa uporabimo kompletno pivotiranje. Pri tem v trenutnem koraku razcepa pivotni element iščemo v celotnem preostanku matrike (in ne samo v trenutnem stolpcu pod diagonalo kot pri delnem pivotiranju). Pivotni element spravimo v zgornji levi rob preostanka matrike s permutacijo vrstic in stolpcev, zato je končni LU razcep oblike $\boldsymbol{PAQ} = \boldsymbol{LU}$. Pri tem sta $\boldsymbol{P}$ in $\boldsymbol{Q}$ permutacijski matriki, ki povzemata zamenjavo vrstic in stolpcev, matriki $\boldsymbol{L}$ in $\boldsymbol{U}$ pa imata enake lastnosti kot matriki, dobljeni pri običajnem LU razcepu. Razcep s kompletnim pivotiranjem daje teoretično zagotovilo za obratno stabilnost postopka, ki je pri delnem pivotiranju nimamo, a zahteva precej več primerjanj, kot jih imamo pri izvedbi z delnim pivotiranjem, zato se v praksi uporablja slednji.

Naloga 3.21. Naj bo

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & -3 \\ 4 & 2 & -6 \\ -3 & 3 & 3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 8 \\ 14 \\ 0 \end{bmatrix}.$$

Izračunajte LU razcep matrike $\boldsymbol{A}$ s kompletnim pivotiranjem in poiščite rešitev sistema $\boldsymbol{Ax} = \boldsymbol{b}$.

Rešitev. Razcep izračunamo po naslednjih korakih.

1. korak:

$$\begin{bmatrix} 1 & 2 & -3 \\ 4 & 2 & -6 \\ -3 & 3 & 3 \end{bmatrix} \underset{(1,1)\leftrightarrow(2,3)}{\overset{\boldsymbol{P}_1,\,\boldsymbol{Q}_1}{\sim}} \begin{bmatrix} -6 & 2 & 4 \\ -3 & 2 & 1 \\ 3 & 3 & -3 \end{bmatrix} \overset{\boldsymbol{L}_1}{\sim} \begin{bmatrix} -6 & 2 & 4 \\ \frac{1}{2} & 1 & -1 \\ -\frac{1}{2} & 4 & -1 \end{bmatrix}$$

2. korak:

$$\begin{bmatrix} -6 & 2 & 4 \\ \frac{1}{2} & 1 & -1 \\ -\frac{1}{2} & 4 & -1 \end{bmatrix} \underset{(2,2)\leftrightarrow(3,2)}{\overset{\boldsymbol{P}_2,\,\boldsymbol{Q}_2}{\sim}} \begin{bmatrix} -6 & 2 & 4 \\ -\frac{1}{2} & 4 & -1 \\ \frac{1}{2} & 1 & -1 \end{bmatrix} \overset{\boldsymbol{L}_2}{\sim} \begin{bmatrix} -6 & 2 & 4 \\ -\frac{1}{2} & 4 & -1 \\ \frac{1}{2} & \frac{1}{4} & -\frac{3}{4} \end{bmatrix}$$

S tem postopkom smo matriko $\boldsymbol{A}$ transformirali v obrnljivo zgornje trikotno matriko $\boldsymbol{U} = \boldsymbol{L}_2\boldsymbol{P}_2\boldsymbol{L}_1\boldsymbol{P}_1\boldsymbol{A}\boldsymbol{Q}_1\boldsymbol{Q}_2$, na podlagi česar dobimo razcep $\boldsymbol{PAQ} = \boldsymbol{LU}$. Trikotni matriki, ki ju razberemo iz izračunane tabele, sta podani z

$$\boldsymbol{L} = \boldsymbol{P}_2\boldsymbol{L}_1^{-1}\boldsymbol{P}_2^{-1}\boldsymbol{L}_2^{-1} = \begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} -6 & 2 & 4 \\ 0 & 4 & -1 \\ 0 & 0 & -\frac{3}{4} \end{bmatrix},$$

permutacijsko matriki, ki ju lahko dobimo z ustreznimi permutacijami vrstic oziroma stolpcev identite, pa z

$$\boldsymbol{P} = \boldsymbol{P}_2\boldsymbol{P}_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \quad \boldsymbol{Q} = \boldsymbol{Q}_1\boldsymbol{Q}_2 = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

Rešitev sistema $\boldsymbol{Ax} = \boldsymbol{b}$ določimo z reševanjem sistema $\boldsymbol{Ly} = \boldsymbol{Pb}$:

$$\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 0 \\ 8 \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 7 \\ -\frac{3}{4} \end{bmatrix};$$

in sistema $\boldsymbol{Uz} = \boldsymbol{y}$:

$$\begin{bmatrix} -6 & 2 & 4 \\ 0 & 4 & -1 \\ 0 & 0 & -\frac{3}{4} \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 14 \\ 7 \\ -\frac{3}{4} \end{bmatrix} \quad \Rightarrow \quad \boldsymbol{z} = \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} -1 \\ 2 \\ 1 \end{bmatrix};$$

ter permutacijo $\boldsymbol{x} = \boldsymbol{Qz} = (1, 2, -1)$.

Pri iskanju rešitve $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{b}$ sistema $\boldsymbol{Ax} = \boldsymbol{b}$ ne računanamo inverza $\boldsymbol{A}^{-1}$ matrike $\boldsymbol{A}$, saj je to računsko zahtevnejše in manj stabilno od določanja $\boldsymbol{x}$ s premimi in obratnimi substitucijami po opravljenem LU razcepu. Vseeno pa obstajajo problemi, kjer je inverz matrike treba izračunati eksplicitno. Pri tem si lahko ponovno pomagamo z LU razcepom. Pomemben pa je tudi razmislek, kako izračunati inverz razširjene ali popravljene matrike na podlagi predhodno izračunanega inverza začetne matrike.

Naloga 3.22. Opišite, kako lahko z uporabo LU razcepa izračunamo inverz matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Koliko operacij v odvisnosti od n je za to potrebnih? Postopek demonstrirajte na matriki

$$\boldsymbol{A} = \begin{bmatrix} 2 & 3 & 1 \\ 4 & 8 & 3 \\ 0 & -2 & 3 \end{bmatrix}.$$

Rešitev. Naj bo $\boldsymbol{X} = \boldsymbol{A}^{-1}$ inverz matrike $\boldsymbol{A}$. Ker je $\boldsymbol{AX} = \boldsymbol{I}$, lahko $\boldsymbol{X}$ določimo z reševanjem sistemov $\boldsymbol{A}\boldsymbol{x}_j = \boldsymbol{e}_j$, $j = 1, 2, \ldots, n$, kjer $\boldsymbol{x}_j$ označuje j-ti stolpec matrike $\boldsymbol{X}$, $\boldsymbol{e}_j$ pa j-ti standardni enotski vektor velikosti n. Najprej izračunamo LU razcep matrike $\boldsymbol{A}$, za kar je potrebnih $\frac{2}{3}n^3 + \mathcal{O}(n^2)$ operacij. Nato s premimi in obratnimi substitucijami rešimo vsakega izmed n sistemov, kar zahtevata $n(2n^2 + n))$ operacij. Celoten postopek je torej zahtevnosti $\frac{8}{3}n^3 + \mathcal{O}(n^2)$.

Poiščimo inverz dane matrike $\boldsymbol{A}$. Najprej izračunamo LU razcep $\boldsymbol{PA} = \boldsymbol{LU}$ matrike $\boldsymbol{A}$ z delnim pivotiranjem. Določen je z matrikami

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} 4 & 8 & 3 \\ 0 & -2 & 3 \\ 0 & 0 & -2 \end{bmatrix}, \quad \boldsymbol{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

Z reševanjem sistemov $\boldsymbol{L}\boldsymbol{y}_j = \boldsymbol{P}\boldsymbol{e}_j$ in $\boldsymbol{U}\boldsymbol{x}_j = \boldsymbol{y}_j$, $j = 1, 2, \ldots, n$, dobimo inverz

$$\boldsymbol{X} = \boldsymbol{A}^{-1} = \begin{bmatrix} \frac{15}{8} & -\frac{11}{16} & \frac{1}{16} \\ -\frac{3}{4} & \frac{3}{8} & -\frac{1}{8} \\ -\frac{1}{2} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}.$$

Naloga 3.23. Dana je obrnljiva matrika $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ ter njen inverz $\boldsymbol{A}^{-1}$. Sestavite učinkovit postopek za izračun inverza matrike

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{u} \\ \boldsymbol{v}^\mathsf{T} & \alpha \end{bmatrix}, \qquad \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n, \ \alpha \in \mathbb{R},$$

pod pogojem, da je $\boldsymbol{B}$ obrnljiva. Koliko operacij je potrebnih za izvedbo postopka in kakšen je potreben in zadosten pogoj za obrnljivost $\boldsymbol{B}$?

Rešitev. Zapišimo inverz $\boldsymbol{B}^{-1}$ matrike $\boldsymbol{B}$ v obliki

$$\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{x} \\ \boldsymbol{y}^\mathsf{T} & \beta \end{bmatrix}, \qquad \boldsymbol{C} \in \mathbb{R}^{n\times n}, \ \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n, \ \beta \in \mathbb{R}.$$

Iz

$$\boldsymbol{B}\boldsymbol{B}^{-1} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{u} \\ \boldsymbol{v}^\mathsf{T} & \alpha \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{x} \\ \boldsymbol{y}^\mathsf{T} & \beta \end{bmatrix} = \boldsymbol{I}_{n+1}$$

sledi

$$\boldsymbol{AC} + \boldsymbol{u}\boldsymbol{y}^\mathsf{T} = \boldsymbol{I}_n, \quad \boldsymbol{A}\boldsymbol{x} + \beta\boldsymbol{u} = 0, \quad \boldsymbol{v}^\mathsf{T}\boldsymbol{C} + \alpha\boldsymbol{y}^\mathsf{T} = 0, \quad \boldsymbol{v}^\mathsf{T}\boldsymbol{x} + \alpha\beta = 1.$$

S pomočjo $\boldsymbol{A}^{-1}$ lahko od tod izrazimo

$$\boldsymbol{C} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{y}^{\mathsf{T}}, \quad \boldsymbol{x} = -\beta\boldsymbol{A}^{-1}\boldsymbol{u}, \quad \beta = \frac{1}{\alpha - \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}, \quad \boldsymbol{y}^{\mathsf{T}} = -\beta\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}.$$

Torej lahko izračun inverza matrike $\boldsymbol{B}$ izvedemo v korakih:

- $\boldsymbol{z} = \boldsymbol{A}^{-1}\boldsymbol{u}$ ($2n^2 - n$ operacij),
- $\beta = 1/(\alpha - \boldsymbol{v}^{\mathsf{T}}\boldsymbol{z})$ ($2n + 1$ operacij),
- $\boldsymbol{y}^{\mathsf{T}} = -\beta\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}$ ($2n^2$ operacij),
- $\boldsymbol{x} = -\beta\boldsymbol{z}$ (n operacij),
- $\boldsymbol{C} = \boldsymbol{A}^{-1} - \boldsymbol{z}\boldsymbol{y}^{\mathsf{T}}$ ($2n^2$ operacij).

Celoten postopek izračuna torej zahteva $6n^2 + 2n + 1$ operacij. Iz postopka je razvidno, da inverz matrike $\boldsymbol{B}$ obstaja, če je $\alpha \neq \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}$. Velja tudi obratno. Če je $\alpha = \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}$, potem je

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{u} \\ \boldsymbol{v}^{\mathsf{T}} & \alpha \end{bmatrix} \begin{bmatrix} \boldsymbol{A}^{-1}\boldsymbol{u} \\ -1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ 0 \end{bmatrix},$$

kar pomeni, da v jedru matrike $\boldsymbol{B}$ obstaja neničeln vektor, zato ni obrnljiva.

Naloga 3.24. Naj bo $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ obrnljiva matrika. Dokažite, da je za vektorja $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ matrika $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$ obrnljiva natanko tedaj, ko je $1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$, ter da v tem primeru velja Sherman–Morrisonova formula

$$\left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right)^{-1} = \boldsymbol{A}^{-1} - \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}}{1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

Rešitev. Iz naloge 3.23 sledi, da je matrika

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{u} \\ \boldsymbol{v}^{\mathsf{T}} & -1 \end{bmatrix}$$

obrnljiva natanko tedaj, ko je $1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$. Opazimo, da je

$$\begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{u} \\ \boldsymbol{0}^{\mathsf{T}} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{A} & \boldsymbol{u} \\ \boldsymbol{v}^{\mathsf{T}} & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{v}^{\mathsf{T}} & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}} & \boldsymbol{0} \\ \boldsymbol{0}^{\mathsf{T}} & -1 \end{bmatrix},$$

pri čemer sta matriki, s katerimi smo množili matriko $\boldsymbol{B}$, obrnljivi:

$$\begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{u} \\ \boldsymbol{0}^{\mathsf{T}} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{u} \\ \boldsymbol{0}^{\mathsf{T}} & 1 \end{bmatrix}, \quad \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ \boldsymbol{v}^{\mathsf{T}} & 1 \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{v}^{\mathsf{T}} & 1 \end{bmatrix}.$$

Sledi, da je tudi matrika $\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}$ obrnljiva natanko tedaj, ko je $1 + \boldsymbol{v}^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{u} \neq 0$. Za inverz glede na oznake iz naloge 3.23 velja

$$\begin{bmatrix} \left(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^{\mathsf{T}}\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0}^{\mathsf{T}} & -1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{I}_n & \boldsymbol{0} \\ -\boldsymbol{v}^{\mathsf{T}} & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{C} & \boldsymbol{x} \\ \boldsymbol{y}^{\mathsf{T}} & \beta \end{bmatrix} \begin{bmatrix} \boldsymbol{I}_n & -\boldsymbol{u} \\ \boldsymbol{0}^{\mathsf{T}} & 1 \end{bmatrix} = \begin{bmatrix} \boldsymbol{C} & \boldsymbol{0} \\ \boldsymbol{0}^{\mathsf{T}} & -1 \end{bmatrix},$$

kar dokazuje Sherman–Morrisonovo formulo.

Naloga 3.25. Denimo, da je znan postopek reševanja sistema $\boldsymbol{Ax} = \boldsymbol{b}$ s pomočjo LU razcepa $\boldsymbol{PA} = \boldsymbol{LU}$ obrnljive matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Kako bi učinkovito izračunali rešitev sistema $\boldsymbol{By} = \boldsymbol{b}$ z obrnljivo matriko $\boldsymbol{B}$, ki se od matrike $\boldsymbol{A}$ razlikuje v samo enem elementu?

Rešitev. Recimo, da se matriki $\boldsymbol{A}$ in $\boldsymbol{B}$ razlikujeta v elementih na mestu (i, j), ki ju označimo z $a_{i,j}$ in $b_{i,j}$. Potem je

$$\boldsymbol{B} = \boldsymbol{A} + (b_{i,j} - a_{i,j})\boldsymbol{e}_i\boldsymbol{e}_j^\mathsf{T},$$

pri čemer sta $\boldsymbol{e}_i, \boldsymbol{e}_j \in \mathbb{R}^{n\times n}$ standardna enotska vektorja. Rešitev sistema $\boldsymbol{By} = \boldsymbol{b}$ lahko izrazimo s pomočjo Sherman–Morrisonove formule iz naloge 3.24 kot

$$\boldsymbol{y} = \boldsymbol{B}^{-1}\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{b} - \frac{\boldsymbol{e}_j^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{b}}{1 + \boldsymbol{e}_j^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{e}_i}\boldsymbol{A}^{-1}\boldsymbol{e}_i = \boldsymbol{x} - \frac{\boldsymbol{e}_j^\mathsf{T}\boldsymbol{x}}{1 + \boldsymbol{e}_j^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{e}_i}\boldsymbol{A}^{-1}\boldsymbol{e}_i.$$

Za izračun $\boldsymbol{y}$ torej zadošča poiskati rešitev sistema $\boldsymbol{Az} = \boldsymbol{e}_i$, ki jo lahko dobimo z $2n^2 + n$ operacijami preme in obratne substitucije ob pomoči LU razcepa matrike $\boldsymbol{A}$. Rešitev $\boldsymbol{y}$ potem ustreza vektorju $\boldsymbol{x} - (\boldsymbol{x}_j/(1 + \boldsymbol{z}_j))\boldsymbol{z}$, ki ga izračunamo z dodatnimi $2n + 2$ operacijami.

3.3. Reševanje sistemov posebne oblike

Sistemi linearnih enačb, ki jih srečamo v praktičnih primerih, imajo pogosto posebne lastnosti, ki se odražajo v matriki $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ sistema $\boldsymbol{Ax} = \boldsymbol{b}$. Take lastnosti lahko včasih izkoristimo za zmanjšanje računske zahtevnosti reševanja sistema ali pa so v pomoč pri analizi stabilnosti.

Če je matrika $\boldsymbol{A}$ simetrično pozitivno definitna ($\boldsymbol{A} = \boldsymbol{A}^\mathsf{T}$ in za vsak $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ velja $\boldsymbol{x}^\mathsf{T}\boldsymbol{Ax} > 0$), pri reševanju sistema namesto LU razcepa uporabljamo razcep Choleskega, ki je zahtevnosti $\frac{1}{3}n^3 + \mathcal{O}(n^2)$. Pri tem razcepu matriko $\boldsymbol{A}$ predstavimo kot produkt spodnje trikotne matrike $\boldsymbol{V}$ in njene transponirane matrike: $\boldsymbol{A} = \boldsymbol{VV}^\mathsf{T}$. Matrika $\boldsymbol{V}$, ki jo imenujemo faktor Choleskega, ima na diagonali same pozitivne elemente. Razcep Choleskega izračunamo s postopkom, ki ga lahko v Matlabu implementiramo s spodnjimi ukazi.

```
n = size(A,1);
V = zeros(n);
for j = 1:n
    V(j,j) = sqrt(A(j,j) - V(j,1:j-1)*V(j,1:j-1)');
    for i = j+1:n
        V(i,j) = (A(i,j)-V(i,1:j-1)*V(j,1:j-1)')/V(j,j);
    end
end
```

Naloga 3.26. Dana je matrika

$$\boldsymbol{A} = \begin{bmatrix} 4 & 6 & 2 & -4 \\ 6 & 18 & 0 & 3 \\ 2 & 0 & 3 & -4 \\ -4 & 3 & -4 & \alpha \end{bmatrix}, \qquad \alpha \in \mathbb{R}.$$

1. Določite vsa števila α, za katere je matrika $\boldsymbol{A}$ simetrična pozitivno definitna.

2. Naj bo $\alpha = 23$. Uporabite razcep Choleskega za reševanje sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, kjer je $\boldsymbol{b} = (6, 15, 2, 1)$.

Rešitev. Za dano matriko $\boldsymbol{A}$ je faktor Choleskega podan z

$$\boldsymbol{V} = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 \\ 1 & -1 & 1 & 0 \\ -2 & 3 & 1 & \sqrt{\alpha - 14} \end{bmatrix}$$

in obstaja natanko tedaj, ko je $\alpha > 14$.

1. Matrika $\boldsymbol{A}$ je simetrična pozitivno definitna natanko tedaj, ko obstaja razcep Choleskega, torej natanko tedaj, ko je $\alpha > 14$.

2. Rešitev sistema poiščemo z reševanjem dveh preprostejših sistemov: $\boldsymbol{V}\boldsymbol{y} = \boldsymbol{b}$ in $\boldsymbol{V}^\mathsf{T}\boldsymbol{x} = \boldsymbol{y}$. Pri prvem uporabimo premo substitucijo, pri drugem pa obratno substitucijo. Izračunamo $\boldsymbol{y} = (3, 2, 1, 0)$ in $\boldsymbol{x} = (-1/2, 1, 1, 0)$.

Naloga 3.27. Naj bo

$$\boldsymbol{B} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \\ \boldsymbol{a}^\mathsf{T} & \alpha \end{bmatrix}, \quad \boldsymbol{A} \in \mathbb{R}^{n\times n},\ \boldsymbol{a} \in \mathbb{R}^n,\ \alpha \in \mathbb{R}.$$

Kakšni so minimalni zadostni pogoji, da je matrika $\boldsymbol{B}$ simetrična pozitivno definitna? Privzemite te pogoje in zapišite razcep Choleskega matrike $\boldsymbol{B}$.

Rešitev. Matrika $\boldsymbol{B}$ je simetrično pozitivno definitna natanko tedaj, ko obstaja razcep Choleskega

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{a} \\ \boldsymbol{a}^\mathsf{T} & \alpha \end{bmatrix} = \begin{bmatrix} \boldsymbol{V} & \boldsymbol{0} \\ \boldsymbol{b}^\mathsf{T} & \beta \end{bmatrix} \begin{bmatrix} \boldsymbol{V}^\mathsf{T} & \boldsymbol{b} \\ \boldsymbol{0}^\mathsf{T} & \beta \end{bmatrix}, \quad \boldsymbol{V} \in \mathbb{R}^{n\times n},\ \boldsymbol{b} \in \mathbb{R}^n,\ \beta \in \mathbb{R},$$

kjer je $\boldsymbol{V}$ spodnje trikotna matrika s pozitivnimi diagonalnimi elementi in $\beta > 0$. Iz tega sledi, da je $\boldsymbol{A} = \boldsymbol{V}\boldsymbol{V}^\mathsf{T}$ in matrika $\boldsymbol{V}$ (če obstaja) je faktor Choleskega matrike $\boldsymbol{A}$. Torej je simetričnost in pozitivna definitnost matrike $\boldsymbol{A}$ potreben pogoj za to, da ima enako lastnost tudi matrika $\boldsymbol{B}$. Poleg tega nastavek implicira, da je $\boldsymbol{V}\boldsymbol{b} = \boldsymbol{a}$ in $\boldsymbol{b}^\mathsf{T}\boldsymbol{b} + \beta^2 = \alpha$ oziroma $\boldsymbol{b} = \boldsymbol{V}^{-1}\boldsymbol{a}$ (matrika $\boldsymbol{V}$ je obrnljiva) in $\beta = \sqrt{\alpha - \boldsymbol{a}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{a}}$. Slednje pomeni, da mora za pozitivno definitnost matrike $\boldsymbol{B}$ veljati še $\alpha > \boldsymbol{a}^\mathsf{T}\boldsymbol{A}^{-1}\boldsymbol{a}$.

Naloga 3.28. Dani sta matriki $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$. Matrika $\boldsymbol{B}$ je simetrična pozitivno definitna. Sestavite učinkovit postopek za izračun sledi matrike $\boldsymbol{A}^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{A}$ in preštejte število osnovnih računskih operacij, ki jih postopek terja.

Rešitev. Naj bo i-ti stolpec matrike $\boldsymbol{A}$ označen z $\boldsymbol{a}_i$, $i = 1, 2, \ldots, n$. Potem je

$$(\boldsymbol{A}^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{A})_{i,j} = \boldsymbol{a}_i^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_j$$

in

$$\operatorname{sled}(\boldsymbol{A}^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{A}) = \boldsymbol{a}_1^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_1 + \boldsymbol{a}_2^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_2 + \ldots + \boldsymbol{a}_n^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_n.$$

Za učinkovit izračun sledi torej zadošča premisliti, kako učinkovito izračunati vrednosti $\boldsymbol{a}_i^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_i$, $i = 1, 2, \ldots, n$. Ker je $\boldsymbol{B}$ simetrična pozitivno definitna matrika, jo lahko zapišemo v obliki $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{V}^\mathsf{T}$, kjer je $\boldsymbol{V}$ faktor Choleskega. Iz razcepa sledi

$$\boldsymbol{a}_i^\mathsf{T}\boldsymbol{B}^{-1}\boldsymbol{a}_i = \boldsymbol{a}_i^\mathsf{T}(\boldsymbol{V}\boldsymbol{V}^\mathsf{T})^{-1}\boldsymbol{a}_i = (\boldsymbol{V}^{-1}\boldsymbol{a}_i)^\mathsf{T}(\boldsymbol{V}^{-1}\boldsymbol{a}_i) = \left\|\boldsymbol{V}^{-1}\boldsymbol{a}_i\right\|_2^2.$$

Izračun matrike $\boldsymbol{V}$ zahteva $\frac{1}{3}n^3 + \mathcal{O}(n^2)$ operacij. Vrednost $\boldsymbol{V}^{-1}\boldsymbol{a}_i$ izračunamo s premimi substitucijami z $n^2 + n$ operacijami. Za izračun kvadrata druge norme je potrebnih še dodatnih $2n - 1$ operacij. Celoten postopek izračuna sledi torej terja

$$\left(\frac{1}{3}n^3 + \mathcal{O}(n^2)\right) + n\left(\left(n^2 + n\right) + (2n - 1)\right) + (n - 1) = \frac{4}{3}n^3 + \mathcal{O}(n^2)$$

osnovnih računskih operacij.

Za matriko pravimo, da je strogo diagonalno dominantna, če je v vsaki njeni vrstici absolutna vrednost diagonalnega elementa večja od vsote absolutnih vrednosti izvendiagonalnih elementov. Pri izračunu LU razcepa tovrstnih matrik pivotiranje ni potrebno. Če je matrika ob tem še tridiagonalna, pa lahko LU razcep bistveno poenostavimo in skrajšamo njegovo izvajanje.

Naloga 3.29. Dokažite, da za vsako strogo diagonalno dominantno matriko obstaja LU razcep (brez pivotiranja).

Rešitev. Preverimo, da je po vrsticah strogo diagonalno dominantna matrika $\boldsymbol{A}$ velikosti $n \times n$ z elementi $a_{i,j}$ nesingularna. Denimo nasprotno, da je $\boldsymbol{A}\boldsymbol{u} = \boldsymbol{0}$ za nek neničeln vektor $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$. Naj bo i indeks po absolutni vrednosti največjega elementa $\boldsymbol{u}$. Ker je

$$\sum_{j=1}^{n} a_{i,j}u_j = 0 \quad \Leftrightarrow \quad a_{i,i}u_i = -\sum_{\substack{j=1\\ j\neq i}}^{n} a_{i,j}u_j \quad \Leftrightarrow \quad a_{i,i} = -\sum_{\substack{j=1\\ j\neq i}}^{n} \frac{u_j}{u_i}a_{i,j},$$

iz izbire indeksa i sledi

$$|a_{i,i}| \leq \sum_{\substack{j=1\\ j\neq i}}^{n} |a_{i,j}|,$$

kar je v protislovju s strogo diagonalno dominantnostjo matrike $\boldsymbol{A}$. Ker je vsaka vodilna podmatrika strogo diagonalno dominantne matrike tudi strogo diagonalno dominantna, so vse vodilne podmatrike $\boldsymbol{A}$ nesingularne, kar zagotavlja obstoj LU razcepa (brez pivotiranja).

Naloga 3.30. Dana je tridiagonalna matrika

$$\boldsymbol{A} = \begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & a_{n-1} & b_{n-1} \\ & & & c_{n-1} & a_n \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

1. Predpostavite, da obstaja LU razcep matrike $\boldsymbol{A}$ (brez pivotiranja) in ga poenostavite. Preštejte število operacij, potrebnih za njegovo izvedbo.

2. Rešitev sistema $\boldsymbol{Ax} = \boldsymbol{z}$ po izračunu razcepa $\boldsymbol{A} = \boldsymbol{LU}$ določimo z uporabo preme in obratne substitucije. Primerjajte število operacij v splošnem postopku s številom operacij, ki so potrebne za izračun rešitve sistema v primeru, ko je $\boldsymbol{A}$ tridiagonalna matrika.

3. Implementirajte postopek za reševanje sistema s tridiagonalno matriko v Matlabu. To je tako imenovani Thomasov postopek. Generirajte naključne podatke za $n = 10^4$ tako, da bo matrika $\boldsymbol{A}$ strogo diagonalno dominantna. Nato izmerite čas, ki ga za reševanje sistema porabi Thomasov postopek, in ga primerjajte s časom, ki je potreben za reševanje sistema z vgrajeno funkcijo.

Rešitev.

1. Pri LU razcepu tridiagonalne matrike sta matriki $\boldsymbol{L}$ in $\boldsymbol{U}$ oblike

$$\boldsymbol{L} = \begin{bmatrix} 1 & & & & \\ l_1 & 1 & & & \\ & \ddots & \ddots & & \\ & & l_{n-2} & 1 & \\ & & & l_{n-1} & 1 \end{bmatrix}, \qquad \boldsymbol{U} = \begin{bmatrix} u_1 & b_1 & & & \\ & u_2 & b_2 & & \\ & & \ddots & \ddots & \\ & & & u_{n-1} & b_{n-1} \\ & & & & u_n \end{bmatrix}.$$

   Velja $u_1 = a_1$, vrednosti l_k in u_{k+1} za $k = 1, 2, \ldots, n-1$ pa se izražajo kot

$$l_k = \frac{c_k}{u_k}, \qquad u_{k+1} = a_{k+1} - l_k b_k.$$

   Za izračun LU razcepa je torej potrebnih vsega $3n - 3$ operacij.

2. Sistem rešujemo s premo in obratno substitucijo. Namesto n^2 operacij je pri premi substituciji potrebnih le $2n-2$ operacij. Podobno je pri obratni substituciji namesto n^2+n operacij potrebnih le $3n-2$ operacij. Celoten postopek reševanja skupaj z LU razcepom torej zahteva $8n - 7$ operacij, kar je za dva reda manj kot pri LU razcepu splošne matrike.

3. Funkcija, ki izračuna rešitev sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$ sprejme obliko $\boldsymbol{A}$ v obliki treh seznamov: seznam $\boldsymbol{a}$ vsebuje elemente na diagonali matrike $\boldsymbol{A}$ in je dolžine n, seznama $\boldsymbol{b}$ in $\boldsymbol{c}$ pa vsebujeta elemente na naddiagonali oziroma poddiagonali matrike $\boldsymbol{A}$ ter sta dolžine $n-1$. Na ta način matriko $\boldsymbol{A}$ namesto z n^2 elementi predstavimo le s seznami v skupni dolžini $3n-2$. Izhodni podatek funkcije je rešitev sistema $\boldsymbol{x}$, pomožna izhodna podatka pa sta še seznam $\boldsymbol{\ell}$ dolžine $n-1$, ki določa matriko $\boldsymbol{L}$, in seznam $\boldsymbol{u}$ dolžine n, ki določa matriko $\boldsymbol{U}$.

```
function [x,l,u] = thomas(a,b,c,z)

n = length(a);
u = zeros(n,1);
l = zeros(n-1,1);

% LU razcep
u(1) = a(1);
for k = 1:n-1
    l(k) = c(k)/u(k);
    u(k+1) = a(k+1) - l(k)*b(k);
end

% prema substitucija
y = z;
for k = 2:n
    y(k) = z(k) - l(k-1)*y(k-1);
end

% obratna substitucija
x = y;
x(n) = y(n)/u(n);
for k = n-1:-1:1
    x(k) = (y(k) - b(k)*x(k+1))/u(k);
end

end
```

Primerjava reševanja sistema z matriko velikosti $n = 10^4$ z uporabo Thomasovega postopka in vgrajene funkcije (`linsolve` oziroma operator \) pokaže, da je naša funkcija približno 500-krat hitrejša.

```
n = 1e4;
a = rand(n,1) + 2;
b = rand(n-1,1);
c = rand(n-1,1);
z = rand(n,1);
```

```
tic; x1 = thomas(a,b,c,z); t1 = toc;

A = diag(a) + diag(b,1) + diag(c,-1);
tic; x2 = A\z; t2 = toc;

t2/t1; % priblizno 500
```

Pivotna rast g pri LU razcepu matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ z elementi $a_{i,j}$ je podana kot

$$g = \frac{\max_{i,j=1,\ldots,n} |u_{i,j}|}{\max_{i,j=1,\ldots,n} |a_{i,j}|},$$

kjer so $u_{i,j}$ elementi obrnljive zgornje trikotne matrike $\boldsymbol{U}$, ki nastopa v razcepu. S pivotno rastjo si lahko pomagamo pri analizi obratne stabilnosti postopka reševanja sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ z uporabo LU razcepa s pivotiranjem. Če je vrednost g majhna, je reševanje sistema obratno stabilno, saj je izračunani $\widehat{\boldsymbol{x}}$, ki ga dobimo namesto $\boldsymbol{x}$, točna rešitev sistema $(\boldsymbol{A}+\boldsymbol{\Delta A})\widehat{\boldsymbol{x}} = \boldsymbol{b}$, kjer je $\boldsymbol{\Delta A}$ matrika, za katero velja

$$\|\boldsymbol{\Delta A}\|_\infty \leq 3nu \|\boldsymbol{L}\|_\infty \|\boldsymbol{U}\|_\infty \leq 3gn^3u \|\boldsymbol{A}\|_\infty .$$

Pri tem je u osnovna zaokrožitvena napaka aritmetike, v kateri računamo. Pivotna rast g je pri LU razcepu z delnim pivotiranjem v splošnem navzgor omejena z 2^{n-1}, zato postopek v teoriji ni obratno stabilen. Pri matrikah s posebno strukturo lahko oceno za pivotno rast izboljšamo.

Naloga 3.31. Matrika $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ z elementi $a_{i,j}$ je zgornja Hessenbergova matrika, če velja velja $a_{i,j} = 0$ za $i > j+1$. Dokažite, da je pivotna rast zgornje Hessenbergove matrike pri LU razcepu z delnim pivotiranjem omejena z n.

Rešitev. Naj bo $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ zgornja Hessenbergova matrika. Zaradi posebne oblike te matrike imamo na vsakem koraku LU razcepa z delnim pivotiranjem dve možnosti za pivot (ohranimo obstoječe stanje ali zamenjamo trenutno in naslednjo vrstico). Naj bo $M = \max_{i,j=1,\ldots,n} |a_{i,j}|$. V prvem koraku postopka je situacija naslednja.

- Če je $|a_{1,1}| \geq |a_{2,1}|$, pivotiranje ni potrebno in novi elementi $a_{2,k}^{(1)}$ na mestih $(2,k)$, $k = 2,3,\ldots,n$, so izračunani kot

  $$a_{2,k}^{(1)} = a_{2,k} - \ell_{2,1}a_{1,k}, \qquad \ell_{2,1} = \frac{a_{2,1}}{a_{1,1}}.$$

  Ker je $|\ell_{2,1}| \leq 1$, je $|a_{2,k}^{(1)}| \leq 2M$.

- Če je $|a_{1,1}| < |a_{2,1}|$, zamenjamo prvo in drugo vrstico in tedaj velja

  $$a_{2,k}^{(1)} = a_{1,k} - \ell_{2,1}a_{2,k}, \qquad \ell_{2,1} = \frac{a_{1,1}}{a_{2,1}}.$$

Ker je $|\ell_{2,1}| < 1$, je tudi v tem primeru $|a_{2,k}^{(1)}| \leq 2M$.

Zaradi lastnosti matrike $\boldsymbol{A}$ preostale vrstice (od tretje dalje) v prvem koraku LU razcepa ostanejo nespremenjene. Recimo, da smo že opravili $r-1$ korakov postopka in predpostavimo, da je $|a_{r,k}^{(r-1)}| \leq rM$, $k = r, \ldots, n$. Ocenimo nove elemente $a_{r+1,k}^{(k)}$ na mestih $(r+1, k)$, $k = r+1, r+2 \ldots, n$.

- Če je $|a_{r,r}^{(1)}| \geq |a_{r+1,r+1}|$, pivotiranje ni potrebno in za

$$a_{r+1,k}^{(r)} = a_{r+1,k} - \ell_{r+1,r} a_{r,k}^{(r-1)}, \qquad \ell_{r+1,r} = \frac{a_{r+1,r}}{a_{r,r}^{(r-1)}},$$

  po predpostavki velja $|a_{r+1,k}^{(r)}| \leq M + rM = (r+1)M$.

- Če je $|a_{r,r}^{(1)}| < |a_{r+1,r+1}|$, pa zamenjamo vrstici z indeksoma r in $r+1$ ter izračunamo

$$a_{r+1,j}^{(r)} = a_{r,k}^{(r-1)} - \ell_{r+1,r} a_{r+1,k}, \qquad \ell_{r+1,r} = \frac{a_{r,r}^{(r-1)}}{a_{r+1,r}}.$$

  Tudi v tem primeru je po predpostavki $|a_{r+1,k}^{(r)}| \leq rM + M = (r+1)M$.

LU razcep izračunamo v $n-1$ korakih. Za element $a_{n,n}^{(n-1)}$ velja $|a_{n,n}^{(n-1)}| \leq nM$. Pokazali smo torej, da so vsi elementi v matriki $\boldsymbol{U}$ manjši od nM, zato je $g \leq n$.

Naloga 3.32. Preverite, da je za zgornjo Hessenbergovo matriko

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 & 1 \\ -1 & 1 & 0 & \ldots & 0 & 1 \\ 0 & -1 & 1 & \ddots & \vdots & 1 \\ \vdots & \ddots & \ddots & \ddots & 0 & \vdots \\ 0 & \ldots & 0 & -1 & 1 & 1 \\ 0 & 0 & \ldots & 0 & -1 & 1 \end{bmatrix} \in \mathbb{R}^{n \times n}$$

pivotna rast g enaka n.

Rešitev. Pri LU razcepu matrike $\boldsymbol{A}$ v nobenem koraku ni potrebe po pivotiranju. Po standardnem postopku izračunamo LU razcep $\boldsymbol{A} = \boldsymbol{LU}$, ki ga določata matriki

$$\boldsymbol{L} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 0 \\ -1 & 1 & \ddots & \vdots & \vdots \\ 0 & -1 & \ddots & 0 & 0 \\ \vdots & \ddots & \ddots & 1 & 0 \\ 0 & \ldots & 0 & -1 & 1 \end{bmatrix}, \quad \boldsymbol{U} = \begin{bmatrix} 1 & 0 & \ldots & 0 & 1 \\ 0 & 1 & \ddots & \vdots & 2 \\ 0 & 0 & \ddots & 0 & 3 \\ \vdots & \ddots & \ddots & 1 & \vdots \\ 0 & \ldots & 0 & 0 & n \end{bmatrix}.$$

Po definiciji pivotne rasti je $g = n$.

4. Sistemi nelinearnih enačb

Sistem n nelinearnih enačb z n neznankami je podan v obliki $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$, pri čemer je $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^n$ preslikava, $\boldsymbol{0} \in \mathbb{R}^n$ vektor samih ničel in $\boldsymbol{x} \in \mathbb{R}^n$ vektor, ki predstavlja rešitev sistema. K reševanju sistema nelinearnih enačb lahko pristopimo na podobne načine kot k reševanju nelinearnih enačb.

4.1. Jacobijeva iteracija

Ideja osnovne metode za reševanje sistema $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ je enaka kot pri navadni iteraciji. Sistem prepišemo v ekvivalentno obliko $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$, kjer je $\boldsymbol{g} : \mathbb{R}^n \to \mathbb{R}^n$ iteracijska preslikava. Rešitev iščemo z vektorjem začetnih približkov $\boldsymbol{x}^{(0)} \in \mathbb{R}^n$ tako, da v vsakem koraku izračunamo nov približek

$$\boldsymbol{x}^{(r+1)} = \boldsymbol{g}\left(\boldsymbol{x}^{(r)}\right), \quad r = 0, 1, \ldots.$$

Ta postopek imenujemo Jacobijeva iteracija. Običajno jo lahko pospešimo tako, da v tekočem koraku iteracije $r+1$ uporabimo že izračunane koordinate $\boldsymbol{x}_j^{(r+1)}$, $j = 1, 2, \ldots, i-1$, novega približka:

$$\boldsymbol{x}_i^{(r+1)} = \boldsymbol{g}_i\left(\boldsymbol{x}_1^{(r+1)}, \ldots, \boldsymbol{x}_{i-1}^{(r+1)}, \boldsymbol{x}_i^{(r)}, \ldots, \boldsymbol{x}_n^{(r)}\right), \quad i = 1, 2, \ldots, n,$$

pri čemer je z $\boldsymbol{g}_i$ označena komponenta preslikave $\boldsymbol{g}$ z indeksom i. Ta postopek se imenuje Seidlova iteracija.

Naloga 4.1. Dan je sistem

$$x = \sin\left(\frac{2x - y}{4}\right), \quad y = \cos\left(\frac{x + 2y}{4}\right)$$

dveh nelinearnih enačb s spremenljivkama x in y v iteracijski obliki. Pri začetnem približku $(x_0, y_0) = (2\pi, 0)$ napravite dva koraka Jacobijeve in Seidlove iteracije. S pomočjo Matlaba napravite še 13 korakov. Primerjajte približke s točko $(-0.43752408, 0.93632532)$, ki je natančen približek rešitve sistema.

Rešitev. Pri Jacobijevi iteraciji za prva približka (x_1, y_1) in (x_2, y_2) dobimo

$$x_1 = \sin\left(\tfrac{2x_0-y_0}{4}\right) = \sin(\pi) = 0, \qquad y_1 = \cos\left(\tfrac{x_0+2y_0}{4}\right) = \cos\left(\tfrac{\pi}{2}\right) = 0,$$
$$x_2 = \sin\left(\tfrac{2x_1-y_1}{4}\right) = \sin(0) = 0, \qquad y_2 = \cos\left(\tfrac{x_1+2y_1}{4}\right) = \cos(0) = 1,$$

pri Seidlovi iteraciji pa

$$x_1 = \sin\left(\tfrac{2x_0-y_0}{4}\right) = \sin(\pi) = 0, \qquad y_1 = \cos\left(\tfrac{x_1+2y_0}{4}\right) = \cos(0) = 1,$$
$$x_2 = \sin\left(\tfrac{2x_1-y_1}{4}\right) = -\sin\left(\tfrac{1}{4}\right), \qquad y_2 = \cos\left(\tfrac{x_2+2y_1}{4}\right) = \cos\left(\frac{-\sin(\frac{1}{4})+2}{4}\right).$$

Naslednjih 13 približkov izračunamo v Matlabu.

```
g1 = @(x)sin((2*x(1)-x(2))/4);
g2 = @(x)cos((x(1)+2*x(2))/4);
x0 = [2*pi; 0];
XJ = x0;
XS = x0;
for r = 1:15
    XJ(1,r+1) = g1(XJ(:,r));
    XJ(2,r+1) = g2(XJ(:,r));
    XS(1,r+1) = g1(XS(:,r));
    XS(2,r+1) = g2([XS(1,r+1); XS(2,r)]);
end
```

Rezultata prikazuje tabela 4.1. Napaka je izračunana kot norma razlike med trenutnim približkom in podanim natančnim približkom. Razvidno je, da so približki Seidlove iteracije boljši od približkov Jacobijeve iteracije.

Konvergenca zaporedja približkov pri Jacobijevi iteraciji je odvisna od lastnosti preslikave $\boldsymbol{g}$. Denimo, da je $\boldsymbol{g}$ na omejenem in zaprtem območju $\Omega \subset \mathbb{R}^n$ zvezno odvedljiva ter da za vsak $\boldsymbol{x} \in \Omega$ velja $\boldsymbol{g}(\boldsymbol{x}) \in \Omega$ in $\|\boldsymbol{J}_{\boldsymbol{g}}(\boldsymbol{x})\| \leq m < 1$. Pri tem je $\boldsymbol{J}_{\boldsymbol{g}}$ Jacobijeva matrika preslikave $\boldsymbol{g}$, $\|\cdot\|$ pa označuje neko matrično normo. Potem je $\boldsymbol{g}$ skrčitev z Lipschitzovo konstanto m in iteracijsko zaporedje po Banachovem skrčitvenem načelu konvergira za vsak začetni približek iz Ω.

Naloga 4.2. Dokažite, da Jacobijeva in Seidlova iteracija za sistem iz naloge 4.1 konvergirata za vsak začetni približek.

Rešitev. Preslikava Jacobijeve iteracije

$$\boldsymbol{g}_J(x, y) = \left(\sin\left(\frac{2x-y}{4}\right), \cos\left(\frac{x+2y}{4}\right)\right)$$

korak	Jacobijeva iteracija			Seidlova iteracija		
r	x_r	y_r	napaka	x_r	y_r	napaka
1	0	0	$1.0 \cdot 10^{0}$	0	1	$4.4 \cdot 10^{-1}$
2	0	1	$4.4 \cdot 10^{-1}$	-0.247404	0.905539	$1.9 \cdot 10^{-1}$
3	-0.247404	0.877583	$2.0 \cdot 10^{-1}$	-0.342979	0.933399	$9.5 \cdot 10^{-2}$
4	-0.336406	0.929795	$1.0 \cdot 10^{-1}$	-0.393871	0.932965	$4.4 \cdot 10^{-2}$
5	-0.390019	0.928369	$4.8 \cdot 10^{-2}$	-0.417032	0.935111	$2.1 \cdot 10^{-2}$
6	-0.414234	0.933523	$2.3 \cdot 10^{-2}$	-0.428013	0.935702	$9.5 \cdot 10^{-3}$
7	-0.42639	0.934764	$1.1 \cdot 10^{-2}$	-0.433103	0.936046	$4.4 \cdot 10^{-3}$
8	-0.432159	0.93562	$5.4 \cdot 10^{-3}$	-0.435472	0.936194	$2.1 \cdot 10^{-3}$
9	-0.434952	0.935978	$2.6 \cdot 10^{-3}$	-0.436572	0.936265	$9.5 \cdot 10^{-4}$
10	-0.436289	0.93616	$1.2 \cdot 10^{-3}$	-0.437082	0.936297	$4.4 \cdot 10^{-4}$
11	-0.436932	0.936246	$6.0 \cdot 10^{-4}$	-0.437319	0.936312	$2.1 \cdot 10^{-4}$
12	-0.43724	0.936287	$2.9 \cdot 10^{-4}$	-0.437429	0.936319	$9.5 \cdot 10^{-5}$
13	-0.437388	0.936307	$1.4 \cdot 10^{-4}$	-0.43748	0.936323	$4.4 \cdot 10^{-5}$
14	-0.437459	0.936317	$6.6 \cdot 10^{-5}$	-0.437504	0.936324	$2.1 \cdot 10^{-5}$
15	-0.437493	0.936321	$3.2 \cdot 10^{-5}$	-0.437515	0.936325	$9.5 \cdot 10^{-6}$

TABELA 4.1: *Jacobijeva in Seidlova iteracija iz naloge 4.1.*

vsak $(x, y) \in \mathbb{R}^2$ preslika v $[-1, 1] \times [-1, 1]$, zato lahko analizo $\boldsymbol{g}_J$ omejimo na to območje. Jacobijeva matrika $\boldsymbol{J}_{\boldsymbol{g}_J}$ preslikave $\boldsymbol{g}_J$ je podana z

$$\boldsymbol{J}_{\boldsymbol{g}_J}(x, y) = \begin{bmatrix} \frac{1}{2}\cos\left(\frac{2x-y}{4}\right) & -\frac{1}{4}\cos\left(\frac{2x-y}{4}\right) \\ -\frac{1}{4}\sin\left(\frac{x+2y}{4}\right) & -\frac{1}{2}\sin\left(\frac{x+2y}{4}\right) \end{bmatrix}.$$

Za vsak (x, y) velja $\|\boldsymbol{J}_{\boldsymbol{g}_J}(x, y)\|_\infty \leq \frac{3}{4} < 1$, zato je $\boldsymbol{g}_J$ skrčitev in iteracijsko zaporedje po Banachovem skrčitvenem načelu konvergira za vsak začetni približek. Enak sklep velja za preslikavo

$$\boldsymbol{g}_S(x, y) = \left(\sin\left(\frac{2x-y}{4}\right), \cos\left(\frac{\sin\left(\frac{2x-y}{4}\right) + 2y}{4}\right)\right)$$

Seidlove iteracije, saj je $\|\boldsymbol{J}_{\boldsymbol{g}_S}(x, y)\|_\infty \leq \frac{3}{4} < 1$.

Z metodami za reševanje sistemov nelinearnih enačb lahko rešujemo tudi sisteme linearnih enačb. To pride v poštev predvsem takrat, kadar je sistem linearnih enačb prevelik, da bi mu bili kos z direktno metodo (na primer z LU razcepom).

Naloga 4.3. Zapišite iteracijsko funkcijo za Jacobijevo in Seidlovo iteracijo pri reševanju sistema linearnih enačb $\boldsymbol{Ax} = \boldsymbol{b}$. Pri tem predpostavite, da so vsi elementi na diagonali matrike $\boldsymbol{A}$ neničelni.

Rešitev. Pri Jacobijevi iteraciji iteracijsko funkcijo zasnujemo tako, da v i-ti enačbi sistema $\boldsymbol{Ax} = \boldsymbol{b}$ izrazimo i-to komponento $\boldsymbol{x}_i$ vektorja $\boldsymbol{x}$. V vektorskem smislu

to pomeni, da sistem zapišemo kot $\boldsymbol{D}\boldsymbol{x} = -(\boldsymbol{A} - \boldsymbol{D})\boldsymbol{x} + \boldsymbol{b}$, kjer je $\boldsymbol{D}$ diagonalna matrika, ki ima enako diagonalo kot $\boldsymbol{A}$. Ker so vsi elementi na diagonali $\boldsymbol{A}$ različni od nič, je $\boldsymbol{D}$ obrnljiva in preurejeno enačbo lahko z leve pomnožimo z inverzom $\boldsymbol{D}$. Na ta način dobimo iteracijsko funkcijo

$$\boldsymbol{g}_J(\boldsymbol{x}) = -\boldsymbol{D}^{-1}(\boldsymbol{A} - \boldsymbol{D})\boldsymbol{x} + \boldsymbol{D}^{-1}\boldsymbol{b}.$$

Pri Seidlovi iteraciji v trenutnem koraku pri računanju posamezne komponente novega približka upoštevamo že izračunane približke v trenutnem koraku in sistem zapišemo kot $\boldsymbol{L}\boldsymbol{x} = -(\boldsymbol{A} - \boldsymbol{L})\boldsymbol{x} + \boldsymbol{b}$, kjer je $\boldsymbol{L}$ spodnje trikotna matrika, ki ima na in pod diagonalo enake elemente kot $\boldsymbol{A}$. Za iteracijsko funkcijo $\boldsymbol{g}_S$ zato vzamemo

$$\boldsymbol{g}_S(\boldsymbol{x}) = -\boldsymbol{L}^{-1}(\boldsymbol{A} - \boldsymbol{L})\boldsymbol{x} + \boldsymbol{L}^{-1}\boldsymbol{b}.$$

Zaradi predpostavke o neničelnosti diagonalnih elementov $\boldsymbol{A}$ je tudi matrika $\boldsymbol{L}$ obrnljiva. Pri izvedbi Seidlove iteracije ne računamo inverza matrike $\boldsymbol{L}$, pač pa korak iteracije izvedemo s premimi substitucijami.

Naloga 4.4. Dan je sistem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ linearnih enačb s strogo diagonalno dominantno matriko $\boldsymbol{A}$. Dokažite, da Jacobijeva in Seidlova iteracija konvergirata za vsak začetni približek.

Rešitev. Jacobijeva matrika iteracijske preslikave $\boldsymbol{g}_J$, ki je definirana v nalogi 4.3, je enaka $-\boldsymbol{D}^{-1}(\boldsymbol{A}-\boldsymbol{D})$. Označimo elemente matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ z $a_{i,j}$, $i, j = 1, 2, \ldots, n$. Iz stroge diagonalne dominantnosti matrike $\boldsymbol{A}$ sledi, da je $a_{i,i} \neq 0$ za vsak i, poleg tega pa za vsak $\boldsymbol{x} \in \mathbb{R}^n$ tudi

$$\|\boldsymbol{J}_{\boldsymbol{g}_J}(\boldsymbol{x})\|_\infty = \left\|-\boldsymbol{D}^{-1}(\boldsymbol{A} - \boldsymbol{D})\right\|_\infty = \max_{i=1,2,\ldots,n} \frac{\sum_{j=1, j\neq i}^{n} |a_{i,j}|}{|a_{i,i}|} < 1,$$

kar pomeni, da je $\boldsymbol{g}_J$ skrčitev.

Pri Seidlovi iteraciji Jacobijeva matrika $\boldsymbol{J}_{\boldsymbol{g}_S}(\boldsymbol{x})$ iteracijske preslikave $\boldsymbol{g}_S$ iz naloge 4.3 za vsak $\boldsymbol{x}$ ustreza matriki $\boldsymbol{J}_{\boldsymbol{g}_S} = -\boldsymbol{L}^{-1}(\boldsymbol{A} - \boldsymbol{L})$. Oglejmo si, v kakšni zvezi je približek $\boldsymbol{x}^{(r)}$ na r-tem koraku iteracije s točno rešitvijo sistema. Ker je

$$\boldsymbol{x}^{(r)} - \boldsymbol{A}^{-1}\boldsymbol{b} = \left(\boldsymbol{J}_{\boldsymbol{g}_S}\boldsymbol{x}^{(r-1)} + \boldsymbol{L}^{-1}\boldsymbol{b}\right) - \left(\boldsymbol{J}_{\boldsymbol{g}_S}\boldsymbol{A}^{-1}\boldsymbol{b} + \boldsymbol{L}^{-1}\boldsymbol{b}\right),$$

sledi $\boldsymbol{x}^{(r)} - \boldsymbol{A}^{-1}\boldsymbol{b} = \boldsymbol{J}_{\boldsymbol{g}_S}\left(\boldsymbol{x}^{(r-1)} - \boldsymbol{A}^{-1}\boldsymbol{b}\right)$. Če to zvezo uporabimo večkrat zapored, dobimo

$$\boldsymbol{x}^{(r)} - \boldsymbol{A}^{-1}\boldsymbol{b} = \boldsymbol{J}_{\boldsymbol{g}_S}\left(\boldsymbol{x}^{(r-1)} - \boldsymbol{A}^{-1}\boldsymbol{b}\right) = \ldots = \left(\boldsymbol{J}_{\boldsymbol{g}_S}\right)^r\left(\boldsymbol{x}^{(0)} - \boldsymbol{A}^{-1}\boldsymbol{b}\right).$$

Torej je $\boldsymbol{A}^{-1}\boldsymbol{b}$ limita iteracijskega zaporedja natanko tedaj, ko je $\lim_{r\to\infty}\left(\boldsymbol{J}_{\boldsymbol{g}_S}\right)^r = \boldsymbol{0}$, to pa velja natanko tedaj, ko so vse lastne vrednosti matrike $\boldsymbol{J}_{\boldsymbol{g}_S}$ po absolutni vrednosti manjše od 1 (v kar se lahko prepričamo s pomočjo Jordanove dekompozicije matrike). Obravnavajmo poljuben lastni par $(\lambda, \boldsymbol{v})$ matrike $\boldsymbol{J}_{\boldsymbol{g}_S}$. Po definiciji je

$$\lambda\boldsymbol{v} = \boldsymbol{J}_{\boldsymbol{g}_S}\boldsymbol{v} = -\boldsymbol{L}^{-1}(\boldsymbol{A} - \boldsymbol{L})\boldsymbol{v}$$

oziroma $\lambda \boldsymbol{L}\boldsymbol{v} = -(\boldsymbol{A} - \boldsymbol{L})\boldsymbol{v}$, iz česar sledi, da za vsak $i \in \{1, 2, \ldots, n\}$ velja

$$\lambda \sum_{j=1}^{i} a_{i,j} \boldsymbol{v}_j = - \sum_{j=i+1}^{n} a_{i,j} \boldsymbol{v}_j .$$

Iz teh enačb lahko izrazimo

$$-\lambda a_{i,i} \boldsymbol{v}_i = \lambda \sum_{j=1}^{i-1} a_{i,j} \boldsymbol{v}_j + \sum_{j=i+1}^{n} a_{i,j} \boldsymbol{v}_j$$

in če izberemo tisto, pri kateri je $|\boldsymbol{v}_i| = \|\boldsymbol{v}\|_\infty$, sledi

$$|\lambda| \, |a_{i,i}| \leq |\lambda| \sum_{j=1}^{i-1} |a_{i,j}| \frac{|\boldsymbol{v}_j|}{|\boldsymbol{v}_i|} + \sum_{j=i+1}^{n} |a_{i,j}| \frac{|\boldsymbol{v}_j|}{|\boldsymbol{v}_i|} \leq |\lambda| \sum_{j=1}^{i-1} |a_{i,j}| + \sum_{j=i+1}^{n} |a_{i,j}| \, .$$

Torej je

$$|\lambda| \leq \frac{\sum_{j=i+1}^{n} |a_{i,j}|}{|a_{i,i}| - \sum_{j=1}^{i-1} |a_{i,j}|}$$

in zaradi diagonalne dominantnosti velja $|\lambda| < 1$. S tem smo dokazali, da so vse lastne vrednosti matrike $\boldsymbol{J}_{\boldsymbol{g}_S}$ po absolutni vrednosti manjše od 1, kar dokazuje konvergenco iteracijskega zaporedja.

4.2. Newtonova metoda

Newtonova metoda za reševanje sistema nelinearnih enačb $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{0}$ je posplošitev tangentne metode. Iteracijiski korak

$$\boldsymbol{x}^{(r+1)} = \boldsymbol{x}^{(r)} - \boldsymbol{J}_{\boldsymbol{f}} \left(\boldsymbol{x}^{(r)} \right)^{-1} \boldsymbol{f} \left(\boldsymbol{x}^{(r)} \right)$$

izvedemo tako, da rešimo sistem linearnih enačb

$$\boldsymbol{J}_{\boldsymbol{f}} \left(\boldsymbol{x}^{(r)} \right) \cdot \boldsymbol{\Delta x}^{(r)} = -\boldsymbol{f} \left(\boldsymbol{x}^{(r)} \right)$$

in iz $\boldsymbol{\Delta x}^{(r)} = \boldsymbol{x}^{(r+1)} - \boldsymbol{x}^{(r)}$ izrazimo novi približek $\boldsymbol{x}^{(r+1)}$. Kvazi-Newtonove metode Jacobijevo matriko $\boldsymbol{J}_{\boldsymbol{f}}$ v zgornjem postopku nadomestijo s hitreje izračunljivimi matrikami. Pri Broydnovi metodi jo v koraku r zamenja matrika $\boldsymbol{B}_r$, ki je določena rekurizvno z

$$\boldsymbol{B}_r = \boldsymbol{B}_{r-1} + \frac{\boldsymbol{f} \left(\boldsymbol{x}^{(r)} \right) \left(\boldsymbol{\Delta x}^{(r-1)} \right)^{\mathsf{T}}}{\left\| \boldsymbol{\Delta x}^{(r-1)} \right\|_2^2},$$

pri čemer je $\boldsymbol{B}_0$ dober približek za $\boldsymbol{J}_{\boldsymbol{f}}(\boldsymbol{x}^{(0)})$. Broydnovo metodo lahko razumemo kot posplošitev sekantne metode za reševanje nelinearnih enačb.

Naloga 4.5. Dan je sistem

$$x^2 + y^2 = 4, \quad x^2 - y^2 = 1$$

dveh nelinearnih enačb za spremenljivki x in y.

1. Geometrijsko interpretirajte sistem. Koliko rešitev ima? Poiščite rešitve analitično.

2. Izpeljite Newtonovo metodo za reševanje sistema in napravite dva koraka metode pri začetnem približku $(x_0, y_0) = (2, 1)$.

3. Poiščite približek za rešitev sistema z dvema korakoma Broydnove metode pri začetnem približku $(x_0, y_0) = (2, 1)$ in $\boldsymbol{B}_0 = \boldsymbol{J}_{\boldsymbol{f}}(x_0, y_0)$.

Rešitev.

1. Prva enačba predstavlja krožnico s polmerom 2, druga enačba pa hiperbolo. Sistem ima štiri rešitve, to so $(\pm\sqrt{5/2}, \pm\sqrt{3/2})$ in $(\pm\sqrt{5/2}, \mp\sqrt{3/2})$

2. Dani sistem enačb zapišemo v obliki $\boldsymbol{f}(x, y) = \boldsymbol{0}$ za preslikavo $\boldsymbol{f}$, podano s predpisom
$$\boldsymbol{f}(x, y) = (x^2 + y^2 - 4, x^2 - y^2 - 1).$$
Jacobijeva matrika preslikave $\boldsymbol{f}$ je
$$\boldsymbol{J}_{\boldsymbol{f}}(x, y) = \begin{bmatrix} 2x & 2y \\ 2x & -2y \end{bmatrix},$$
zato v prvem koraku metode pri začetnem približku $(x_0, y_0) = (2, 1)$ rešujemo sistem
$$\begin{bmatrix} 4 & 2 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} \Delta x_0 \\ \Delta y_0 \end{bmatrix} = - \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$
Izračunamo $(\Delta x_0, \Delta y_0) = (-\frac{3}{8}, \frac{1}{4})$, kar pomeni, da je novi približek
$$(x_1, y_1) = (x_0, y_0) + (\Delta x_0, \Delta y_0) = \left(\tfrac{13}{8}, \tfrac{5}{4}\right).$$
V drugem koraku z reševanjem sistema
$$\begin{bmatrix} \frac{13}{4} & \frac{5}{2} \\ \frac{13}{4} & -\frac{5}{2} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \Delta y_1 \end{bmatrix} = - \begin{bmatrix} \frac{13}{64} \\ \frac{5}{64} \end{bmatrix}$$
dobimo $(\Delta x_1, \Delta y_1) = (-\frac{9}{208}, -\frac{1}{40})$, torej je
$$(x_2, y_2) = (x_1, y_1) + (\Delta x_1, \Delta y_1) = \left(\tfrac{329}{208}, \tfrac{49}{40}\right) \approx (1.5817, 1.2250).$$
Iteracijsko zaporedje konvergira k rešitvi $(\sqrt{5/2}, \sqrt{3/2}) \approx (1.5811, 1.2247)$.

3. Zaradi izbire $\boldsymbol{B}_0$ je prvi korak Broydnove metode enak kot pri Newtonovi metodi. Za izvedbo drugega koraka najprej izračunamo

$$\boldsymbol{B}_1 = \boldsymbol{B}_0 + \frac{\boldsymbol{f}(x_1, y_1) \cdot (\Delta x_0, \Delta y_0)^\mathsf{T}}{\|(\Delta x_0, \Delta y_0)\|_2^2} = \begin{bmatrix} \frac{29}{8} & \frac{9}{4} \\ \frac{401}{104} & -\frac{99}{52} \end{bmatrix}.$$

Nato rešimo sistem $\boldsymbol{B}_1 \cdot (\Delta x_1, \Delta y_1) = -\boldsymbol{f}(x_1, y_1)$ in približek (x_1, y_1) izračunamo iz rešitve $(\Delta x_1, \Delta y_1) = (-\frac{13}{360}, -\frac{13}{405})$. Dobimo

$$(x_2, y_2) = \left(\tfrac{143}{90}, \tfrac{1973}{1620}\right) \approx (1.5589, 1.2179).$$

Naloga 4.6. Dane so točke (x_i, y_i), $i = 1, 2, \ldots, n$, v ravnini. Radi bi poiskali funkcijo f oblike $f(x) = \alpha e^{\beta x}$, $\alpha, \beta \in \mathbb{R}$, ki minimizira izraz

$$\varphi(\alpha, \beta) = \sum_{i=1}^{n} (y_i - f(x_i))^2 = \sum_{i=1}^{n} \left(y_i - \alpha e^{\beta x_i}\right)^2.$$

Za tak f pravimo, da se točkam najbolje prilega po metodi najmanjših kvadratov.

1. Iskanje minimuma izraza $\varphi(\alpha, \beta)$ prevedite na reševanje sistema nelinearnih enačb s spremenljivkama α in β.

2. Izpeljite Newtonovo metodo za dobljeni sistem.

3. Če so ordinate y_i pozitivne, lahko dober začetni približek za Newtonovo metodo poiščemo z linearizacijo $\varphi(\alpha, \beta)$ v obliko

$$\psi(\alpha, \beta) = \sum_{i=1}^{n} (\ln(y_i) - \ln(\alpha) - \beta x_i)^2.$$

Izpeljite in rešite sistem linearnih enačb za parametra α in β.

4. V Matlabu s pomočjo linearizacije in Newtonove metode poiščite in nariše f, ki se po metodi najmanjših kvadratov najbolje prilega točkam

$$(1, 8),\ (2, 9),\ (3, 6),\ (4, 7),\ (5, 4),\ (6, 3),\ (7, 2),\ (8, 3),\ (9, 2),\ (10, 1).$$

Rešitev primerjajte z rešitvijo prvotnega problema, izračunano s pomočjo vgrajene funkcije `fminsearch`.

Rešitev.

1. Z odvajanjem φ dobimo

$$\frac{\partial \varphi}{\partial \alpha}(\alpha, \beta) = \sum_{i=1}^{n} 2\left(y_i - \alpha e^{\beta x_i}\right)\left(-e^{\beta x_i}\right),$$
$$\frac{\partial \varphi}{\partial \beta}(\alpha, \beta) = \sum_{i=1}^{n} 2\left(y_i - \alpha e^{\beta x_i}\right)\left(-\alpha x_i e^{\beta x_i}\right),$$

kar pomeni, da lahko minimum funkcije φ iščemo z reševanjem sistema nelinearnih enačb $\boldsymbol{f}(\alpha,\beta)=\mathbf{0}$ za

$$\boldsymbol{f}(\alpha,\beta)=\left(\sum_{i=1}^{n} e^{\beta x_i}\left(y_i-\alpha e^{\beta x_i}\right), \alpha\sum_{i=1}^{n} x_i e^{\beta x_i}\left(y_i-\alpha e^{\beta x_i}\right)\right).$$

2. Za izpeljavo Newtonove metode je treba določiti Jacobijevo matriko $\boldsymbol{J_f}$ preslikave $\boldsymbol{f}$. Izračunamo

$$\boldsymbol{J_f}(\alpha,\beta)=\begin{bmatrix} -\sum_{i=1}^{n} e^{2\beta x_i} & \sum_{i=1}^{n} x_i e^{\beta x_i}\left(y_i-2\alpha e^{\beta x_i}\right) \\ \sum_{i=1}^{n} x_i e^{\beta x_i}\left(y_i-2\alpha e^{\beta x_i}\right) & \alpha\sum_{i=1}^{n} x_i^2 e^{\beta x_i}\left(y_i-2\alpha e^{\beta x_i}\right) \end{bmatrix}.$$

Na vsakem koraku metode novi približek $(\alpha_{r+1},\beta_{r+1})$ izračunamo iz prejšnjega (α_r,β_r) tako, da rešimo sistem

$$\boldsymbol{J_f}(\alpha_r,\beta_r)\cdot(\Delta\alpha_r,\Delta\beta_r)=-\boldsymbol{f}(\alpha_r,\beta_r)$$

in vzamemo

$$(\alpha_{r+1},\beta_{r+1})=(\alpha_r,\beta_r)+(\Delta\alpha_r,\Delta\beta_r).$$

3. Naj bo $X_i=x_i$, $Y_i=\ln(y_i)$, $A=\ln(\alpha)$ in $B=\beta$. Tedaj je

$$\psi(A,B)=\sum_{i=1}^{n}\left(Y_i-A-BX_i\right)^2$$

in iz

$$\frac{\partial\psi}{\partial A}(A,B)=\sum_{i=1}^{n} 2\left(Y_i-A-BX_i\right)(-1),$$
$$\frac{\partial\psi}{\partial B}(A,B)=\sum_{i=1}^{n} 2\left(Y_i-A-BX_i\right)(-X_i)$$

podobno kot v prvi točki naloge sledi

$$\begin{bmatrix} n & \sum_{i=1}^{n} X_i \\ \sum_{i=1}^{n} X_i & \sum_{i=1}^{n} X_i^2 \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix}=\begin{bmatrix} \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_iY_i \end{bmatrix}.$$

To ni nič drugega kot normalni sistem za predoločen sistem, ki določa najboljšo aproksimacijo parov (X_i,Y_i) s premico $x\mapsto A+Bx$ po metodi najmanjših kvadratov. Prvi približek za parametra α in β lahko torej določimo na podlagi rešitve

sistema

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \ln(\alpha) \\ \beta \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \ln(y_i) \\ \sum_{i=1}^{n} x_i \ln(y_i) \end{bmatrix}.$$

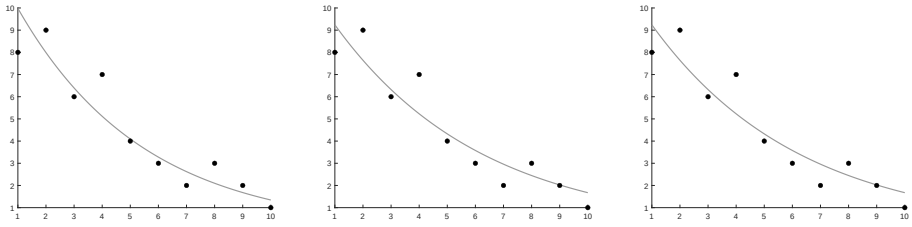
4. V Matlabu pripravimo implementacijo Newtonove metode, ki za dano preslikavo $\boldsymbol{f}$ in njeno Jacobijevo matriko vrne približek za rešitev sistema $\boldsymbol{f}(\boldsymbol{x}) = \mathbf{0}$. Funkcija deluje tako, da pri podanem začetnem približku izvede toliko iteracij, da je norma razlike dveh zaporednih približkov manjša od podane tolerance, razen če število iteracij preseže maksimalno predpisano število le-teh.

```
function [x,X,k] = newton(f,Jf,x0,tol,N)
% Opis:
%  newton izvede Newtonovo metodo za reševanje sistema
%  nelinearnih enačb
%
% Definicija:
%  [x,X,k] = newton(f,Jf,x0,tol,N)
%
% Vhodni podatki:
%  f    preslikava, ki določa sistem f(x) = 0,
%  Jf   Jacobijeva matrika preslikave f,
%  x0   začetni približek (stolpec),
%  tol  toleranca norme razlike dveh zaporednih
%       približkov,
%  N    maksimalno število korakov iteracije
%
% Izhodni podatki:
%  x    končni približek za rešitev sistema f(x) = 0,
%  X    tabela vseh izračunanih približkov,
%  k    število opravljenih korakov

X = x0;
dx = Inf;
k = 0;
while norm(dx) > tol && k < N
   k = k+1;
   dx = -Jf(X(:,k))\f(X(:,k));
   X(:,k+1) = X(:,k) + dx;
end
x = X(:,k+1);

end
```

Začetni približek za Newtonovo metodo poiščemo z linearizacijo.

```
X = (1:10)';
Y = [8; 9; 6; 7; 4; 3; 2; 3; 2; 1];

A = [10 sum(X); sum(X) X'*X];
logY = log(Y);
logx0 = A\[sum(logY); X'*logY];
x0 = [exp(logx0(1)); logx0(2)];
```

Za začetni približek vzamemo par, ki je rezultat zgornjega izračuna in je približno enak $(12.4733, -0.2228)$. Nato z ustrezno definiranima funkcijama, ki za dani vektor $\boldsymbol{x}$ vrneta vektor $\boldsymbol{f}(\boldsymbol{x})$ in matriko $\boldsymbol{J_f}(\boldsymbol{x})$, izvedemo funkcijo `newton`. Pri toleranci 10^{-10} v 5 korakih iteracije dobimo približek, ki je približno enak $(11.1713, -0.1898)$. Skoraj povsem enak približek dobimo, če rešitev problema poiščemo z vgrajeno funkcijo `fminsearch`. Kot vhodna podatka ji podamo funkcijo φ in začetni približek, ki smo ga dobili z linearizacijo. Rezultati so grafično predstavljeni na sliki 4.1.



(A) *linearizacija* (B) *Newtonova metoda* (C) *funkcija* `fminsearch`

SLIKA 4.1: *Grafi funkcije f, ki jih dobimo pri reševanju naloge 4.6.*

5. Predoločeni sistemi enačb

Sistem linearnih enačb $\boldsymbol{Ax} = \boldsymbol{b}$ ima včasih več enačb kot neznank. To pomeni, da je $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ pravokotna matrika, ki ima več vrstic kot stolpcev ($m > n$). Vektor $\boldsymbol{x} \in \mathbb{R}^n$, ki bi za poljuben vektor $\boldsymbol{b} \in \mathbb{R}^m$ zadostil sistemu enačb, ne obstaja, zato navadno iščemo $\boldsymbol{x}$, ki minimizira drugo normo ostanka $\boldsymbol{Ax} - \boldsymbol{b}$.

5.1. Normalni sistem

Problem iskanja vektorja $\boldsymbol{x} \in \mathbb{R}^n$, ki minimizira vrednost $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2$, ima v primeru, ko je matrika $\boldsymbol{A}$ polnega ranga ($\operatorname{rang}(\boldsymbol{A}) = n$), enolično rešitev, ki jo imenujemo rešitev po metodi najmanjših kvadratov. Določena je kot rešitev normalnega sistema $\boldsymbol{A}^\mathsf{T}\boldsymbol{Ax} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$. Matrika $\boldsymbol{A}^\mathsf{T}\boldsymbol{A} \in \mathbb{R}^{n\times n}$ je simetrična pozitivno definitna, zato lahko normalni sistem rešujemo z uporabo razcepa Choleskega.

Naloga 5.1. Dokažite, da ima sistem

$$\begin{bmatrix} \boldsymbol{I} & \boldsymbol{A} \\ \boldsymbol{A}^\mathsf{T} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{r} \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \boldsymbol{b} \\ \boldsymbol{0} \end{bmatrix}$$

rešitev, ki ustreza rešitvi predoločenega sistema $\boldsymbol{Ax} = \boldsymbol{b}$ po metodi najmanjših kvadratov. Geometrijsko interpretirajte sistem.

Rešitev. Vektorja $\boldsymbol{r}$ in $\boldsymbol{x}$ ustrezata enakostma $\boldsymbol{r} + \boldsymbol{Ax} = \boldsymbol{b}$ in $\boldsymbol{A}^\mathsf{T}\boldsymbol{r} = \boldsymbol{0}$. Če v prvi izrazimo $\boldsymbol{r}$ in ga vstavimo v drugo, dobimo normalni sistem $\boldsymbol{A}^\mathsf{T}\boldsymbol{Ax} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$. Vektor $\boldsymbol{r}$ predstavlja razliko med vektorjem $\boldsymbol{b}$ in vektorjem $\boldsymbol{Ax}$, ki leži v prostoru $\operatorname{im}(\boldsymbol{A})$. Z zahtevo $\boldsymbol{A}^\mathsf{T}\boldsymbol{r} = \boldsymbol{0}$ predpišemo, da je $\boldsymbol{r}$ pravokoten na $\operatorname{im}(\boldsymbol{A})$ oziroma da je $\boldsymbol{Ax}$ pravokotna projekcija $\boldsymbol{b}$ na $\operatorname{im}(\boldsymbol{A})$.

Naloga 5.2. Funkcija f je podana v štirih točkah:

$$f(-1) = \tfrac{11}{4}, \quad f(0) = \tfrac{7}{4}, \quad f(1) = \tfrac{1}{4}, \quad f(2) = \tfrac{13}{4}.$$

Poiščite parabolo, ki se po metodi najmanjših kvadratov najbolje prilega danim vrednostim funkcije f. Izpeljite normalni sistem in ga rešite s pomočjo razcepa Choleskega.

Rešitev. Parabolo predstavimo v obliki $p(x) = a_0 + a_1 x + a_2 x^2$. Veljati mora $p(x) = f(x)$ za $x \in \{-1, 0, 1, 2\}$, torej so koeficienti a_0, a_1, a_2 določeni s predoločenim sistemom

$$\begin{bmatrix} 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \frac{1}{4} \begin{bmatrix} 11 \\ 7 \\ 1 \\ 13 \end{bmatrix}.$$

Naj bo matrika sistema označena z $\boldsymbol{A}$, desna stran sistema pa z $\boldsymbol{b}$. Normalni sistem $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$ je podan kot

$$\begin{bmatrix} 4 & 2 & 6 \\ 2 & 6 & 8 \\ 6 & 8 & 18 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} = \begin{bmatrix} 8 \\ 4 \\ 16 \end{bmatrix}.$$

Njegova rešitev ustreza vektorju $\boldsymbol{x}$, ki minimizira $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ oziroma vsoto kvadratov razlik med vrednostmi parabole p in funkcije f v točkah -1, 0, 1, 2. Faktor Choleskega za matriko $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ je podan z

$$\boldsymbol{V} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & \sqrt{5} & 0 \\ 3 & \sqrt{5} & 2 \end{bmatrix}.$$

Ker je $\boldsymbol{A}^\mathsf{T}\boldsymbol{A} = \boldsymbol{V}\boldsymbol{V}^\mathsf{T}$, lahko koeficiente parabole izračunamo z reševanjem sistemov $\boldsymbol{V}\boldsymbol{y} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$ in $\boldsymbol{V}^\mathsf{T}\boldsymbol{x} = \boldsymbol{y}$. Dobimo $\boldsymbol{y} = (4, 0, 2)$ in $\boldsymbol{x} = (1, -1, 1)$, kar pomeni, da je $p(x) = x^2 - x + 1$ parabola, ki se po metodi najmanjših kvadratov najbolje danim podatkom.

Naloga 5.3. Zapišite normalni sistem, ki določa polinom stopnje n, ki se po metodi najmanjših kvadratov najbolje prilega vrednostim funkcije f v $m+1$ paroma različnih točkah $x_0, x_1, \ldots, x_m$.

Rešitev. Polinom zapišimo v obliki

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

Predoločen sistem določajo enačbe $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, m$, kar lahko v vektorski obliki zapišemo kot

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_m & x_m^2 & \ldots & x_m^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_m) \end{bmatrix}.$$

Normalni sistem je zato

$$\left[\sum_{k=0}^{m} x_k^{j+i} \right]_{i,j=0}^{n} \cdot \left[\; a_i \; \right]_{j=0}^{n} = \left[\sum_{k=0}^{m} x_k^j \, f(x_k) \right]_{j=0}^{n}.$$

Naloga 5.4. Planet se giblje po eliptični orbiti. Znanih je nekaj položajev planeta, ki so v spodnji tabeli podani v obliki kartezičnih koordinat (x_i, y_i), $i = 1, 2, \ldots, 10$, glede na ravnino orbite.

x_i	-0.5	-0.3	0.1	0.8	0.9	0.5	0.2	-0.1	-0.3	-0.6
y_i	-0.3	-0.7	-0.8	-0.7	-0.1	0.5	0.6	0.7	0.6	0.2

Določite koeficiente a, b, c, d, e v kvadratni formi

$$ax^2 + bxy + cy^2 + dx + ey = 1,$$

ki minimizirajo vrednost

$$\sum_{i=1}^{10} \left(ax_i^2 + bx_iy_i + cy_i^2 + dx_i + ey_i - 1\right)^2.$$

Interpretirajte problem kot reševanje predoločenega sistema enačb in v Matlabu narišite tirnico planeta, ki jo določa njegova rešitev po metodi najmanjših kvadratov.

Rešitev. Iskani koeficienti a, b, c, d, e ustrezajo rešitvi predoločenega sistema

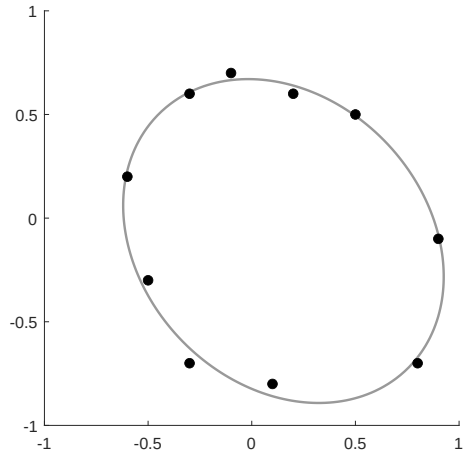
$$\begin{bmatrix} x_1^2 & x_1y_1 & y_1^2 & x_1 & y_1 \\ x_2^2 & x_2y_2 & y_2^2 & x_2 & y_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{10}^2 & x_{10}y_{10} & y_{10}^2 & x_{10} & y_{10} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \\ d \\ e \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

po metodi najmanjših kvadratov. V Matlabu lahko rešitev predoločenega sistema izračunamo kar s pomočjo operatorja \, enako kot da bi reševali sistem linearnih enačb z enakim številom enačb in neznank.

```
X = [-0.5;-0.3;0.1;0.8;0.9;0.5;0.2;-0.1;-0.3;-0.6];
Y = [-0.3;-0.7;-0.8;-0.7;-0.1;0.5;0.6;0.7;0.6;0.2];

A = [X.^2 X.*Y Y.^2 X Y];
b = ones(10,1);

k = A\b;
```

Kvadratno formo, ki je določena s koeficienti vektorja $\boldsymbol{k} = (a, b, c, d, e)$, lahko v Matlabu narišemo z uporabo vgrajenih funkcij `meshgrid` in `contour`. Skupaj s podanimi položaji planeta je prikazana na sliki 5.1.

Problemi, ki jih prevedemo na reševanje predoločenega sistema, pogosto vključujejo dodatne omejitve, zaradi katerih moramo osnovni model reševanja nekoliko prilagoditi.



SLIKA 5.1: *Položaji planeta, podani v nalogi 5.4, in kvadratna forma, ki se jim najbolje prilega po metodi najmanjših kvadratov.*

Naloga 5.5. Predoločeni sistem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ z matriko $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m > n$, in vektorjem $\boldsymbol{b} \in \mathbb{R}^m$ vključuje dodatne omejitve, ki so izražane z matriko $\boldsymbol{C} \in \mathbb{R}^{p\times n}$, $p < n$, in vektorjem $\boldsymbol{d} \in \mathbb{R}^p$, v obliki nedoločenega sistema enačb $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$. Za matriki $\boldsymbol{A}$ in $\boldsymbol{C}$ predpostavljamo, da sta polnega ranga. Iščemo $\boldsymbol{x} \in \mathbb{R}^n$, ki minimizira $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2$ ob omejitvi, da je $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$.

1. Naj bo $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n \times \mathbb{R}^p$ rešitev sistema
$$\begin{bmatrix} \boldsymbol{A}^\mathsf{T}\boldsymbol{A} & \boldsymbol{C}^\mathsf{T} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{A}^\mathsf{T}\boldsymbol{b} \\ \boldsymbol{d} \end{bmatrix}.$$
Dokažite, da je $\boldsymbol{x}$ rešitev minimizacijskega problema.

2. Utemeljite, da rešitev $\boldsymbol{x}$ obstaja in je enolična.

Rešitev.

1. Naj bo $\widetilde{\boldsymbol{x}} \in \mathbb{R}^n$ poljuben vektor, za katerega velja $\boldsymbol{C}\widetilde{\boldsymbol{x}} = \boldsymbol{d}$. Tudi vektor $\boldsymbol{x}$, ki je delna rešitev zgornjega bločnega sistema, zadošča pogoju $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{d}$. Dokažimo, da je $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 \le \|\boldsymbol{A}\widetilde{\boldsymbol{x}} - \boldsymbol{b}\|_2$. Kvadrat $\|\boldsymbol{A}\widetilde{\boldsymbol{x}} - \boldsymbol{b}\|_2$ lahko zapišemo kot
$$\|\boldsymbol{A}(\widetilde{\boldsymbol{x}} - \boldsymbol{x}) + \boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{A}(\widetilde{\boldsymbol{x}} - \boldsymbol{x})\|_2^2 + 2(\widetilde{\boldsymbol{x}} - \boldsymbol{x})^\mathsf{T}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) + \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 .$$
Sedaj upoštevamo, da za $\boldsymbol{x}$ velja $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} + \boldsymbol{C}^\mathsf{T}\boldsymbol{y} = \boldsymbol{A}^\mathsf{T}\boldsymbol{b}$ in za člen
$$2(\widetilde{\boldsymbol{x}} - \boldsymbol{x})^\mathsf{T}\boldsymbol{A}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = 2(\widetilde{\boldsymbol{x}} - \boldsymbol{x})^\mathsf{T}\boldsymbol{C}^\mathsf{T}\boldsymbol{y}$$
sklepamo, da je enak 0, saj je $\boldsymbol{C}\boldsymbol{x} = \boldsymbol{C}\widetilde{\boldsymbol{x}}$. Od tod potem takoj sledi
$$\|\boldsymbol{A}\widetilde{\boldsymbol{x}} - \boldsymbol{b}\|_2^2 = \|\boldsymbol{A}(\widetilde{\boldsymbol{x}} - \boldsymbol{x})\|_2^2 + \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 \ge \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2^2 .$$

2. Zadošča utemeljitev, da je bločna matrika, ki nastopa v sistemu iz prve točke, obrnljiva. Denimo, da ni. Potem obstaja neničeln vektor $(\widehat{\boldsymbol{x}}, \widehat{\boldsymbol{y}}) \in \mathbb{R}^n \times \mathbb{R}^p$, da je

$$\begin{bmatrix} \boldsymbol{A}^\mathsf{T}\boldsymbol{A} & \boldsymbol{C}^\mathsf{T} \\ \boldsymbol{C} & \boldsymbol{0} \end{bmatrix} \begin{bmatrix} \widehat{\boldsymbol{x}} \\ \widehat{\boldsymbol{y}} \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}.$$

Od tod sledi

$$\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\widehat{\boldsymbol{x}} + \boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{y}} = \boldsymbol{0}, \qquad \boldsymbol{C}\widehat{\boldsymbol{x}} = \boldsymbol{0}.$$

Jasno je, da je $\widehat{\boldsymbol{x}}^\mathsf{T}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}\widehat{\boldsymbol{x}} + \boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{y}}) = 0$, kar pomeni, da je $\|\boldsymbol{A}\widehat{\boldsymbol{x}}\|_2 = 0$, saj je $\boldsymbol{C}\widehat{\boldsymbol{x}} = \boldsymbol{0}$. Torej je $\boldsymbol{A}\widehat{\boldsymbol{x}} = \boldsymbol{0}$ in ker je $\boldsymbol{A}$ polnega ranga, je $\widehat{\boldsymbol{x}} = \boldsymbol{0}$. Sedaj prva enačba implicira $\boldsymbol{C}^\mathsf{T}\widehat{\boldsymbol{y}} = \boldsymbol{0}$; tudi matrika $\boldsymbol{C}$ je polnega ranga, zato mora biti $\widehat{\boldsymbol{y}} = \boldsymbol{0}$, kar je v protislovju z začetno predpostavko.

Naloga 5.6. Naj bo $\boldsymbol{C} \in \mathbb{R}^{m\times m}$ simetrična pozitivno definitna matrika, ki določa normo

$$\|\boldsymbol{x}\|_{\boldsymbol{C}} = \sqrt{\boldsymbol{x}^\mathsf{T}\boldsymbol{C}\boldsymbol{x}}.$$

Zapišite normalni sistem, ki določa rešitev problema

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_{\boldsymbol{C}}.$$

Rešitev. Ker je $\boldsymbol{C}$ simetrična pozitivno definitna, jo lahko zapišemo kot $\boldsymbol{C} = \boldsymbol{V}\boldsymbol{V}^\mathsf{T}$, kjer je $\boldsymbol{V}$ faktor Choleskega. Kvadrat norme $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_{\boldsymbol{C}}$ je zato enak

$$(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^\mathsf{T}\boldsymbol{C}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \left(\boldsymbol{V}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right)^\mathsf{T}\left(\boldsymbol{V}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right),$$

torej ustreza kvadratu norme $\left\|\boldsymbol{V}^\mathsf{T}\boldsymbol{A}\boldsymbol{x} - \boldsymbol{V}^\mathsf{T}\boldsymbol{b}\right\|_2$. To pomeni, da je minimalna vrednost $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_{\boldsymbol{C}}$ dosežena pri vektorju $\boldsymbol{x} \in \mathbb{R}^n$, ki je rešitev sistema

$$\left(\boldsymbol{V}^\mathsf{T}\boldsymbol{A}\right)^\mathsf{T}\left(\boldsymbol{V}^\mathsf{T}\boldsymbol{A}\right)\boldsymbol{x} = \left(\boldsymbol{V}^\mathsf{T}\boldsymbol{A}\right)^\mathsf{T}\boldsymbol{V}^\mathsf{T}\boldsymbol{b}$$

in ga lahko poenostavimo v $\boldsymbol{A}^\mathsf{T}\boldsymbol{C}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^\mathsf{T}\boldsymbol{C}\boldsymbol{b}$. To je normalni sistem, ki določa rešitev problema po metodi najmanjših kvadratov v normi $\|\cdot\|_{\boldsymbol{C}}$.

5.2. QR razcep

Iskanja rešitve predoločenega sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ se navadno ne lotevamo z reševanjem normalnega sistema, saj je postopek v primeru, ko je par stolpcev matrike $\boldsymbol{A}$ skoraj linearno odvisen, nestabilen. Težavam se izognemo, če stolpce matrike $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ nadomestimo z ortonormirano bazo za $\operatorname{im}(\boldsymbol{A})$, kar dosežemo z razcepom $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ matrike $\boldsymbol{A}$ na ortogonalno matriko $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$ ($\boldsymbol{Q}^\mathsf{T}\boldsymbol{Q} = \boldsymbol{I}$) in zgornje trikotno matriko $\boldsymbol{R} \in \mathbb{R}^{n\times n}$.

Naloga 5.7. S pomočjo QR razcepa matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ dokažite Hadamardovo neenakost

$$|\det(\boldsymbol{A})| \leq \prod_{j=1}^{n} \|\boldsymbol{a}_j\|_2 ,$$

kjer so $\boldsymbol{a}_j$, $j = 1, 2, \ldots, n$, stolpci matrike $\boldsymbol{A}$.

Rešitev. Če je matrika $\boldsymbol{A}$ singularna, je trditev trivialna, zato privzemimo, da je $\operatorname{rang}(\boldsymbol{A}) = n$. Naj bosta $\boldsymbol{Q}$ in $\boldsymbol{R}$ matriki QR razcepa $\boldsymbol{A} = \boldsymbol{QR}$ matrike $\boldsymbol{A}$. Ker je $\boldsymbol{Q}$ ortogonalna, je absolutna vrednost determinante $\boldsymbol{A}$ enaka determinanti $\boldsymbol{R}$, to je produktu diagonalnih elementov $r_{j,j}$, $j = 1, 2, \ldots, n$, matrike $\boldsymbol{R}$. Poleg tega iz razvoja

$$\boldsymbol{a}_j = \sum_{i=1}^{j} r_{i,j}\boldsymbol{q}_i$$

zaradi ortonormiranosti vektorjev $\boldsymbol{q}_i$ sledi, da je $|r_{j,j}| \leq \|\boldsymbol{a}_j\|_2$. S tem je neenakost dokazana.

Pri modificiranem Gram–Schmidtovem postopku QR razcep $\boldsymbol{A} = \boldsymbol{QR}$ matrike $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ izračunamo v n korakih. Naj bodo stolpci matrike $\boldsymbol{A}$ označeni z $\boldsymbol{a}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$, stolpci matrike $\boldsymbol{Q}$ pa s $\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_n$. V koraku $j \in \{1, 2, \ldots, n\}$ začnemo z vektorjem $\boldsymbol{a}_j$ in nato postopno odštejemo pravokotne projekcije trenutnega vektorja na vektorje $\boldsymbol{q}_i$, $i = 1, 2, \ldots, j-1$:

$$\boldsymbol{q}_j^{(1)} = \boldsymbol{a}_j, \qquad \boldsymbol{q}_j^{(i+1)} = \boldsymbol{q}_j^{(i)} - r_{i,j}\boldsymbol{q}_i, \quad r_{i,j} = \boldsymbol{q}_i^{\mathsf{T}}\boldsymbol{q}_j^{(i)}.$$

Na koncu dobljeni vektor normiramo:

$$\boldsymbol{q}_j = \frac{1}{r_{j,j}}\boldsymbol{q}_j^{(j)}, \qquad r_{j,j} = \left\|\boldsymbol{q}_j^{(j)}\right\|_2 .$$

S tem postopkom, ki zahteva približno $2mn^2$ operacij, zagotovimo, da so vektorji $\boldsymbol{q}_j$ normirani in paroma pravokotni, koeficienti $r_{i,j}$, $i = 1, 2, \ldots, j$, ki nastopajo v zgornje trikotni matriki $\boldsymbol{R}$, pa določajo razvoj vektorja $\boldsymbol{a}_j$ po vektorjih $\boldsymbol{q}_i$. Iz načina izračuna sledi še, da so diagonalni elementi matrike $\boldsymbol{R}$ pozitivni.

Naloga 5.8. Naj bo

$$\boldsymbol{A} = \begin{bmatrix} 1 & 0 & -1 \\ 1 & 0 & -3 \\ 0 & 1 & 1 \\ 0 & -1 & 1 \end{bmatrix}.$$

Izračunajte QR razcep matrike $\boldsymbol{A}$ z modificiranim Gram–Schmidtovim postopkom.

Rešitev. QR razcep $\boldsymbol{A} = \boldsymbol{QR}$ matrike $\boldsymbol{A}$ s stolpci

$$\boldsymbol{a}_1 = (1, 1, 0, 0), \quad \boldsymbol{a}_2 = (0, 0, 1, -1), \quad \boldsymbol{a}_3 = (-1, -3, 1, 1)$$

izračunamo v treh korakih.

1. korak: Definiramo $\boldsymbol{q}_1^{(1)} = \boldsymbol{a}_1$ in izračunamo

$$i = 1: \quad r_{1,1} = \left\|\boldsymbol{q}_1^{(1)}\right\|_2 = \sqrt{2}, \qquad \boldsymbol{q}_1 = \frac{1}{r_{1,1}}\boldsymbol{q}_1^{(1)} = \frac{\sqrt{2}}{2}(1,1,0,0).$$

2. korak: Definiramo $\boldsymbol{q}_2^{(1)} = \boldsymbol{a}_2$ in izračunamo

$$i = 1: \quad r_{1,2} = \boldsymbol{q}_1^{\mathsf{T}}\boldsymbol{q}_2^{(1)} = 0, \qquad \boldsymbol{q}_2^{(2)} = \boldsymbol{q}_2^{(1)} - r_{1,2}\boldsymbol{q}_1 = \boldsymbol{q}_2^{(1)},$$

$$i = 2: \quad r_{2,2} = \left\|\boldsymbol{q}_2^{(2)}\right\|_2 = \sqrt{2}, \qquad \boldsymbol{q}_2 = \frac{1}{r_{2,2}}\boldsymbol{q}_2^{(2)} = \frac{\sqrt{2}}{2}(0,0,1,-1).$$

3. korak: Definiramo $\boldsymbol{q}_3^{(1)} = \boldsymbol{a}_3$ in izračunamo

$$i = 1: \quad r_{1,3} = \boldsymbol{q}_1^{\mathsf{T}}\boldsymbol{q}_3^{(1)} = -2\sqrt{2}, \qquad \boldsymbol{q}_3^{(2)} = \boldsymbol{q}_3^{(1)} - r_{1,3}\boldsymbol{q}_1 = (1,-1,1,1),$$

$$i = 2: \quad r_{2,3} = \boldsymbol{q}_2^{\mathsf{T}}\boldsymbol{q}_3^{(2)} = 0, \qquad \boldsymbol{q}_3^{(3)} = \boldsymbol{q}_3^{(2)} - r_{2,3}\boldsymbol{q}_2 = (1,-1,1,1),$$

$$i = 3: \quad r_{3,3} = \left\|\boldsymbol{q}_3^{(3)}\right\|_2 = 2, \qquad \boldsymbol{q}_3 = \frac{1}{r_{3,3}}\boldsymbol{q}_3^{(3)} = \frac{1}{2}(1,-1,1,1).$$

Ortogonalna matrika $\boldsymbol{Q}$ je določena s stolpci $\boldsymbol{q}_1$, $\boldsymbol{q}_2$, $\boldsymbol{q}_3$, zgornje trikotna matrika $\boldsymbol{R}$ pa z elementi $r_{i,j}$, $1 \le i \le j \le 3$:

$$\boldsymbol{Q} = \frac{1}{2}\begin{bmatrix} \sqrt{2} & 0 & 1 \\ \sqrt{2} & 0 & -1 \\ 0 & \sqrt{2} & 1 \\ 0 & -\sqrt{2} & 1 \end{bmatrix}, \quad \boldsymbol{R} = \begin{bmatrix} \sqrt{2} & 0 & -2\sqrt{2} \\ & \sqrt{2} & 0 \\ & & 2 \end{bmatrix}.$$

V teoriji bi lahko koeficient $r_{i,j}$ matrike $\boldsymbol{R}$ izračunali tudi kot skalarni produkt vektorjev $\boldsymbol{q}_i$ in $\boldsymbol{a}_j$. To je klasični Gram–Schmidtov postopek, ki pa pri numeričnem računanju daje slabše rezultate, saj se izkaže, da je ortogonalnost bolje popravljati glede na vektorje, ki že vsebujejo zaokrožitvene napake. Iz istega razloga sistema $\boldsymbol{Ax} = \boldsymbol{b}$ ne rešujemo z razcepom matrike $\boldsymbol{A}$ in prevedbo na (prilagojeni normalni) sistem $\boldsymbol{Rx} = \boldsymbol{Q}^{\mathsf{T}}\boldsymbol{b}$, pač pa poiščemo QR razcep razširjene matrike

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{q}_{n+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{z} \\ & \rho \end{bmatrix},$$

kjer je $\boldsymbol{Q} \in \mathbb{R}^{m\times n}$, $\boldsymbol{q}_{n+1} \in \mathbb{R}^m$, $\boldsymbol{R} \in \mathbb{R}^{n\times n}$, $\boldsymbol{z} \in \mathbb{R}^n$ in $\rho \in \mathbb{R}$. Iz

$$\begin{aligned} \boldsymbol{Ax} - \boldsymbol{b} &= \begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ -1 \end{bmatrix} \\ &= \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{q}_{n+1} \end{bmatrix} \begin{bmatrix} \boldsymbol{R} & \boldsymbol{z} \\ & \rho \end{bmatrix} \begin{bmatrix} \boldsymbol{x} \\ -1 \end{bmatrix} = \boldsymbol{Q}(\boldsymbol{Rx} - \boldsymbol{z}) - \rho\boldsymbol{q}_{n+1} \end{aligned}$$

in dejstva, da je $\boldsymbol{q}_{n+1}$ po konstrukciji pravokoten na stolpce $\boldsymbol{Q}$, potem sledi, da je vrednost $\|\boldsymbol{Ax} - \boldsymbol{b}\|_2$ minimalna natanko tedaj, ko je $\boldsymbol{x}$ rešitev sistema $\boldsymbol{Rx} = \boldsymbol{z}$. Vrednost minimuma je enaka $|\rho| = \rho$.

Naloga 5.9. Naj bo $\boldsymbol{A}$ matrika iz naloge 5.8 in $\boldsymbol{b} = (4, 6, -1, 2)$. Dopolnite modificirani Gram–Schmidtov postopek za matriko $\boldsymbol{A}$ in poiščite QR razcep z vektorjem $\boldsymbol{b}$ razširjene matrike $\boldsymbol{A}$. S pomočjo rezultata izračunajte rešitev predoločenega sistema $\boldsymbol{Ax} = \boldsymbol{b}$.

Rešitev. Nadaljujemo postopek iz naloge 5.8 s še enim korakom.

4. korak: Definiramo $\boldsymbol{q}_4^{(1)} = \boldsymbol{b}$ in izračunamo

$$\begin{aligned}
i = 1: &\quad r_{1,4} = \boldsymbol{q}_1^\mathsf{T} \boldsymbol{q}_4^{(1)} = 5\sqrt{2}, &&\quad \boldsymbol{q}_4^{(2)} = \boldsymbol{q}_4^{(1)} - r_{1,4}\boldsymbol{q}_1 = (-1, 1, -1, 2),\\
i = 2: &\quad r_{2,4} = \boldsymbol{q}_2^\mathsf{T} \boldsymbol{q}_4^{(2)} = -\frac{3\sqrt{2}}{2}, &&\quad \boldsymbol{q}_4^{(3)} = \boldsymbol{q}_4^{(2)} - r_{2,4}\boldsymbol{q}_2 = \frac{1}{2}(-2, 2, 1, 1),\\
i = 3: &\quad r_{3,4} = \boldsymbol{q}_3^\mathsf{T} \boldsymbol{q}_4^{(3)} = -\frac{1}{2}, &&\quad \boldsymbol{q}_4^{(4)} = \boldsymbol{q}_4^{(3)} - r_{3,4}\boldsymbol{q}_3 = \frac{1}{4}(-3, 3, 3, 3),\\
i = 4: &\quad r_{4,4} = \left\| \boldsymbol{q}_4^{(4)} \right\|_2 = \frac{3}{2}, &&\quad \boldsymbol{q}_4 = \frac{1}{r_{4,4}} \boldsymbol{q}_4^{(4)} = \frac{1}{2}(-1, 1, 1, 1).
\end{aligned}$$

Velja torej

$$\begin{bmatrix} \boldsymbol{A} & \boldsymbol{b} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} \sqrt{2} & 0 & 1 & -1 \\ \sqrt{2} & 0 & -1 & 1 \\ 0 & \sqrt{2} & 1 & 1 \\ 0 & -\sqrt{2} & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} \sqrt{2} & 0 & -2\sqrt{2} & 5\sqrt{2} \\ & \sqrt{2} & 0 & -\frac{3\sqrt{2}}{2} \\ & & 2 & -\frac{1}{2} \\ & & & \frac{3}{2} \end{bmatrix},$$

iz česar sklepamo, da je $\min_{\boldsymbol{x} \in \mathbb{R}^3} \|\boldsymbol{Ax} - \boldsymbol{b}\|_2 = \frac{3}{2}$. Vektor $\boldsymbol{x} \in \mathbb{R}^3$, pri katerem je minimum dosežen, izračunamo z reševanjem sistema

$$\begin{bmatrix} \sqrt{2} & 0 & -2\sqrt{2} \\ & \sqrt{2} & 0 \\ & & 2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} 5\sqrt{2} \\ -\frac{3\sqrt{2}}{2} \\ -\frac{1}{2} \end{bmatrix}$$

in je enak $\boldsymbol{x} = \frac{1}{4}(18, -6, -1)$.

Naloga 5.10. Z uporabo modificiranega Gram–Schmidtovega postopka določite funkcijo f oblike

$$f(x) = a + bx^2 + c\sin(\tfrac{\pi x}{3}), \quad a, b, c \in \mathbb{R},$$

ki po metodi najmanjših kvadratov najbolje aproksimira točke $(-2, 2)$, $(-1, -2)$, $(1, 2)$, $(2, 4)$. Ali so parametri a, b, c enolično določeni? Ali enako velja tudi v primeru, če funkcijo $x \mapsto \sin(\frac{\pi x}{3})$ v predpisu za f zamenjate z $x \mapsto \cos(\frac{\pi x}{3})$?

Rešitev. V prvem delu naloge iščemo rešitev predoločenega sistema, ki je določen z enačbami

$$a + 4b - \frac{\sqrt{3}}{2}c = 2, \quad a + b - \frac{\sqrt{3}}{2}c = -2, \quad a + b + \frac{\sqrt{3}}{2}c = 2, \quad a + 4b + \frac{\sqrt{3}}{2}c = 4.$$

Z modificiranim Gram–Schmidtovim postopkom ga prevedemo v (prilagojeni normalni) sistem

$$\begin{bmatrix} 2 & 5 & 0 \\ 0 & 3 & 0 \\ 0 & 0 & \sqrt{3} \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

z enolično rešitvijo $a = -1$, $b = 1$, $c = \sqrt{3}$. Torej je $f(x) = -1 + x^2 + \sqrt{3}\sin(\frac{\pi x}{3})$ iskana funkcija.

Če v funkciji f zamenjamo $x \mapsto \sin(\frac{\pi x}{3})$ z $x \mapsto \cos(\frac{\pi x}{3})$, dobimo predoločeni sistem

$$a + 4b - \frac{1}{2}c = 2, \quad a + b + \frac{1}{2}c = -2, \quad a + b + \frac{1}{2}c = 2, \quad a + 4b - \frac{1}{2}c = 4,$$

ki ni polnega ranga, saj leve strani enačb ustrezajo le dvema različnima izrazoma. Če uporabimo modificirani Gram–Schmidtov postopek, dobimo (prilagojeni normalni) sistem

$$\begin{bmatrix} 2 & 5 & 0 \\ 0 & 3 & -1 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 3 \\ 3 \end{bmatrix},$$

kar pomeni, da rešitvi problema ustreza funkcija $f(x) = a + bx^2 + c\cos(\frac{\pi x}{3})$ pri vseh parametrih a, b, c, ki zadoščajo zvezama $2a + 5b = 3$ in $3b - c = 3$.

Pri računanju QR razcepa matrike $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ ali reševanju predoločenega sistema z matriko $\boldsymbol{A}$ si lahko pomagamo tudi z Givensovimi rotacijami. Osnovna ideja je, da z ravninsko rotacijo matriko preoblikujemo tako, da izničimo enega izmed njenih elementov pod diagonalo. Če želimo na primer izničiti element $a_{i,j}$ matrike $\boldsymbol{A}$ na mestu (i, j), $i > j$, lahko to napravimo tako, da vektor $(a_{j,j}, a_{i,j})$, kjer $a_{j,j}$ označuje diagonalni element matrike $\boldsymbol{A}$ na mestu (j, j), rotiramo v vektor $(r, 0)$, $r = \|(a_{j,j}, a_{i,j})\|_2$. To rotacijo opišemo z matriko $\boldsymbol{R}_{j,i}^{\mathsf{T}} \in \mathbb{R}^{m\times m}$, ki je enaka identični matriki povsod, razen v vrsticah i in j. V teh vrsticah so neničelni elementi podani z

$$\boldsymbol{R}_{j,i}^{\mathsf{T}}([j, i], [j, i]) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

pri čemer je $c = a_{j,j}/r$ in $s = a_{i,j}/r$. Vrednosti c in s lahko interpretiramo kot $\cos(\varphi)$ in $\sin(\varphi)$, kjer je φ kot rotacije v negativni smeri (to je v smeri urinega kazalca). Matrika $\boldsymbol{R}_{j,i}^{\mathsf{T}}$ deluje le na vrsticah i in j matrike $\boldsymbol{A}$, elementi v ostalih vrsticah ostanejo nespremenjeni. Zato lahko matriko $\boldsymbol{A}$ z več zaporednimi Givensovimi rotacijami preoblikujemo v zgornje trapezno matriko, v kateri zgornji trikotnik določa zgornjo trikotno matriko iz QR razcepa matrike $\boldsymbol{A}$.

Naloga 5.11. Naj bo $\boldsymbol{a} = (3, -4, 12, -84)$ in $\boldsymbol{b} = (7, -1, 5, 0)$. Z uporabo Givensovih rotacij izračunajte $\min_{x\in\mathbb{R}} \|\boldsymbol{a}x - \boldsymbol{b}\|_2$ in določite $x \in \mathbb{R}$, pri katerem je minimum dosežen.

Rešitev. Konstruiramo tri Givensove rotacije, s katerimi v vektorju $\boldsymbol{a}$ po vrsti izničimo elemente na drugem, tretjem in četrtem mestu.

1. korak: Z rotacijo $\boldsymbol{R}_{1,2}^{\mathsf{T}}$ izničimo element $\boldsymbol{a}$ na drugem mestu.

$$\boldsymbol{R}_{1,2}^{\mathsf{T}} = \begin{bmatrix} \frac{3}{5} & -\frac{4}{5} & & \\ \frac{4}{5} & \frac{3}{5} & & \\ & & 1 & \\ & & & 1 \end{bmatrix} : \quad \begin{aligned} \boldsymbol{a}^{(1)} &= \boldsymbol{R}_{1,2}^{\mathsf{T}}\boldsymbol{a} = (5, 0, 12, -84), \\ \boldsymbol{b}^{(1)} &= \boldsymbol{R}_{1,2}^{\mathsf{T}}\boldsymbol{b} = (5, 5, 5, 0). \end{aligned}$$

2. korak: Z rotacijo $\boldsymbol{R}_{1,3}^{\mathsf{T}}$ izničimo element $\boldsymbol{a}^{(1)}$ na tretjem mestu.

$$\boldsymbol{R}_{1,3}^{\mathsf{T}} = \begin{bmatrix} \frac{5}{13} & & \frac{12}{13} & \\ & 1 & & \\ -\frac{12}{13} & & \frac{5}{13} & \\ & & & 1 \end{bmatrix} : \quad \begin{aligned} \boldsymbol{a}^{(2)} &= \boldsymbol{R}_{1,3}^{\mathsf{T}}\boldsymbol{a}^{(1)} = (13, 0, 0, -84), \\ \boldsymbol{b}^{(2)} &= \boldsymbol{R}_{1,3}^{\mathsf{T}}\boldsymbol{b}^{(1)} = (85/13, 5, -35/13, 0). \end{aligned}$$

3. korak: Z rotacijo $\boldsymbol{R}_{1,4}^{\mathsf{T}}$ izničimo element $\boldsymbol{a}^{(2)}$ na četrtem mestu.

$$\boldsymbol{R}_{1,4}^{\mathsf{T}} = \begin{bmatrix} \frac{13}{85} & & & -\frac{84}{85} \\ & 1 & & \\ & & 1 & \\ \frac{84}{85} & & & \frac{13}{85} \end{bmatrix} : \quad \begin{aligned} \boldsymbol{a}^{(3)} &= \boldsymbol{R}_{1,4}^{\mathsf{T}}\boldsymbol{a}^{(2)} = (85, 0, 0, 0), \\ \boldsymbol{b}^{(3)} &= \boldsymbol{R}_{1,4}^{\mathsf{T}}\boldsymbol{b}^{(2)} = (1, 5, -35/13, 84/13). \end{aligned}$$

Ker so matrike rotacij ortogonalne matrike, je

$$\|\boldsymbol{a}x - \boldsymbol{b}\|_2 = \left\|\boldsymbol{R}_{1,4}^{\mathsf{T}}\boldsymbol{R}_{1,3}^{\mathsf{T}}\boldsymbol{R}_{1,2}^{\mathsf{T}}(\boldsymbol{a}x - \boldsymbol{b})\right\|_2 = \left\|\boldsymbol{a}^{(3)}x - \boldsymbol{b}^{(3)}\right\|_2 .$$

Sledi, da je $\min_{x\in\mathbb{R}} \|\boldsymbol{a}x - \boldsymbol{b}\|_2$ dosežen pri rešitvi enačbe $85x = 1$, to je $x = 1/85$. Minimum ustreza normi vektorja $(5, -35/13, 84/13)$, ki je enaka $\sqrt{74}$.

Naloga 5.12. Sestavite postopek za reševanje sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$ s tridiagonalno matriko

$$\boldsymbol{A} = \begin{bmatrix} a_1 & b_1 & & & \\ c_1 & a_2 & b_2 & & \\ & \ddots & \ddots & \ddots & \\ & & c_{n-2} & a_{n-1} & b_{n-1} \\ & & & c_{n-1} & a_n \end{bmatrix} \in \mathbb{R}^{n\times n},$$

ki temelji na Givensovih rotacijah, in ga implementirajte v Matlabu. Koliko osnovnih operacij je potrebnih za izračun $\boldsymbol{x}$?

Rešitev. Razširjeno matriko $[\boldsymbol{A}\ \boldsymbol{z}]$ želimo z Givensovimi rotacijami preoblikovati tako, da bo na mestu matrike $\boldsymbol{A}$ zgornje trikotna matrika. Reševanje sistema, ki ustreza postopku izračuna QR razcepa razširjene matrike, lahko torej zastavimo

tako, da najprej določimo rotacijo $\boldsymbol{R}_{1,2}^{\mathsf{T}}$, s katero izničimo element c_1 na mestu $(2,1)$. Spremenita se prva in druga vrstica razširjene matrike, ostale vrstice ostanejo nespremenjene. Zato lahko nadaljujemo z rotacijami $\boldsymbol{R}_{j,j+1}^{\mathsf{T}}$, $j = 2, 3, \ldots, n-1$, ki jih po enakem razmisleku določimo tako, da izničijo elemente c_j na mestih $(j+1, j)$.

Pri implementaciji metode za reševanje sistema $\boldsymbol{Ax} = \boldsymbol{z}$ v Matlabu pazimo tako na računsko kot prostorsko učinkovitost. Metoda sprejme matriko $\boldsymbol{A}$ v obliki treh seznamov: seznama diagonalnih elementov $\boldsymbol{a}$ ter seznamov $\boldsymbol{b}$ in $\boldsymbol{c}$, ki predstavljata naddiagonalo in poddiagonalo matrike $\boldsymbol{A}$. Pri transformaciji matrike $\boldsymbol{A}$ v zgornje trikotno ne konstruiramo matrik Givensovih rotacij, temveč v vsakem izmed $n-1$ korakov določimo le c in s ter popravimo 5 elementov matrike (v koraku $n-1$ samo 3). Zgornje trikotna matrika $\boldsymbol{R}$, ki v razširjeni matriki na koncu nadomesti matriko $\boldsymbol{A}$, ima neničelne elemente le na diagonali in na dveh naddiagonalah, zato jo lahko predstavimo s seznami $\boldsymbol{r}_1$, $\boldsymbol{r}_2$ in $\boldsymbol{r}_3$, ki so po vrsti dolžin n, $n-1$ in $n-2$. Na vsakem koraku postopka na podoben način kot elemente spreminjajoče se matrike $\boldsymbol{A}$ popravimo tudi dva elementa v vektorju $\boldsymbol{z}$.

```
n = length(a);

% diagonala in naddiagonali matrike R
r1 = a;
r2 = b;
r3 = zeros(n-2,1);

for j = 1:n-1
    r = sqrt(r1(j)^2 + c(j)^2);
    if r > 0
        C = r1(j)/r; S = c(j)/r;

        % transformacija matrike sistema
        r1(j) = r;
        r1(j+1) = -S*r2(j) + C*a(j+1);
        r2(j) = C*r2(j) + S*a(j+1);
        if j < n-1
            r2(j+1) = C*b(j+1);
            r3(j) = S*b(j+1);
        end

        % transformacija desne strani sistema
        z([j j+1]) = [C S; -S C]*z([j j+1]);
    end
end
```

Na koncu iz razširjene matrike $[\boldsymbol{A}\ \boldsymbol{z}]$ dobimo razširjeno matriko $[\boldsymbol{R}\ \widetilde{\boldsymbol{z}}]$. Ortogonalna matrika $\boldsymbol{Q}$ iz QR razcepa ustreza transponirani matriki, ki je produkt uporabljenih Givensovih rotacij, a je ni treba eksplicitno izračunati. Matrika $\boldsymbol{Q}$ podaja povezavo

med razširjenima matrikama $[\boldsymbol{A}\ \boldsymbol{z}]$ in $[\boldsymbol{R}\ \widetilde{\boldsymbol{z}}]$, saj velja $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ in $\boldsymbol{z} = \boldsymbol{Q}\widetilde{\boldsymbol{z}}$. Rešitev sistema $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{z}$ lahko tako izračunamo z reševanjem sistema $\boldsymbol{R}\boldsymbol{x} = \widetilde{\boldsymbol{z}}$ z obratno subsitucijo.

```
x = zeros(n,1);

x(n) = z(n)/r1(n);

if n > 1
    x(n-1) = (z(n-1) - r2(n-1)*x(n))/r1(n-1);
end

for i = n-2:-1:1
    x(i) = (z(i) - r2(i)*x(i+1) - r3(i)*x(i+2))/r1(i);
end
```

Preštejemo, da je za celoten postopek izračuna $\boldsymbol{x}$ po zgornji implementaciji potrebnih $25n-28$ osnovnih računskih operacij ($n > 1$). Tudi prostorska zahtevnost je linearna glede na n.

Naloga 5.13. Dan je QR razcep $\boldsymbol{A} = \boldsymbol{Q}\boldsymbol{R}$ matrike $\boldsymbol{A} \in \mathbb{R}^{m\times n}$, $m \geq n$. S pomočjo Givensovih rotacij sestavite učinkovit postopek za izračun QR razcepa matrike

$$\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{a}^\mathsf{T} \\ \boldsymbol{A} \end{bmatrix}, \qquad \boldsymbol{a} \in \mathbb{R}^n.$$

Preštejte število operacij, ki jih zahteva predlagani postopek.

Rešitev. Matriko $\boldsymbol{A}'$ zapišemo kot

$$\boldsymbol{A}' = \begin{bmatrix} \boldsymbol{a}^\mathsf{T} \\ \boldsymbol{Q}\boldsymbol{R} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & \\ & \boldsymbol{Q} \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} \boldsymbol{a}^\mathsf{T} \\ \boldsymbol{R} \end{bmatrix}}_{\boldsymbol{H}}, \qquad \boldsymbol{U} \in \mathbb{R}^{(m+1)\times(n+1)},\ \boldsymbol{H} \in \mathbb{R}^{(n+1)\times n}.$$

Matrika, označena z $\boldsymbol{U}$, je sestavljena iz ortogonalnih stolpcev. Matrika $\boldsymbol{H}$ je zgornja Hessenbergova (to pomeni, da so vsi elementi pod diagonalo matrike enaki nič, razen tistih, ki se nahajajo neposredno pod diagonalo) in jo lahko z uporabo Givensovih rotacij transformiramo v zgornje trapezno. Naj bo $\boldsymbol{H}^{(0)} = \boldsymbol{H}$ in

$$\boldsymbol{H}^{(j)} = \boldsymbol{R}_{j,j+1}^\mathsf{T} \boldsymbol{H}^{(j-1)}, \quad j = 1, 2, \ldots, n,$$

kjer je $\boldsymbol{R}_{j,j+1}^\mathsf{T} \in \mathbb{R}^{(n+1)\times(n+1)}$ Givensova rotacija, ki v matriki $\boldsymbol{H}^{(j-1)}$ izniči element na mestu $(j+1, j)$. Pri $j = n$ dobimo

$$\boldsymbol{H}^{(n)} = \boldsymbol{R}_{n,n+1}^\mathsf{T} \ldots \boldsymbol{R}_{2,3}^\mathsf{T} \boldsymbol{R}_{1,2}^\mathsf{T} \boldsymbol{H} = \begin{bmatrix} \boldsymbol{R}' \\ 0 \end{bmatrix},$$

kjer $\boldsymbol{R}' \in \mathbb{R}^{n\times n}$ označuje zgornje trikotno matriko z neničelnimi diagonalnimi elementi. Vpeljimo še matrike $\boldsymbol{U}^{(0)} = \boldsymbol{U}$ in

$$\boldsymbol{U}^{(j)} = \boldsymbol{U}^{(j-1)}\boldsymbol{R}_{j,j+1}, \quad j = 1, 2, \ldots, n.$$

Pri $j = n$ velja

$$\boldsymbol{U}^{(n)} = \boldsymbol{U}\boldsymbol{R}_{1,2}\boldsymbol{R}_{2,3}\ldots\boldsymbol{R}_{n,n+1} = \begin{bmatrix} \boldsymbol{Q}' & \boldsymbol{q}' \end{bmatrix},$$

kjer $\boldsymbol{Q}' \in \mathbb{R}^{(m+1)\times n}$ označuje ortogonalno matriko in $\boldsymbol{q}' \in \mathbb{R}^{m+1}$ vektor, ki je pravokoten na stolpce matrike $\boldsymbol{Q}'$. Zaradi ortogonalnosti Givensovih rotacij velja

$$\boldsymbol{A}' = \boldsymbol{U}\boldsymbol{H} = \boldsymbol{U}^{(n)}\boldsymbol{H}^{(n)} = \boldsymbol{Q}'\boldsymbol{R}',$$

torej matriki $\boldsymbol{Q}'$ in $\boldsymbol{R}'$ določata QR razcep matrike $\boldsymbol{A}'$.

Povzemimo postopek izračuna matrik $\boldsymbol{Q}'$ in $\boldsymbol{R}'$ ter preštejmo število računskih operacij, potrebnih za njegovo izvedbo. Za vsak $j \in \{1, 2, \ldots, n\}$ pripravimo Givensovo rotacijo $\boldsymbol{R}^{\mathsf{T}}_{j,j+1}$, ki je določena z vrednostma c in s v vrsticah in stolpcih j in $j+1$:

$$\boldsymbol{R}^{\mathsf{T}}_{j,j+1}([j, j+1], [j, j+1]) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix}.$$

Izračun vrednosti c in s terja 6 operacij. Rotacijo uporabimo za izračun matrike $\boldsymbol{H}^{(j)}$, ki jo dobimo z

$$\boldsymbol{H}^{(j)}([j, j+1], j : n) = \begin{bmatrix} c & s \\ -s & c \end{bmatrix} \boldsymbol{H}^{(j-1)}([j, j+1], j : n),$$

za kar je potrebnih $6(n-j)$ operacij. Poleg tega s to rotacijo izračunamo tudi $\boldsymbol{U}^{(j)}$ po formuli

$$\boldsymbol{U}^{(j)}(1 : (m+1), [j, j+1]) = \boldsymbol{U}^{(j-1)}(1 : (m+1), [j, j+1]) \begin{bmatrix} c & -s \\ s & c \end{bmatrix},$$

za kar je potrebnih $6(m+1)$ operacij. Celoten postopek torej zahteva

$$\sum_{j=1}^{n} (6 + 6(n-j) + 6(m+1)) = 6mn + 3n^2 + 9n$$

operacij.

Pri uporabi Givensovih rotacij se nakazuje alternativna oblika QR razcepa, tako imenovani razširjeni QR razcep $\boldsymbol{A} = \widetilde{\boldsymbol{Q}}\widetilde{\boldsymbol{R}}$ matrike $\boldsymbol{A} \in \mathbb{R}^{m\times n}$. Pri tej obliki je $\widetilde{\boldsymbol{Q}} \in \mathbb{R}^{m\times m}$ ortogonalna matrika s stolpci, ki določajo ortonormirano bazo za prostor $\mathbb{R}^m$, $\widetilde{\boldsymbol{R}} \in \mathbb{R}^{m\times n}$ pa zgornja trapezna matrika. Ko rešujemo predoločen sistem $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$, matrike $\widetilde{\boldsymbol{Q}}$ običajno ne izračunamo, je pa implicitno konstruirana v postopku reševanja. Kljub temu je postopek z Givensovimi rotacijami računsko zahtevnejši od Gram–Schmidtovega postopka, razen kadar ima matrika $\boldsymbol{A}$ veliko ničelnih elementov. Lahko pa namesto rotacij uporabimo zrcaljenja, ki izničijo več elementov matrike naenkrat.

Predoločen sistem $\boldsymbol{Ax} = \boldsymbol{b}$ s pomočjo Householderjevih zrcaljenj rešujemo tako, da razširjeno matriko $[\boldsymbol{A}\ \boldsymbol{b}]$ z leve množimo z ortogonalnimi transformacijami. V koraku $k \in \{1, 2, \ldots, n\}$ izničimo vse elemente stolpca trenutne matrike z indeksom k, ki ležijo pod diagonalnim elementom. Na ta način postopno dobimo zgornje trapezno matriko, ki predstavlja razširjeno zgornje trikotno matriko iz QR razcepa, medtem ko kompozicija transformacij določa ortogonalno matriko. Posamezno ortogonalno transformacijo definiramo kot zrcaljenje čez ustrezno izbrano hiperravnino. Če želimo v vektorju $\boldsymbol{a} = (a_1, a_2, \ldots, a_p)$, $p > 1$, izničiti vse, razen prve komponente, ga prezrcalimo čez $\operatorname{im}(\boldsymbol{w})^{\perp}$, kjer je

$$\boldsymbol{w} = (a_1 + \operatorname{sign}(a_1)\,\|\boldsymbol{a}\|_2\,, a_2, \ldots, a_p).$$

Matrika zrcaljenja je podana kot

$$\boldsymbol{P} = \boldsymbol{I} - \frac{2}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}}$$

in je običajno ni treba eksplicitno izračunati, saj za transformacijo $\boldsymbol{Px}$ vektorja $\boldsymbol{x}$ velja

$$\boldsymbol{Px} = \boldsymbol{x} - \frac{2}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}}(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x})\boldsymbol{w}.$$

Vektor $\boldsymbol{a}$ se prezrcali v $\boldsymbol{Pa} = (-\operatorname{sign}(a_1)\,\|\boldsymbol{a}\|_2\,, 0, \ldots, 0)$.

Naloga 5.14. Naj bo

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 & 1 \\ 1 & 2 & -1 \\ 1 & -4 & 1 \\ 1 & -4 & 3 \end{bmatrix}, \qquad \boldsymbol{b} = \begin{bmatrix} 0 \\ 3 \\ 0 \\ 3 \end{bmatrix}.$$

S pomočjo Householderjevih zrcaljenj poiščite rešitev predoločenega sistema $\boldsymbol{Ax} = \boldsymbol{b}$. Kakšna je vrednost $\min_{\boldsymbol{x}\in\mathbb{R}^3} \|\boldsymbol{Ax} - \boldsymbol{b}\|_2$?

Rešitev.

1. korak: Najprej zrcalimo prvi stolpec matrike $\boldsymbol{A}$. Householderjevo zrcaljenje $\boldsymbol{P}_1$ je določeno z vektorjem $\boldsymbol{w}_1 = (3, 1, 1, 1)$. Ker je

$$\begin{aligned}
\boldsymbol{P}_1\boldsymbol{a}_1 &= \boldsymbol{a}_1 - \frac{2}{\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{w}_1}\boldsymbol{w}_1(\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{a}_1) = (-2, 0, 0, 0), && \boldsymbol{a}_1 = (1, 1, 1, 1),\\
\boldsymbol{P}_1\boldsymbol{a}_2 &= \boldsymbol{a}_2 - \frac{2}{\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{w}_1}\boldsymbol{w}_1(\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{a}_2) = (2, 2, -4, -4), && \boldsymbol{a}_2 = (2, 2, -4, -4),\\
\boldsymbol{P}_1\boldsymbol{a}_3 &= \boldsymbol{a}_3 - \frac{2}{\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{w}_1}\boldsymbol{w}_1(\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{a}_3) = (-2, -2, 0, 2), && \boldsymbol{a}_3 = (1, -1, 1, 3),\\
\boldsymbol{P}_1\boldsymbol{a}_4 &= \boldsymbol{a}_4 - \frac{2}{\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{w}_1}\boldsymbol{w}_1(\boldsymbol{w}_1^{\mathsf{T}}\boldsymbol{a}_4) = (-3, 2, -1, 2), && \boldsymbol{a}_4 = (0, 3, 0, 3),
\end{aligned}$$

se prvotni predoločeni sistem preoblikuje v

$$\begin{bmatrix} -2 & 2 & -2 \\ 0 & 2 & -2 \\ 0 & -4 & 0 \\ 0 & -4 & 2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} -3 \\ 2 \\ -1 \\ 2 \end{bmatrix}$$

2. korak: Zrcalimo vektor $(2, -4, -4)$. Householderjevo zrcaljenje $\boldsymbol{P}_2$ je določeno z vektorjem $\boldsymbol{w}_2 = (8, -4, -4)$. Iz

$$\begin{aligned} \boldsymbol{P}_2\boldsymbol{a}_1 &= \boldsymbol{a}_1 - \frac{2}{\boldsymbol{w}_2^\mathsf{T}\boldsymbol{w}_2}\boldsymbol{w}_2(\boldsymbol{w}_2^\mathsf{T}\boldsymbol{a}_1) = (-6, 0, 0), & \boldsymbol{a}_1 &= (2, -4, -4), \\ \boldsymbol{P}_2\boldsymbol{a}_2 &= \boldsymbol{a}_2 - \frac{2}{\boldsymbol{w}_2^\mathsf{T}\boldsymbol{w}_2}\boldsymbol{w}_2(\boldsymbol{w}_2^\mathsf{T}\boldsymbol{a}_2) = (2, -2, 0), & \boldsymbol{a}_2 &= (-2, 0, 2), \\ \boldsymbol{P}_2\boldsymbol{a}_3 &= \boldsymbol{a}_3 - \frac{2}{\boldsymbol{w}_2^\mathsf{T}\boldsymbol{w}_2}\boldsymbol{w}_2(\boldsymbol{w}_2^\mathsf{T}\boldsymbol{a}_3) = (0, 0, 3), & \boldsymbol{a}_3 &= (2, -1, 2), \end{aligned}$$

sledi, da je reševanje prvotnega predoločenega sistema ekvivalentno reševanju sistema

$$\begin{bmatrix} -2 & 2 & -2 \\ 0 & -6 & 2 \\ 0 & 0 & -2 \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} -3 \\ 0 \\ 0 \\ 3 \end{bmatrix}.$$

3. korak: Zrcaljenje v tretjem koraku ni potrebno, saj je matrika predoločenega sistema iz prejšnjega koraka že zgornje trapezna.

Kompozicija zrcaljenj v obratnem vrstnem redu, kot smo jih uporabili na matriki $\boldsymbol{A}$, določa matriko $\widetilde{\boldsymbol{Q}} = [\boldsymbol{Q}\ \boldsymbol{q}]$ razširjenega QR razcepa matrike $\boldsymbol{A}$. Zgornje trapezno matriko, dobljeno v drugem koraku, označimo z $\widetilde{\boldsymbol{R}}$, njen zgornji trikotnik pa z $\boldsymbol{R}$. Ker je

$$\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 = \left\|\widetilde{\boldsymbol{Q}}^\mathsf{T}(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})\right\|_2 = \left\|\widetilde{\boldsymbol{R}}\boldsymbol{x} - \widetilde{\boldsymbol{Q}}^\mathsf{T}\boldsymbol{b}\right\|_2 = \left\|\begin{bmatrix} \boldsymbol{R}\boldsymbol{x} - \boldsymbol{Q}^\mathsf{T}\boldsymbol{b} \\ -\boldsymbol{q}^\mathsf{T}\boldsymbol{b} \end{bmatrix}\right\|_2,$$

je

$$\min_{\boldsymbol{x}\in\mathbb{R}^3} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|_2 = \left|-\boldsymbol{q}^\mathsf{T}\boldsymbol{b}\right| = 3,$$

ta vrednost pa je dosežena pri vektorju $\boldsymbol{x}$, ki je rešitev sistema

$$\begin{bmatrix} -2 & 2 & -2 \\ 0 & -6 & 2 \\ 0 & 0 & -2 \end{bmatrix} \boldsymbol{x} = \begin{bmatrix} -3 \\ 0 \\ 0 \end{bmatrix}.$$

Izračunamo, da je $\boldsymbol{x} = (\frac{3}{2}, 0, 0)$.

6. Lastne vrednosti matrik

Lastna vrednost $\lambda \in \mathbb{C}$ matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ je število, za katerega velja $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$ za nek neničeln vektor $\boldsymbol{x} \in \mathbb{C}^n$. Slednjega imenujemo lastni vektor, dvojici $(\lambda, \boldsymbol{x})$ pa rečemo lastni par matrike $\boldsymbol{A}$.

6.1. Schurova forma

Lastne vrednosti matrike lahko določimo s pomočjo Jordanovega razcepa. V numerični linearni algebri se uporablja njegova računsko stabilnejša alternativa, obstoj katere zagotavlja Schurov izrek: za vsako matriko $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ obstaja unitarna matrika $\boldsymbol{U} \in \mathbb{C}^{n\times n}$ ($\boldsymbol{U}^{\mathsf{H}}\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^{\mathsf{H}} = \boldsymbol{I}$) in zgornja trikotna matrika $\boldsymbol{T} \in \mathbb{C}^{n\times n}$, da je $\boldsymbol{U}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{U} = \boldsymbol{T}$. To pomeni, da lahko matriko $\boldsymbol{A}$ zapišemo v obliki $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^{\mathsf{H}}$, kar imenujemo Schurov razcep matrike $\boldsymbol{A}$, medtem ko matriki $\boldsymbol{T}$ pravimo Schurova forma matrike $\boldsymbol{A}$. Ker sta si $\boldsymbol{A}$ in $\boldsymbol{T}$ podobni, lahko lastne vrednosti $\boldsymbol{A}$ preberemo iz diagonale $\boldsymbol{T}$.

Naloga 6.1. Naj bo λ enostavna lastna vrednost matrike $\boldsymbol{A}$. Ugotovite, kako lahko s pomočjo Schurove forme $\boldsymbol{T}$ izračunamo pripadajoči lastni vektor. Postopek demonstrirajte z izračunom lastnega vektorja za najmanjšo lastno vrednost matrike $\boldsymbol{A}$, ki jo določata matriki Schurovega razcepa

$$\boldsymbol{U} = \frac{1}{3}\begin{bmatrix} 2 & -2 & 0 & 1 \\ 1 & 2 & 0 & 2 \\ 0 & 0 & 3 & 0 \\ 2 & 1 & 0 & -2 \end{bmatrix}, \qquad \boldsymbol{T} = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 2 & 1 & -1 \\ 0 & 0 & -1 & 2 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

Rešitev. Matrika $\boldsymbol{T}$ je zgornja trikotna, zato elementi na diagonali ustrezajo njenim lastnim vrednostim. Ker je matrika $\boldsymbol{A}$ podobna matriki $\boldsymbol{T}$, so to tudi lastne vrednosti $\boldsymbol{A}$. Lastni vektor $\boldsymbol{x}$ za matriko $\boldsymbol{A}$, ki pripada lastni vrednosti λ, lahko izrazimo v odvisnosti od lastnega vektorja $\boldsymbol{y}$ za matriko $\boldsymbol{T}$, ki pripada λ, kot $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y}$. Premislimo torej, kako poiskati lastni vektor $\boldsymbol{y}$.

Denimo, da λ ustreza i-temu elementu $t_{i,i}$ na diagonali matrike $\boldsymbol{T}$. Pišimo

$$\boldsymbol{T} = \begin{bmatrix} \boldsymbol{T}_{1,1} & \boldsymbol{T}_{1,2} & \boldsymbol{T}_{1,3} \\ & t_{i,i} & \boldsymbol{T}_{2,3} \\ & & \boldsymbol{T}_{3,3} \end{bmatrix}, \qquad \boldsymbol{y} = \begin{bmatrix} \boldsymbol{y}_1 \\ y_2 \\ \boldsymbol{y}_3 \end{bmatrix}.$$

Ker je $\boldsymbol{T}\boldsymbol{y} = \lambda\boldsymbol{y}$, mora veljati

$$(\boldsymbol{T}_{1,1} - \lambda\boldsymbol{I})\boldsymbol{y}_1 + \boldsymbol{T}_{1,2}y_2 + \boldsymbol{T}_{1,3}\boldsymbol{y}_3 = \boldsymbol{0}, \qquad \boldsymbol{T}_{2,3}\boldsymbol{y}_3 = \boldsymbol{0}, \qquad (\boldsymbol{T}_{3,3} - \lambda\boldsymbol{I})\boldsymbol{y}_3 = \boldsymbol{0}.$$

Ker je λ enostavna lastna vrednost, je $\boldsymbol{T}_{3,3} - \lambda\boldsymbol{I}$ nesingularna matrika, zato je $\boldsymbol{y}_3 = \boldsymbol{0}$. Ostane nam enačba

$$(\boldsymbol{T}_{1,1} - \lambda\boldsymbol{I})\boldsymbol{y}_1 + \boldsymbol{T}_{1,2}y_2 = \boldsymbol{0}.$$

Ker je lastni vektor $\boldsymbol{y}$ določen le do konstante natačno, lahko vzamemo, da je $y_2 = 1$. Z reševanjem sistema $(\boldsymbol{T}_{1,1} - \lambda\boldsymbol{I})\boldsymbol{y}_1 = -\boldsymbol{T}_{1,2}$ določimo še $\boldsymbol{y}_1$. Vektor $(\boldsymbol{y}_1, 1, \boldsymbol{0})$ je lastni vektor za $\boldsymbol{T}$, ki pripada lastni vrednosti λ.

Za podano Schurovo formo matrike $\boldsymbol{A}$ lahko lastni vektor, ki pripada najmanjši lastni vrednosti -1, dobimo z reševanjem sistema

$$\left(\begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} - (-1) \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right) \boldsymbol{y}_1 = - \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Izračunamo $\boldsymbol{y}_1 = (-7, -2)/6$, kar pomeni, da je $\boldsymbol{y} = (-7, -2, 6, 0)/6$ lastni vektor za matriko $\boldsymbol{T}$, $\boldsymbol{x} = \boldsymbol{U}\boldsymbol{y} = (-10, -11, 18, -16)/18$ pa lastni vektor za matriko $\boldsymbol{A}$, ki pripada lastni vrednosti -1.

Naloga 6.2. Dana je bločna matrika

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{X} \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix}, \qquad \boldsymbol{A}_1 \in \mathbb{R}^{m\times m},\ \boldsymbol{A}_2 \in \mathbb{R}^{n\times n},\ \boldsymbol{X} \in \mathbb{R}^{m\times n}.$$

1. Denimo, da poznamo Schurova razcepa $\boldsymbol{A}_1 = \boldsymbol{U}_1\boldsymbol{T}_1\boldsymbol{U}_1^{\mathsf{H}}$ in $\boldsymbol{A}_2 = \boldsymbol{U}_2\boldsymbol{T}_2\boldsymbol{U}_2^{\mathsf{H}}$ matrik $\boldsymbol{A}_1$ in $\boldsymbol{A}_2$. Določite Schurov razcep matrike $\boldsymbol{A}$ in dokažite, da so lastne vrednosti matrike $\boldsymbol{A}$ enake uniji lastnih vrednosti matrik $\boldsymbol{A}_1$ in $\boldsymbol{A}_2$.

2. Z uporabo Householderjevega zrcaljenja izračunajte zgornjo Hessenbergovo matriko, ki ima obliko matrike $\boldsymbol{A}$ in je podobna matriki

$$\boldsymbol{B} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \end{bmatrix}.$$

Nato določite Schurov razcep matrike $\boldsymbol{B}$.

Rešitev.

1. Opazimo, da je

$$\boldsymbol{A} = \begin{bmatrix} \boldsymbol{U}_1\boldsymbol{T}_1\boldsymbol{U}_1^{\mathsf{H}} & \boldsymbol{X} \\ \boldsymbol{0} & \boldsymbol{U}_2\boldsymbol{T}_2\boldsymbol{U}_2^{\mathsf{H}} \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_2 \end{bmatrix}}_{\boldsymbol{U}} \underbrace{\begin{bmatrix} \boldsymbol{T}_1 & \boldsymbol{U}_1^{\mathsf{H}}\boldsymbol{X}\boldsymbol{U}_2 \\ \boldsymbol{0} & \boldsymbol{T}_2 \end{bmatrix}}_{\boldsymbol{T}} \underbrace{\begin{bmatrix} \boldsymbol{U}_1^{\mathsf{H}} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_2^{\mathsf{H}} \end{bmatrix}}_{\boldsymbol{U}^{\mathsf{H}}}.$$

Ker je $\boldsymbol{T}$ zgornja trikotna matrika in $\boldsymbol{U}$ unitarna matrika, je $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^{\mathsf{H}}$ Schurov razcep matrike $\boldsymbol{A}$. Lastne vrednosti matrike $\boldsymbol{A}$ so enake lastnim vrednostim matrike $\boldsymbol{T}$ in se nahajajo na diagonali matrike $\boldsymbol{T}$. Torej so enake lastnim vrednostim matrik $\boldsymbol{T}_1$ in $\boldsymbol{T}_2$, te pa ustrezajo lastnim vrednostim matrik $\boldsymbol{A}_1$ in $\boldsymbol{A}_2$.

2. Konstruirajmo ortogonalno transformacijo

$$\boldsymbol{Q} = \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix}, \qquad \boldsymbol{P} = \boldsymbol{I} - \frac{2}{\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^{\mathsf{T}},$$

kjer je $\boldsymbol{P}$ Householderjevo zrcaljenje, določeno z vektorjem $\boldsymbol{w} = (1, 0, -1)$, ki poskrbi, da je $\boldsymbol{P} \cdot (0, 0, -1) = (-1, 0, 0)$. Izračunamo, da je

$$\boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q}^{\mathsf{T}} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ -1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1 & 0 \end{bmatrix} \cdot \boldsymbol{Q}^{\mathsf{T}} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & -1 & 0 \end{bmatrix}.$$

Vidimo, da ima matrika $\boldsymbol{Q}\boldsymbol{B}\boldsymbol{Q}^{\mathsf{T}}$ obliko matrike $\boldsymbol{A}$. Ker je

$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \underbrace{\frac{\sqrt{2}}{2}\begin{bmatrix} 1 & i \\ i & 1 \end{bmatrix}}_{\boldsymbol{U}_1} \cdot \underbrace{\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}}_{\boldsymbol{T}_1} \cdot \underbrace{\frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -i \\ -i & 1 \end{bmatrix}}_{\boldsymbol{U}_1^{\mathsf{H}}},$$

je Schurov razcep $\boldsymbol{B} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^{\mathsf{H}}$ matrike $\boldsymbol{B}$ podan z matrikama

$$\boldsymbol{T} = \begin{bmatrix} i & 0 & 0 & 0 \\ 0 & -i & 0 & 0 \\ 0 & 0 & i & 0 \\ 0 & 0 & 0 & -i \end{bmatrix}, \qquad \boldsymbol{U} = \boldsymbol{Q}^{\mathsf{T}} \begin{bmatrix} \boldsymbol{U}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{U}_1 \end{bmatrix} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & i & 0 & 0 \\ 0 & 0 & i & 1 \\ 0 & 0 & 1 & i \\ i & 1 & 0 & 0 \end{bmatrix}.$$

Splošno Schurovo formo lahko nadomestimo z realno Schurovo formo, ki ustreza realni kvazi zgornje trikotni matriki. To je zgornje trikotna matrika, le na diagonali so dovoljeni bloki velikosti 2×2.

Naloga 6.3. Dokažite, da za vsako matriko $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ obstajata ortogonalna matrika $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ in kvazi zgornja trikotna matrika $\boldsymbol{T} \in \mathbb{R}^{n\times n}$, da je $\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{T}$.

Rešitev. Dokazujemo z indukcijo na velikost matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$. Če je $n = 1$, trditev velja, saj lahko izberemo $\boldsymbol{Q} = 1$ in $\boldsymbol{T} = \boldsymbol{A}$. Denimo, da trditev velja za matrike velikosti manjše ali enake $n-1$, in dokažimo, da velja tudi za matrike velikosti n. Naj bo $(\lambda, \boldsymbol{x})$ lastni par matrike $\boldsymbol{A}$. Če je λ realna lastna vrednost, $\boldsymbol{x}$ razpenja realni enodimenzionalni invariantni podprostor $\mathcal{N} = \{\alpha\boldsymbol{x};\, \alpha \in \mathbb{R}\}$ za $\boldsymbol{A}$, saj je $\boldsymbol{A}\boldsymbol{y} \in \mathcal{N}$ za vsak $\boldsymbol{y} \in \mathcal{N}$. Če je po drugi strani λ kompleksna lastna vrednost, je tudi $(\overline{\lambda}, \overline{\boldsymbol{x}})$ lastni par za $\boldsymbol{A}$. Realni in imaginarni del vektorjev $\boldsymbol{x}$ in $\overline{\boldsymbol{x}}$,

$$\boldsymbol{x}_{\mathrm{R}} = \frac{1}{2}\left(\boldsymbol{x} + \overline{\boldsymbol{x}}\right), \qquad \boldsymbol{x}_{\mathrm{I}} = \frac{1}{2i}\left(\boldsymbol{x} - \overline{\boldsymbol{x}}\right),$$

sta vektorja iz prostora $\mathbb{R}^n$, ki razpenjata realni dvodimenzionalni invariantni podprostor $\mathcal{N} = \{\alpha\boldsymbol{x}_{\mathrm{R}} + \beta\boldsymbol{x}_{\mathrm{I}};\, \alpha, \beta \in \mathbb{R}\}$ za $\boldsymbol{A}$. V obeh primerih obstaja taka ortogonalna matrika $\boldsymbol{Q}_1 \in \mathbb{R}^{n\times n}$, da njen prvi oziroma njena prva dva stolpca določata bazo za $\mathcal{N}$. Iz invariantnosti podprostora $\mathcal{N}$ potem sledi

$$\boldsymbol{A}\boldsymbol{Q}_1 = \boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{T}_1 & \times \\ \boldsymbol{0} & \boldsymbol{A}_2 \end{bmatrix},$$

pri čemer je $\boldsymbol{T}_1$ skalar ali matrika velikosti 2×2, matrika $\boldsymbol{A}_2$ pa v odvisnosti od tega velikosti $n-1$ ali $n-2$. Po indukcijski predpostavki obstaja ortogonalna matrika $\boldsymbol{Q}_2$, da je $\boldsymbol{T}_2 = \boldsymbol{Q}_2^{\mathsf{T}}\boldsymbol{A}_2\boldsymbol{Q}_2$ kvazi zgornja trikotna. Iz tega sledi

$$\left(\boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{I} & \\ & \boldsymbol{Q}_2 \end{bmatrix}\right)^{\mathsf{T}} \boldsymbol{A} \left(\boldsymbol{Q}_1 \begin{bmatrix} \boldsymbol{I} & \\ & \boldsymbol{Q}_2 \end{bmatrix}\right) = \begin{bmatrix} \boldsymbol{T}_1 & \times \\ \boldsymbol{0} & \boldsymbol{T}_2 \end{bmatrix},$$

kar potrjuje obstoj razcepa $\boldsymbol{Q}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Q} = \boldsymbol{T}$.

6.2. Potenčna metoda

Običajno lahko kakšno izmed lastnih vrednosti matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ poiščemo z izvedbo iteracije

$$\boldsymbol{x}^{(r+1)} = \boldsymbol{A}\boldsymbol{x}^{(r)}, \quad r = 0, 1, \ldots.$$

Pri tem je $\boldsymbol{x}^{(0)} \in \mathbb{C}^n$ izbran začetni približek za lastni vektor matrike $\boldsymbol{A}$.

Naloga 6.4. Naj bo $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ diagonalizabilna matrika z lastnimi vrednostmi $\lambda_1, \lambda_2, \ldots, \lambda_n$, za katere velja

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|.$$

To pomeni, da je λ_1 dominantna lastna vrednost. Dokažite, da je

$$\lim_{r\to\infty} \frac{\boldsymbol{x}_i^{(r+1)}}{\boldsymbol{x}_i^{(r)}} = \lambda_1,$$

kjer je $i \in \{1, 2, \ldots, n\}$ indeks, pri katerem velja $|\boldsymbol{x}_i^{(r)}| = \|\boldsymbol{x}^{(r)}\|_\infty$.

Rešitev. Naj bodo $(\lambda_i, \boldsymbol{v}_i)$, $i = 1, 2, \ldots, n$, lastni pari matrike $\boldsymbol{A}$, ki so številčeni tako, da velja

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \ldots \geq |\lambda_n|\,.$$

Ker se da matriko $\boldsymbol{A}$ diagonalizirati, so lastni vektorji linearno neodvisni in začetni vektor $\boldsymbol{x}^{(0)}$ lahko razvijemo kot

$$\boldsymbol{x}^{(0)} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_n \boldsymbol{v}_n.$$

To pomeni, da za približek $\boldsymbol{x}^{(r)}$ velja

$$\boldsymbol{x}^{(r)} = \boldsymbol{A}^r \boldsymbol{x}^{(0)} = \alpha_1 \lambda_1^r \boldsymbol{v}_1 + \alpha_2 \lambda_2^r \boldsymbol{v}_2 + \ldots + \alpha_n \lambda_n^r \boldsymbol{v}_n$$

in za vsak $i \in \{1, 2, \ldots, n\}$, pri katerem je $\boldsymbol{x}_i^{(r)} \neq 0$, lahko razmerje $\boldsymbol{x}_i^{(r+1)}/\boldsymbol{x}_i^{(r)}$ zapišemo kot

$$\frac{\boldsymbol{x}_i^{(r+1)}}{\boldsymbol{x}_i^{(r)}} = \lambda_1 \frac{\left(\alpha_1 \boldsymbol{v}_1 + \alpha_2 (\frac{\lambda_2}{\lambda_1})^{r+1} \boldsymbol{v}_2 + \ldots + \alpha_n (\frac{\lambda_n}{\lambda_1})^{r+1} \boldsymbol{v}_n\right)_i}{\left(\alpha_1 \boldsymbol{v}_1 + \alpha_2 (\frac{\lambda_2}{\lambda_1})^{r} \boldsymbol{v}_2 + \ldots + \alpha_n (\frac{\lambda_n}{\lambda_1})^{r} \boldsymbol{v}_n\right)_i}.$$

Ker je λ_1 dominantna lastna vrednost ($|\lambda_1| > |\lambda_2|$), od tod sledi, da razmerja konvergirajo k λ_1, ko gre r proti neskončno.

Naloga 6.5. Denimo, da za lastne vrednosti $\lambda_1, \lambda_2, \ldots, \lambda_n$ diagonalizabilne matrike $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ velja

$$|\lambda_1| = |\lambda_2| > |\lambda_3| \geq \ldots \geq |\lambda_n|$$

in $\lambda_2 = -\lambda_1$. Sestavite postopek za izračun lastnih vrednosti λ_1 in λ_2 ter njunih pripadajočih lastnih vektorjev.

Rešitev. V primeru, ko je $\lambda_2 = -\lambda_1$, lahko na podoben način kot v nalogi 6.4 izpeljemo, da za vsak indeks $i \in \{1, 2, \ldots, n\}$ velja

$$\lim_{r \to \infty} \frac{\boldsymbol{x}_i^{(r+2)}}{\boldsymbol{x}_i^{(r)}} = \lambda_1^2.$$

Če torej vemo, da sta po absolutni vrednosti največji lastni vrednosti matrike $\boldsymbol{A}$ nasprotno predznačeni, ju lahko določimo na podlagi razmerij komponent vsakih dveh drugih zaporednih približkov. Nadalje, ker je

$$\frac{\boldsymbol{x}^{(r)}}{\left\|\boldsymbol{x}^{(r)}\right\|} = \frac{\alpha_1 \boldsymbol{v}_1 + \alpha_2 (-1)^r \boldsymbol{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^r \boldsymbol{v}_3 + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^r \boldsymbol{v}_n}{\left\|\alpha_1 \boldsymbol{v}_1 + \alpha_2 (-1)^r \boldsymbol{v}_2 + \left(\frac{\lambda_3}{\lambda_1}\right)^r \boldsymbol{v}_3 + \ldots + \left(\frac{\lambda_n}{\lambda_1}\right)^r \boldsymbol{v}_n\right\|},$$

je $\boldsymbol{x}^{(r)}$ za dovolj velik r približno enak $\beta_1 \boldsymbol{v}_1 + \beta_2 \boldsymbol{v}_2$ za neki konstanti β_1 in β_2. Za naslednji približek potem velja $\boldsymbol{x}^{(r+1)} = \lambda_1 \beta_1 \boldsymbol{v}_1 - \lambda_1 \beta_2 \boldsymbol{v}_2$. Iz tega izrazimo

$$\boldsymbol{v}_1 = \frac{1}{2\lambda_1 \beta_1} \left(\lambda_1 \boldsymbol{x}^{(r)} + \boldsymbol{x}^{(r+1)}\right), \qquad \boldsymbol{v}_2 = \frac{1}{2\lambda_1 \beta_2} \left(\lambda_1 \boldsymbol{x}^{(r)} - \boldsymbol{x}^{(r+1)}\right).$$

Vektorja $\boldsymbol{v}_1$ in $\boldsymbol{v}_2$ lahko torej določimo tako, da na podlagi λ_1 izračunamo vektorja $\lambda_1 \boldsymbol{x}^{(r)} \pm \boldsymbol{x}^{(r+1)}$ in ju normiramo.

Pri potenčni metodi dominantno lastno vrednost matrike $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ računamo tako, da na vsakem koraku iteracije približek za lastni vektor normiramo:

$$\widetilde{\boldsymbol{x}}^{(r+1)} = \boldsymbol{A}\boldsymbol{x}^{(r)}, \quad \boldsymbol{x}^{(r+1)} = \frac{\widetilde{\boldsymbol{x}}^{(r+1)}}{\left\|\widetilde{\boldsymbol{x}}^{(r+1)}\right\|}, \quad r = 0, 1, \ldots.$$

Pri tem je tudi začetni približek $\boldsymbol{x}^{(0)} \in \mathbb{C}^n$ normiran. Z normiranjem zagotovimo, da pri računanju ne pride do prekoračitve ali podkoračitve. Najboljši približek λ_r za dominantno lastno vrednost pri danem približku $\boldsymbol{x}^{(r)}$ za lastni vektor je Rayleighjev kvocient

$$\rho\left(\boldsymbol{A}, \boldsymbol{x}^{(r)}\right) = \frac{{\boldsymbol{x}^{(r)}}^{\mathsf{T}} \boldsymbol{A}\boldsymbol{x}^{(r)}}{{\boldsymbol{x}^{(r)}}^{\mathsf{T}} \boldsymbol{x}^{(r)}} = {\boldsymbol{x}^{(r)}}^{\mathsf{T}} \boldsymbol{A}\boldsymbol{x}^{(r)}.$$

Naloga 6.6. V Matlabu sestavite funkcijo, ki izvede potenčno metodo za dano matriko $\boldsymbol{A}$ in začetni vektor $\boldsymbol{x}^{(0)}$. Zaustavitveni kriterij za iteracijo naj določata toleranca za drugo normo napake $\boldsymbol{A}\boldsymbol{x}^{(r)} - \lambda_r \boldsymbol{x}^{(r)}$ in maksimalno število korakov. Preizkusite funkcijo z matriko

$$\boldsymbol{A} = \begin{bmatrix} 4 & 1 & -2 & 1 \\ 1 & 3 & -1 & 8 \\ -1 & -1 & 5 & 2 \\ 0 & 3 & 1 & -6 \end{bmatrix}$$

in z vgrajenim ukazom `eig` preverite pravilnost dobljenega približka.

Rešitev. Vhodni podatki implementacije potenčne metode so matrika $\boldsymbol{A}$, začetni približek $\boldsymbol{x}^{(0)}$ ter parametra, ki določata zaustavitveni kriterij. Metoda vrne približek za dominantno lastno vrednost matrike $\boldsymbol{A}$, ki je izračunan kot Rayleighjev kvocient drugega izhodnega podatka: približka za dominantni lastni vektor. Tretji izhodni podatek je število korakov, ki jih je potrebno izvesti, da je izpolnjen zaustavitveni kriterij. Pri implementaciji smo pozorni na to, da v vsakem koraku približek z matriko množimo le enkrat, saj to predstavlja najzahtevnejšo operacijo v metodi.

```
function [e,x,k] = potencna(A,x0,tol,N)

x = x0/norm(x0);
Ax = A*x;
e = x'*Ax;
k = 0;
while norm(Ax-e*x) >= tol && k < N
```

```
    x = Ax/norm(Ax);
    Ax = A*x;
    e = x'*Ax;
    k = k+1;
end

end
```

Metodo preizkusimo s spodnjimi ukazi in ugotovimo, da se zaključi po 146 korakih pri približku -8.3602 za lastno vrednost.

```
x0 = ones(4,1); tol = 1e-12; N = 200;
[e1,~,k] = potencna(A,x0,tol,N);
```

Uporaba vgrajene metode `eig` na matriki $\boldsymbol{A}$ potrdi pravilnost izračuna, saj so približki za lastne vrednosti matrike $\boldsymbol{A}$ enaki -8.3602, 6.6817, 4.8349 in 2.8436.

Naloga 6.7. Prilagodite implementacijo potenčne metode iz naloge 6.6 tako, da lahko namesto matrike sprejme funkcijo množenja vektorja z matriko. Z uporabo množenja z inverzom $\boldsymbol{A}^{-1}$ izračunajte po absolutni vrednosti najmanjšo lastno vrednost dane matrike $\boldsymbol{A}$. S pomočjo ustreznih premikov poskusite s potenčno metodo dobiti še vse ostale lastne vrednosti matrike $\boldsymbol{A}$.

Rešitev. Popravek implementacije metode, ki omogoča množenje približka z inverzom matrike $\boldsymbol{A}$, je preprost; treba je le zamenjati ukaz `A*x` z ukazom `A(x)` ter metodi za prvi vhodni podatek namesto matrike podati funkcijo množenja z matriko. Za definicijo funkcije množenja z inverzom matrike $\boldsymbol{A}$ ni treba izračunati inverza matrike $\boldsymbol{A}$. Računanje $\boldsymbol{A}^{-1}\boldsymbol{x}$ interpretiramo kot reševanje sistema $\boldsymbol{A}\boldsymbol{z} = \boldsymbol{x}$. Ker se tekom iteracije matrika $\boldsymbol{A}$ ne spreminja, vnaprej izračunamo LU razcep $\boldsymbol{P}\boldsymbol{A} = \boldsymbol{L}\boldsymbol{U}$ in na vsakem koraku izvedemo le reševanje sistemov $\boldsymbol{L}\boldsymbol{y} = \boldsymbol{P}\boldsymbol{x}$ in $\boldsymbol{U}\boldsymbol{z} = \boldsymbol{y}$ s premimi in obratnimi substitucijami. Rezultat potenčne metode je recipročna vrednost približka za absolutno najmanjšo lastno vrednost matrike $\boldsymbol{A}$, saj je λ lastna vrednost matrike $\boldsymbol{A}$ natanko tedaj, ko je λ^{-1} lastna vrednost matrike $\boldsymbol{A}^{-1}$. Iskanju absolutno najmanjše lastne vrednosti matrike $\boldsymbol{A}$ z izvajanjem potenčne metode za $\boldsymbol{A}^{-1}$ pravimo inverzna potenčna metoda. S spodnjimi ukazi dobimo lastno vrednost 2.8436 pri prejšnjih nastavitvah vhodnih podatkov v 49 korakih.

```
[L,U,p] = lu(A,'vector');
[e4,~,k] = potencna(@(x)U\(L\x(p)),x0,tol,N);
e4 = 1/e4;
```

Podoben trik kot pri računanju absolutno najmanjše lastne vrednosti uporabimo tudi za računanje srednjih dveh lastnih vrednosti matrike $\boldsymbol{A}$. Za poljubno realno število μ je η lastna vrednost matrike $\boldsymbol{A}-\mu\boldsymbol{I}$ natanko tedaj, ko je η^{-1} lastna vrednost matrike $(\boldsymbol{A}-\mu\boldsymbol{I})^{-1}$. Če torej vzamemo $\mu = 6.5$, bo lastna vrednost $\lambda \approx 6.6817$ matrike $\boldsymbol{A}$ (določili smo jo z uporabo `eig`) z absolutno najmanjšo lastno vrednostjo η matrike $\boldsymbol{A}-\mu\boldsymbol{I}$ povezana z zvezo $\lambda = \eta + \mu$ in jo lahko zato poiščemo z inverzno potenčno

metodo. Na podoben način dobimo lastno vrednost $\lambda \approx 4.8349$ matrike $\boldsymbol{A}$ z uporabo premika $\mu = 5$. Prvi rezultat dobimo v 14, drugega pa v 13 korakih. Korakov je manj kot pri prejšnjih izvedbah metode, ker smo zaradi poznavanja dejanskih lastnih vrednosti matrike $\boldsymbol{A}$ lahko izbrali primerna premika.

```
[L,U,p] = lu(A-6.5*I,'vector');
[e2,~,k] = potencna(@(x)U\(L\x(p)),x0,tol,N);
e2 = 1/e2 + 6.5;

[L,U,p] = lu(A-5*I,'vector');
[e3,~,k] = potencna(@(x)U\(L\x(p)),x0,tol,N);
e3 = 1/e3 + 5;
```

Naloga 6.8. Sestavite metodo, ki s potenčno metodo in Householderjevimi redukcijami poišče vse lastne vrednosti matrike. Predpostavite, da ima vhodna matrika po absolutni vrednosti različne lastne vrednosti. Preizkusite metodo z matriko $\boldsymbol{A}$ iz naloge 6.6.

Rešitev. Pri Householderjevi redukciji s Householderjevim zrcaljenjem matrike izločimo lastno vrednost, ki smo jo izračunali s potenčno metodo, in nadaljujemo z matriko manjše velikosti, ki ima enake lastne vrednosti (z izjemo izločene) kot prvotna matrika. V osnovi je postopek naslednji.

```
n = size(A,1);
A = @(x)A*x;
E = zeros(n,1);
for k = 1:n-1
    % potencna metoda
    [E(k),v] = potencna(A,rand(n-k+1,1),tol,N);

    % Householderjeva redukcija
    A = @(x)redukcija(A,v,x);
end
E(n) = A(1);
```

Postavimo $\boldsymbol{A}_1 = \boldsymbol{A} \in \mathbb{R}^{n\times n}$. V koraku $k \in \{1, 2, \ldots, n-1\}$ s potenčno metodo izračunamo dominantno lastno vrednost λ_k matrike $\boldsymbol{A}_k \in \mathbb{R}^{(n-k+1)\times(n-k+1)}$. Nato funkcijo množenja z matriko popravimo z redukcijo. V manjkajoči metodi `redukcija` se na podlagi dominantnega lastnega vektorja $\boldsymbol{v}_k$ matrike $\boldsymbol{A}_k$ določi vektor $\boldsymbol{w}_k$ Householderjevega zrcaljenja $\boldsymbol{P}_k = \boldsymbol{I} - 2/(\boldsymbol{w}_k^\mathsf{T}\boldsymbol{w}_k)\boldsymbol{w}_k\boldsymbol{w}_k^\mathsf{T}$ z lastnostjo $\boldsymbol{P}_k\boldsymbol{v}_k = \pm\boldsymbol{e}_1$ (kjer je $\boldsymbol{e}_1$ standardni enotski vektor z enko v prvi komponenti), ki zagotavlja, da je

$$\boldsymbol{P}_k\boldsymbol{A}_k\boldsymbol{P}_k = \begin{bmatrix} \lambda_k & \times \\ \boldsymbol{0} & \boldsymbol{A}_{k+1} \end{bmatrix}.$$

Matrika $\boldsymbol{A}_{k+1} \in \mathbb{R}^{(n-k)\times(n-k)}$ ima enake lastne vrednosti kot $\boldsymbol{A}_k$ z izjemo λ_k. Množenje matrike $\boldsymbol{A}_{k+1}$ z vektorjem $\boldsymbol{x} \in \mathbb{R}^{n-k}$ se izvede brez dejanskega izračuna

matrik $\boldsymbol{P}_k$ in $\boldsymbol{A}_{k+1}$ na podlagi opazke

$$\boldsymbol{P}_k\boldsymbol{A}_k\boldsymbol{P}_k \begin{bmatrix} 0 \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \lambda_k & \times \\ \boldsymbol{0} & \boldsymbol{A}_{k+1} \end{bmatrix} \begin{bmatrix} 0 \\ \boldsymbol{x} \end{bmatrix} = \begin{bmatrix} \times \\ \boldsymbol{A}_{k+1}\boldsymbol{x} \end{bmatrix}.$$

Absolutno najmanjša lastna vrednost matrike $\boldsymbol{A}$ je kar $\boldsymbol{A}_n$. Z izvedbo opisanega postopka na matriki iz naloge 6.6 dobimo dobre približke za vse lastne vrednosti.

Naloga 6.9. Dana je realna simetrična matrika $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ z lastnimi pari $(\lambda_i, \boldsymbol{v}_i)$, $i = 1, 2, \ldots, n$, kjer so $\boldsymbol{v}_i \in \mathbb{R}^n$ normirani vektorji, za lastne vrednosti $\lambda_i \in \mathbb{R}$ pa velja

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|\,.$$

Lastni par $(\lambda_1, \boldsymbol{v}_1)$ lahko izračunamo s potenčno metodo. Obravnavajte lastne pare matrike $\boldsymbol{A} - \lambda_1\boldsymbol{v}_1\boldsymbol{v}_1^\mathsf{T}$ in na podlagi ugotovitev zasnujte postopek za izračun vseh lastnih parov matrike $\boldsymbol{A}$.

Rešitev. Za produkt matrike $\boldsymbol{A}_1 = \boldsymbol{A} - \lambda_1\boldsymbol{v}_1\boldsymbol{v}_1^\mathsf{T}$ in vektorja $\boldsymbol{v}_i$, $i \in \{1, 2, \ldots, n\}$, velja

$$\boldsymbol{A}_1\boldsymbol{v}_i = \boldsymbol{A}\boldsymbol{v}_i - \lambda_1(\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_i)\boldsymbol{v}_1 = \lambda_1(1 - \boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_i)\boldsymbol{v}_1.$$

Obravnavajmo najprej $i = 1$. Ker je $\boldsymbol{v}_1$ normiran vektor, je $1 - \boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_1 = 0$. Torej je $\boldsymbol{A}_1\boldsymbol{v}_1 = \boldsymbol{0} = 0 \cdot \boldsymbol{v}_1$, kar pomeni, da je $(0, \boldsymbol{v}_1)$ lastni par matrike $\boldsymbol{A}_1$. Za $i > 1$ iz simetrije matrike $\boldsymbol{A}$ sledi

$$\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_i = \frac{1}{\lambda_1}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{v}_i = \frac{1}{\lambda_1}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{A}\boldsymbol{v}_i = \frac{\lambda_i}{\lambda_1}\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_i.$$

Lastni vrednosti λ_1 in λ_i sta različni, zato je $\boldsymbol{v}_1^\mathsf{T}\boldsymbol{v}_i = 0$. Potemtakem je $\boldsymbol{A}_1\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$ in $(\lambda_i, \boldsymbol{v}_i)$ je lastni par matrike $\boldsymbol{A}_1$.

Po tem premisleku je $(\lambda_2, \boldsymbol{v}_2)$ dominantni lastni par matrike $\boldsymbol{A}_1$. Izračunamo ga lahko s potenčno metodo. Nato definiramo matriko $\boldsymbol{A}_2 = \boldsymbol{A}_1 - \lambda_2\boldsymbol{v}_2\boldsymbol{v}_2^\mathsf{T}$. Zanjo po zgornjih sklepih velja $\boldsymbol{A}_2\boldsymbol{v}_1 = \boldsymbol{A}_2\boldsymbol{v}_2 = \boldsymbol{0}$ ter $\boldsymbol{A}_2\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i$, $i = 3, 4, \ldots, n$. Lastni par $(\lambda_3, \boldsymbol{v}_3)$ izračunamo s potenčno metodo in enak postopek nadaljujemo rekurzivno z matrikami $\boldsymbol{A}_i = \boldsymbol{A}_{i-1} - \lambda_i\boldsymbol{v}_i\boldsymbol{v}_i^\mathsf{T}$. Na ta način dobimo vse lastne pare matrike $\boldsymbol{A}$. Ta postopek se imenuje Hotellingova redukcija.

Naloga 6.10. Naj bo λ enostavna lastna vrednost matrike $\boldsymbol{A}$, ki ji pripadata (desni) lastni vektor $\boldsymbol{v}$ (velja $\boldsymbol{A}\boldsymbol{v} = \lambda\boldsymbol{v}$) in levi lastni vektor $\boldsymbol{u}$ (velja $\boldsymbol{u}^\mathsf{H}\boldsymbol{A} = \lambda\boldsymbol{u}^\mathsf{H}$). Dokažite, da se lastne vrednosti matrike

$$\boldsymbol{B} = \boldsymbol{A} - \frac{\lambda}{\boldsymbol{u}^\mathsf{H}\boldsymbol{v}}\boldsymbol{v}\boldsymbol{u}^\mathsf{H}$$

ujemajo z lastnimi vrednostmi matrike $\boldsymbol{A}$, le da ima namesto λ lastno vrednost 0. Opišite, kako lahko na podlagi te ugotovite z uporabo potenčne metode izračunamo po absolutni vrednosti drugo največjo lastno vrednost μ matrike $\boldsymbol{A}$ ob predpostavki, da je $|\lambda| > |\mu|$ in da je μ po absolutni vrednosti večja od vseh ostalih lastnih vrednosti matrike $\boldsymbol{A}$. Postopek demonstrirajte v Matlabu na primeru iz naloge 6.6.

Rešitev. Ker je

$$\boldsymbol{B}\boldsymbol{v} = \boldsymbol{A}\boldsymbol{v} - \frac{\lambda}{\boldsymbol{u}^{\mathsf{H}}\boldsymbol{v}}\boldsymbol{v}\boldsymbol{u}^{\mathsf{H}}\boldsymbol{v} = \lambda\boldsymbol{v} - \lambda\boldsymbol{v} = \boldsymbol{0} = 0 \cdot \boldsymbol{v},$$

je 0 lastna vrednost matrike $\boldsymbol{B}$. Naj bo $(\nu, \boldsymbol{w})$ poljuben lastni par matrike $\boldsymbol{A}$, različen od $(\lambda, \boldsymbol{v})$. Tedaj velja

$$\lambda\boldsymbol{u}^{\mathsf{H}}\boldsymbol{w} = \boldsymbol{u}^{\mathsf{H}}\boldsymbol{A}\boldsymbol{w} = \nu\boldsymbol{u}^{\mathsf{H}}\boldsymbol{w}$$

in ker je $\lambda \neq \nu$, je $\boldsymbol{u}^{\mathsf{H}}\boldsymbol{w} = 0$. To pomeni, da je $\boldsymbol{B}\boldsymbol{w} = \nu\boldsymbol{w}$, zato je ν tudi lastna vrednost matrike $\boldsymbol{B}$.

Če je λ dominantna lastna vrednost matrike $\boldsymbol{A}$, lahko s potenčno metodo izračunamo približek za par $(\lambda, \boldsymbol{v})$. Poleg tega lahko s potenčno metodo za matriko $\boldsymbol{A}^{\mathsf{T}}$ izračunamo približek za $\boldsymbol{u}$, saj je $\boldsymbol{u}$ (desni) lastni vektor za $\boldsymbol{A}^{\mathsf{T}}$:

$$\boldsymbol{A}^{\mathsf{T}}\boldsymbol{u} = \left(\boldsymbol{u}^{\mathsf{H}}\boldsymbol{A}\right)^{\mathsf{H}} = (\lambda\boldsymbol{u})^{\mathsf{H}} = \overline{\lambda}\boldsymbol{u}^{\mathsf{H}}.$$

S tem je določena matrika $\boldsymbol{B}$. Če je μ po absolutni vrednosti druga največja lastna vrednost matrike $\boldsymbol{A}$, je μ dominantna lastna vrednost matrike $\boldsymbol{B}$, ki jo lahko izračunamo z novo izvedbo potenčne metode.

Na podlagi teh ugotovitev lahko v Matlabu največjo in drugo največjo lastno vrednost matrike $\boldsymbol{A}$ iz naloge 6.6 izračunamo z naslednjimi ukazi.

```
x0 = ones(4,1); N = 200; tol = 1e-12;

[lam,v] = potencna(A,x0,tol,N);    % lam = 8.3602
[~,u] = potencna(A',x0,tol,N);

s = lam/(u'*v)*v;
Bx = @(x) A*x-(u'*x)*s;
mu = potencna(Bx,x0,tol,N);        % mu = 6.6817
```

Pri izvedbi potenčne metode za matriko $\boldsymbol{B}$ podamo le funkcijo $\boldsymbol{x} \mapsto \boldsymbol{B}\boldsymbol{x}$, ki zahteva manj operacij od izračuna matrike $\boldsymbol{B}$. Vidimo, da s tem postopkom dobimo dober približek za μ. Omeniti velja še, da je ob izračunu natančnejšega približka za λ vektor $\boldsymbol{u}$ učinkoviteje poiskati z inverzno potenčno metodo za matriko $\boldsymbol{A}^{\mathsf{T}} - \lambda\boldsymbol{I}$, saj lahko na ta način približek za $\boldsymbol{u}$ dobimo z bistveno manj koraki iteracije.

Naloga 6.11. Naj bo $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ matrika z lastnimi vrednostmi $\lambda_1, \lambda_2, \ldots, \lambda_n$, za katere velja

$$|\lambda_1| > |\lambda_2| > \ldots > |\lambda_n|.$$

1. Dokažite, da za naključno izbran vektor $\boldsymbol{x} \in \mathbb{R}^n$ Rayleighjevi kvocienti v potenčni metodi z začetnim približkom $(\boldsymbol{A} - \lambda_1\boldsymbol{I})\boldsymbol{x}$ v teoriji konvergirajo k lastni vrednosti λ_2. Kakšne težave gre pričakovati v praksi?

2. Recimo, da s potenčno metodo z začetnim približkom $\boldsymbol{x}$ izračunamo lastno vrednost λ_1 in nato z začetnim približkom $(\boldsymbol{A} - \lambda_1\boldsymbol{I})\boldsymbol{x}$ še lastno vrednost λ_2. Kako bi v slogu prve točke nadaljevali postopek in izračunali preostale lastne vrednosti matrike $\boldsymbol{A}$?

Rešitev.

1. Po predpostavki so lastne vrednosti paroma različne, zato so pripadajoči lastni vektorji $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_n$ linearno neodvisni. Naj bodo $\alpha_1, \alpha_2, \ldots, \alpha_n$ take realne vrednosti, da je
$$\boldsymbol{x} = \alpha_1 \boldsymbol{v}_1 + \alpha_2 \boldsymbol{v}_2 + \ldots + \alpha_n \boldsymbol{v}_n.$$
Za začetni približek $\boldsymbol{x}_0 = (\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{x}$ potem velja
$$(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{x} = (\lambda_2 - \lambda_1)\alpha_2 \boldsymbol{v}_2 + \ldots + (\lambda_n - \lambda_1)\alpha_n \boldsymbol{v}_n.$$
Pišimo $\beta_i = (\lambda_i - \lambda_1)\alpha_i$, $i = 2, 3, \ldots, n$. Ker je $\boldsymbol{x}$ naključno izbran, lahko predpostavimo, da je $\alpha_2 \neq 0$ oziroma $\beta_2 \neq 0$. Za približek $\boldsymbol{x}_k$ na k-tem koraku potenčne metode potem velja
$$\boldsymbol{x}_k = \frac{\boldsymbol{A}^k \boldsymbol{x}_0}{\|\boldsymbol{A}^k \boldsymbol{x}_0\|} = \frac{\beta_2 \boldsymbol{v}_2 + \beta_2 (\frac{\lambda_3}{\lambda_2})^k \boldsymbol{v}_3 + \ldots + \beta_n (\frac{\lambda_n}{\lambda_2})^k \boldsymbol{v}_n}{\left\| \beta_2 \boldsymbol{v}_2 + \beta_2 (\frac{\lambda_3}{\lambda_2})^k \boldsymbol{v}_3 + \ldots + \beta_n (\frac{\lambda_n}{\lambda_2})^k \boldsymbol{v}_n \right\|}$$
in po smeri konvergira k lastnemu vektorju $\boldsymbol{v}_2$, zato Rayleighjevi kvocienti konvergirajo k λ_2. V praksi zaradi numeričnih napak koeficient razvoja vektorja $(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{x}$ pri vektorju $\boldsymbol{v}_1$ ni enak 0, zato lahko Rayleighjevi kvocienti v potenčni metodi sčasoma skonvergirajo k lastni vrednosti λ_1.

2. Če za začetni približek pri potenčni metodi vzamemo $(\boldsymbol{A} - \lambda_2 \boldsymbol{I})(\boldsymbol{A} - \lambda_1 \boldsymbol{I})\boldsymbol{x}$, po podobnem razmisleku kot v prvi točki Rayleighjevi kvocienti konvergirajo k λ_3. Če torej po izračunu lastne vrednosti λ_k začetni približek, uporabljen pri izvedbi potenčne metode, pomnožimo z $(\boldsymbol{A} - \lambda_k \boldsymbol{I})$, z naslednjo izvedbo potenčne metode dobimo λ_{k+1}. Na ta način nadaljujemo, dokler ne dobimo lastne vrednosti λ_n.

6.3. QR iteracija

Pri QR iteraciji lastne vrednosti matrike $\boldsymbol{A}$ izračunamo tako, da matriko z zaporedjem iteracij transformiramo v realno Schurovo formo, iz katere lahko razberemo njene lastne vrednosti. To navadno napravimo tako, da $\boldsymbol{A}$ najprej s podobnostno transformacijo reduciramo v zgornjo Hessenbergovo matriko $\boldsymbol{H}$. Nato postavimo $\boldsymbol{A}_0 = \boldsymbol{H}$ in iterativno izračuamo
$$\boldsymbol{A}_{k+1} = \boldsymbol{R}_k \boldsymbol{Q}_k + \sigma_k \boldsymbol{I}, \qquad k = 0, 1, \ldots,$$
kjer sta $\boldsymbol{Q}_k$ in $\boldsymbol{R}_k$ matriki QR razcepa $\boldsymbol{A}_k - \sigma_k \boldsymbol{I} = \boldsymbol{Q}_k \boldsymbol{R}_k$ za izbran premik σ_k. Redukcijo na Hessenbergovo matriko uporabljamo, da zmanjšamo časovno zahtevnost koraka QR iteracije, premike σ_k pa uvedemo, da pospešimo hitrost konvergence matrik $\boldsymbol{A}_k$ k realni Schurovi formi matrike $\boldsymbol{A}$.

Naloga 6.12. Naj bo

$$\boldsymbol{A} = \begin{bmatrix} 6 & 1 & 1 & 1 \\ 1 & 6 & 1 & 1 \\ 1 & 1 & 6 & 1 \\ 1 & 1 & 1 & 6 \end{bmatrix}.$$

1. Z uporabo Householderjevih zrcaljenj reducirajte matriko $\boldsymbol{A}$ v zgornjo Hessenbergovo matriko $\boldsymbol{H}$.
2. Napravite korak QR iteracije za $\boldsymbol{H}$ s premikom $\sigma_0 = 5$. QR razcep izračunajte z Givensovo rotacijo.

Rešitev.

1. Poskrbimo, da se zadnja dva elementa v prvem stolpcu matrike $\boldsymbol{A}$ spremenita v 0. To lahko dosežemo s Householderjevim zrcaljenjem

$$\boldsymbol{P} = \boldsymbol{I} - \frac{2}{\boldsymbol{w}^\mathsf{T}\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}^\mathsf{T}, \qquad \boldsymbol{w} = (1+\sqrt{3}, 1, 1).$$

Z nekaj računanja v enem koraku redukcije dobimo Hessenbergovo matriko

$$\boldsymbol{H} = \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix} \cdot \boldsymbol{A} \cdot \begin{bmatrix} 1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{P} \end{bmatrix}^\mathsf{T} = \begin{bmatrix} 6 & -\sqrt{3} & 0 & 0 \\ -\sqrt{3} & 8 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 5 \end{bmatrix}.$$

Ker je matrika $\boldsymbol{A}$ simetrična, je matrika $\boldsymbol{H}$ po pričakovanjih tridiagonalna.

2. Hessenbergova matrika $\boldsymbol{H}$ iz prejšnje točke je razcepna. Sklepamo lahko, da je 5 vsaj dvojna lastna vrednost matrike $\boldsymbol{H}$, preostali pa sta določeni s podmatriko

$$\boldsymbol{A}_0 = \begin{bmatrix} 6 & -\sqrt{3} \\ -\sqrt{3} & 8 \end{bmatrix}.$$

S pomočjo Givensove rotacije določimo QR razcep matrike $\boldsymbol{A}_0 - 5\boldsymbol{I}$. Rotacija, ki izniči drugi element v stolpcu $(1, -\sqrt{3})$, je podana z

$$\boldsymbol{Q}_0^\mathsf{T} = \frac{1}{2}\begin{bmatrix} 1 & -\sqrt{3} \\ \sqrt{3} & 1 \end{bmatrix}.$$

Velja

$$\boldsymbol{R}_0 = \boldsymbol{Q}_0^\mathsf{T}(\boldsymbol{A}_0 - 5\boldsymbol{I}) = \begin{bmatrix} 2 & -2\sqrt{3} \\ 0 & 0 \end{bmatrix}$$

in kot rezultat prvega koraka QR iteracije dobimo

$$\boldsymbol{A}_1 = \boldsymbol{R}_0\boldsymbol{Q}_0 + 5\boldsymbol{I} = \begin{bmatrix} 4 & 0 \\ 0 & 0 \end{bmatrix} + \begin{bmatrix} 5 & 0 \\ 0 & 5 \end{bmatrix} = \begin{bmatrix} 9 & 0 \\ 0 & 5 \end{bmatrix}.$$

Izkazalo se je, da premik $\sigma_0 = 5$ ustreza lastni vrednosti matrike $\boldsymbol{A}_0$, zato smo že po prvem koraku QR iteracije dobili obe lastni vrednosti matrike $\boldsymbol{A}_0$: 5 in 9. Zaključimo, da je 5 trojna, 9 pa enojna lastna vrednost matrike $\boldsymbol{A}$.

Naloga 6.13. Naj bo $\boldsymbol{A}$ simetrična pozitivno definitna matrika.

1. Utemeljite, da obstaja matrika $\sqrt{\boldsymbol{A}}$, za katero velja $\sqrt{\boldsymbol{A}}^2 = \boldsymbol{A}$.
2. Naj bo $\boldsymbol{B}_0 = \sqrt{\boldsymbol{A}}$ in naj bodo $\boldsymbol{B}_k$, $k = 1, 2, \ldots$, matrike, ki jih dobimo z izvajanjem QR iteracije za matriko $\sqrt{\boldsymbol{A}}$. Dokažite, da $\boldsymbol{B}_k^2$ ustreza matriki $\boldsymbol{A}_k$, ki jo dobimo z iteracijo
$$\boldsymbol{A}_0 = \boldsymbol{A}, \qquad \boldsymbol{A}_k = \boldsymbol{V}_{k-1}^{\mathsf{T}} \boldsymbol{V}_{k-1}, \quad k = 1, 2, \ldots,$$
kjer je $\boldsymbol{A}_{k-1} = \boldsymbol{V}_{k-1}\boldsymbol{V}_{k-1}^{\mathsf{T}}$ razcep Choleskega matrike $\boldsymbol{A}_{k-1}$. Iz tega sklepajte, da lahko lastne vrednosti matrike $\boldsymbol{A}$ izračunamo z zgornjo iteracijo, ki QR razcep zamenja z razcepom Choleskega.

Rešitev.

1. Ker je matrika $\boldsymbol{A}$ simetrična, ima realne lastne vrednosti in ortogonalne lastne vektorje, zato jo lahko predstavimo z razcepom $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$, kjer je $\boldsymbol{U}$ ortogonalna matrika s stolpci, ki ustrezajo normiranim lastnim vektorjem matrike $\boldsymbol{A}$, $\boldsymbol{D}$ pa diagonalna matrika, ki je določena z lastnimi vrednostmi matrike $\boldsymbol{A}$. Ker je $\boldsymbol{A}$ pozitivno definitna, so njene lastne vrednosti pozitivne in matriko $\boldsymbol{D}$ lahko razcepimo v $\boldsymbol{D} = \sqrt{\boldsymbol{D}}\sqrt{\boldsymbol{D}}$, kjer je $\sqrt{\boldsymbol{D}}$ diagonalna matrika, v kateri nastopajo koreni lastnih vrednosti matrike $\boldsymbol{A}$. Sedaj lahko iz
$$\left(\boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}^{\mathsf{T}}\right)\left(\boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}^{\mathsf{T}}\right) = \boldsymbol{U}\left(\sqrt{\boldsymbol{D}}\left(\boldsymbol{U}^{\mathsf{T}}\boldsymbol{U}\right)\sqrt{\boldsymbol{D}}\right)\boldsymbol{U}^{\mathsf{T}} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$$
sklepamo, da je $\sqrt{\boldsymbol{A}} = \boldsymbol{U}\sqrt{\boldsymbol{D}}\boldsymbol{U}^{\mathsf{T}}$ matrika z želeno lastnostjo.
2. Pri QR iteraciji za matriko $\sqrt{\boldsymbol{A}}$ v koraku $k \in \{1, 2, \ldots\}$ izračunamo QR razcep $\boldsymbol{B}_{k-1} = \boldsymbol{Q}_{k-1}\boldsymbol{R}_{k-1}$ matrike $\boldsymbol{B}_{k-1}$ in nadaljujemo z matriko $\boldsymbol{B}_k = \boldsymbol{R}_{k-1}\boldsymbol{Q}_{k-1}$. Zanjo velja
$$\boldsymbol{B}_k = \boldsymbol{R}_{k-1}\boldsymbol{Q}_{k-1} = \boldsymbol{Q}_{k-1}^{\mathsf{T}}\boldsymbol{B}_{k-1}\boldsymbol{Q}_{k-1},$$
zato iz dejstva, da je $\boldsymbol{B}_0$ simetrična, po indukciji sledi, da je tudi $\boldsymbol{B}_k$ simetrična. Na podlagi tega lahko sklepamo, da je
$$\boldsymbol{B}_k^2 = \boldsymbol{B}_k\boldsymbol{B}_k^{\mathsf{T}} = \boldsymbol{R}_{k-1}\boldsymbol{Q}_{k-1}\boldsymbol{Q}_{k-1}^{\mathsf{T}}\boldsymbol{R}_{k-1}^{\mathsf{T}} = \boldsymbol{R}_{k-1}\boldsymbol{R}_{k-1}^{\mathsf{T}},$$
in ker je $\boldsymbol{R}_{k-1}$ zgornje trikotna matrika s pozitivnimi diagonalnimi elementi, matrika $\boldsymbol{R}_{k-1}^{\mathsf{T}}$ ustreza (enolično določenemu) faktorju Choleskega za matriko $\boldsymbol{B}_k^2$. Tudi matrika $\boldsymbol{B}_{k-1}$ je simetrična, zato zanjo velja
$$\boldsymbol{B}_{k-1}^2 = \boldsymbol{B}_{k-1}^{\mathsf{T}}\boldsymbol{B}_{k-1} = \boldsymbol{R}_{k-1}^{\mathsf{T}}\boldsymbol{Q}_{k-1}^{\mathsf{T}}\boldsymbol{Q}_{k-1}\boldsymbol{R}_{k-1} = \boldsymbol{R}_{k-1}^{\mathsf{T}}\boldsymbol{R}_{k-1}$$
in po definiciji matrike $\boldsymbol{A}_k$ je $\boldsymbol{A}_k = \boldsymbol{B}_k^2$, kot smo želeli dokazati. Iz lastnosti QR iteracije zaključimo, da zaporedje matrik $\boldsymbol{B}_0, \boldsymbol{B}_1, \boldsymbol{B}_2, \ldots$ konvergira k Schurovi formi za $\sqrt{\boldsymbol{A}}$, torej zaporedje matrik $\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2$ konvergira k Schurovi formi za $\boldsymbol{A}$. Pri tem velja dodati, da je zaporedje matrik $\boldsymbol{A}_0, \boldsymbol{A}_1, \boldsymbol{A}_2, \ldots$ dobro definiramo, saj je $\boldsymbol{A}_k = \boldsymbol{V}_{k-1}^{-1}\boldsymbol{A}_{k-1}\boldsymbol{V}_{k-1}$ za vsak $k > 0$ in so zato po indukciji vse matrike v zaporedju simetrične pozitivno definitne.

7. Polinomska interpolacija

Osnovna naloga polinomske interpolacije je za dano funkcijo f poiskati polinom stopnje $n \in \mathbb{N}_0$, ki se s funkcijo f ujema v $n+1$ paroma različnih točkah. Za tak polinom pravimo, da interpolira funkcijo oziroma da je interpolacijski polinom za f. V širšem kontekstu pa se teorija interpolacije ukvarja s konstrukcijo in analizo polinomov, ki se s funkcijo ujemajo v določenih lastnostih.

7.1. Lagrangeeva oblika

Priročna oblika predstavitve polinoma stopnje n, ki interpolira vrednosti funkcije v paroma različnih točkah x_i, $i = 0, 1, \ldots, n$, temelji na Lagrangeevih baznih polinomih $\ell_{n,i}$. Definirani so s predpisi

$$\ell_{n,i}(x) = \prod_{\substack{k=0 \\ k \neq i}}^{n} \frac{x - x_k}{x_i - x_k}, \qquad i = 0, 1, \ldots, n,$$

iz katerih je razvidno, da so vsi polinomi $\ell_{n,i}$ stopnje n.

Naloga 7.1. Utemeljite, da je $\ell_{n,i}(x_i) = 1$ in $\ell_{n,i}(x_j) = 0$ za vsak $j \neq i$. Na podlagi tega sklepajte, da je

$$p = \sum_{k=0}^{n} f(x_k)\ell_{n,k}$$

polinom stopnje največ n, za katerega velja $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$.

Rešitev. Če izvrednotimo $\ell_{n,i}$ v točki x_i, so vsi členi produkta, ki določa $\ell_{n,i}$, enaki 1, zato je $\ell_{n,i}(x_i) = 1$. Če po drugi strani $\ell_{n,i}$ izvrednotimo v točki x_j, $j \neq i$, je eden izmed členov produkta enak 0, zato je $\ell_{n,i}(x_j) = 0$. Za vrednost polinoma p v poljubni interpolacijski točki x_i torej velja

$$p(x_i) = \sum_{k=0}^{n} f(x_k)\ell_{n,k}(x_i) = f(x_i).$$

Ker so vsi Lagrangeeevi bazni polinomi $\ell_{n,k}$ stopnje n, je tudi p, ki je določen kot njihova linearna kombinacija, največ stopnje n.

Naloga 7.2. Dokažite, da Lagrangeevi bazni polinomi $\ell_{n,i}$, $i = 0, 1, \ldots, n$, sestavljajo bazo prostora polinomov stopnje največ n.

Rešitev. Število Lagrangeevih baznih polinomov $\ell_{n,i}$ je enako $n + 1$, kar ustreza dimenziji prostora polinomov stopnje največ n. Zato zadošča dokaz, da so funkcije $\ell_{n,i}$ linearno neodvisne. Denimo, da za neka realna števila a_k, $k = 0, 1, \ldots, n$, velja

$$\sum_{k=0}^{n} a_k \ell_{n,k} = 0.$$

Ker po nalogi 7.1 za vsak $i \in \{0, 1, \ldots, n\}$ velja $\ell_{n,i}(x_i) = 1$ in $\ell_{n,i}(x_j) = 0$, $j \neq i$, z izračunom vrednosti linearne kombinacije v točki x_i dobimo $a_i = 0$.

Naloga 7.3. Z uporabo Lagrangeevih baznih polinomov določite polinom stopnje največ 3, ki interpolira funkcijo $f(x) = 40/(x + 1)$ v točkah 0, $\frac{1}{3}$, $\frac{2}{3}$, 1. Polinom izrazite tudi v potenčni bazi.

Rešitev. Lagrangeeevi bazni polinomi so za ta primer podani z

$$\ell_{3,0}(x) = -\frac{1}{2}(3x-1)(3x-2)(x-1), \qquad \ell_{3,1}(x) = \frac{9}{2}x(3x-2)(x-1),$$
$$\ell_{3,2}(x) = -\frac{9}{2}x(3x-1)(x-1), \qquad \ell_{3,3}(x) = \frac{1}{2}x(3x-1)(3x-2).$$

Prikazani so na sliki 7.1a. Izračunamo $f(0) = 40$, $f(\frac{1}{3}) = 30$, $f(\frac{2}{3}) = 24$, $f(1) = 20$ in polinom p, ki interpolira funkcijo, predstavimo v obliki

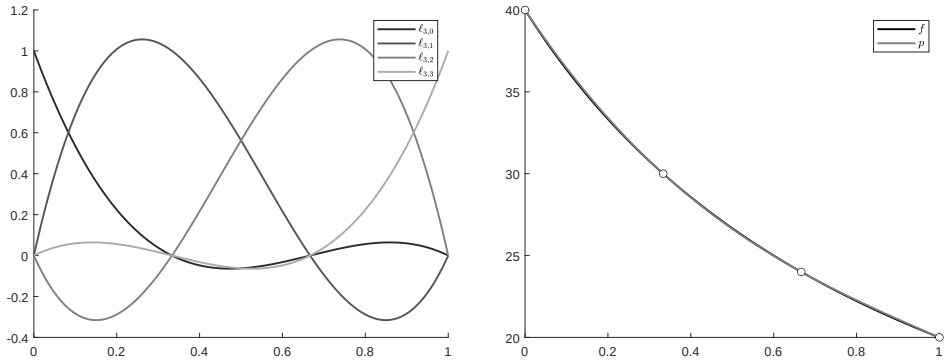
$$p = 40\ell_{3,0} + 30\ell_{3,1} + 24\ell_{3,2} + 20\ell_{3,3}.$$

Z upoštevanjem predpisov za Lagrangeeve bazne polinome lahko p v potenčni bazi izrazimo kot $p(x) = -9x^3 + 27x^2 - 38x + 40$. Polinom p je, skupaj s funkcijo f, prikazan na sliki 7.1b.

Naloga 7.4. Dokažite, da lahko polinom p stopnje največ n, ki interpolira vrednosti funkcije f v točkah $x_0, x_1, \ldots, x_n$, predstavimo v baricentrični obliki

$$p(x) = \frac{\displaystyle\sum_{k=0}^{n} \frac{w_k}{x - x_k} f(x_k)}{\displaystyle\sum_{k=0}^{n} \frac{w_k}{x - x_k}},$$

kjer je $w_k = 1/\omega'(x_k)$ za $\omega(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$. Utemeljite, da lahko ob vnaprej izračunanih vrednostih w_k, $k = 0, 1, \ldots, n$, vrednost polinoma p v poljubni točki x izračunamo z $\mathcal{O}(n)$ osnovnimi računskimi operacijami, neodvisno od vrednosti $f(x_k)$.



(A) *Lagrangeevi bazni polinomi.* (B) *Interpolacijski polinom.*

SLIKA 7.1: *Primer polinomske interpolacije iz naloge 7.3.*

Rešitev. Za odvod polinoma ω velja

$$\omega'(x) = \sum_{k=0}^{n} \prod_{\substack{i=0 \\ i \neq k}}^{n} (x - x_i),$$

torej je

$$w_k = \frac{1}{\prod\limits_{\substack{i=0 \\ i \neq k}} (x_k - x_i)}, \quad k = 0, 1, \ldots, n,$$

in po definiciji Lagrangeevih baznih polinomov lahko p zapišemo v obliki

$$p(x) = \sum_{k=0}^{n} f(x_k) \ell_{n,k}(x) = \sum_{k=0}^{n} \frac{\omega(x)}{x - x_k} w_k f(x_k).$$

Ta zveza velja za vsako funkcijo f, tudi za konstanto $x \mapsto 1$, iz česar sledi enakost

$$1 = \sum_{k=0}^{n} \frac{\omega(x)}{x - x_k} w_k.$$

Ob upoštevanju $p(x) = p(x)/1$ iz teh dveh zvez dobimo iskano obliko polinoma p.

Za izračun vrednosti w_k, $k = 0, 1, \ldots, n$, je potrebnih $(n+1)(2n-1) = \mathcal{O}(n^2)$ osnovnih računskih operacij. Ko te vrednosti, ki so odvisne le od interpolacijskih točk, imamo, lahko ob podanih vrednostih $f(x_k)$, $k = 0, 1, \ldots, n$, vrednost polinoma p v poljubni točki x izračunamo le s $5n+4 = \mathcal{O}(n)$ osnovnimi računskimi operacijami. Pri tem moramo biti pazljivi v primeru, ko je x enak eni izmed interpolacijskih točk x_k. Tedaj velja $p(x_k) = f(x_k)$.

Vrednost interpolacijskega polinoma lahko izračunamo tudi brez njegove eksplicitne konstrukcije. Tovrstne postopke združujemo pod imenom iterativna interpolacija. Eden izmed njih je računanje vrednosti s pomočjo Nevillove sheme, ki je predstavljena v nadaljevanju.

Naloga 7.5. Naj bodo $x_0, x_1, \ldots$ paroma različne točke in $p_{i,k}$, $i \geq 0$, $k \geq 0$, polinom stopnje največ k, ki interpolira vrednosti funkcije f v točkah $x_i, x_{i+1}, \ldots, x_{i+k}$.

1. Denimo, da sta vrednosti polinomov $p_{i,k-1}$ in $p_{i+1,k-1}$ v točki x znani. Dokažite, da za vrednost $p_{i,k}$ v točki x velja
$$p_{i,k}(x) = \frac{1}{x_{i+k} - x_i} \begin{vmatrix} x - x_i & p_{i,k-1}(x) \\ x - x_{i+k} & p_{i+1,k-1}(x) \end{vmatrix}.$$

2. Izračunajte vrednost polinoma, ki interpolira točke $(0, 2)$, $(2, 4)$ in $(4, 8)$, v točki $x = 1$. Pomagajte si z zvezo iz prejšnje točke tako, da začnete s konstantnimi polinomi in v dveh korakih izračunate vrednost parabole, ki interpolira dane točke.

3. Izračunajte vrednost polinoma, ki interpolira točke $(0, 2)$, $(2, 4)$, $(4, 8)$ in $(6, 10)$, v točki $x = 1$. Uporabite izračune iz prejšnje točke.

Rešitev.

1. Ker je
$$\begin{aligned} p_{i,k-1}(x_j) &= f(x_j), & j &= i, i+1, \ldots, i+k-1, \\ p_{i+1,k-1}(x_j) &= f(x_j), & j &= i+1, \ldots, i+k-1, i+k, \end{aligned}$$
za polinom
$$q(x) = \frac{x - x_i}{x_{i+k} - x_i} p_{i+1,k-1}(x) + \frac{x_{i+k} - x}{x_{i+k} - x_i} p_{i,k-1}(x)$$
velja
$$q(x_j) = f(x_j), \qquad j = i, i+1, \ldots, i+k.$$
Iz enoličnosti interpolacijskega polinoma sledi $p_{i,k} = q$, kar zaključuje dokaz.

2. Naj bo $x_0 = 0$, $x_1 = 2$ in $x_3 = 4$. Vemo, da je
$$p_{0,0}(x) = 2, \qquad p_{1,0}(x) = 4, \qquad p_{2,0}(x) = 8.$$
Nato za $x = 1$ izračunamo
$$p_{0,1}(x) = \frac{1}{x_1 - x_0} \begin{vmatrix} x - x_0 & p_{0,0}(x) \\ x - x_1 & p_{1,0}(x) \end{vmatrix} = \frac{1}{2} \begin{vmatrix} 1 & 2 \\ -1 & 4 \end{vmatrix} = 3,$$
$$p_{1,1}(x) = \frac{1}{x_2 - x_1} \begin{vmatrix} x - x_1 & p_{1,0}(x) \\ x - x_2 & p_{2,0}(x) \end{vmatrix} = \frac{1}{2} \begin{vmatrix} -1 & 4 \\ -3 & 8 \end{vmatrix} = 2$$

ter še

$$p_{0,2}(x) = \frac{1}{x_2 - x_0} \begin{vmatrix} x - x_0 & p_{0,1}(x) \\ x - x_2 & p_{1,1}(x) \end{vmatrix} = \frac{1}{4} \begin{vmatrix} 1 & 3 \\ -3 & 2 \end{vmatrix} = \frac{11}{4}.$$

3. Naj bo še $x_3 = 6$. Izračune iz prejšnje točke lahko zberemo v naslednji shemi, ki jo dopolnimo z izračunom vrednosti $p_{2,1}(x)$, $p_{1,2}(x)$ in $p_{0,3}(x)$ pri $x = 1$.

$x_\cdot$	$x - x_\cdot$	$p_{\cdot,0}(x)$	$p_{\cdot,1}(x)$	$p_{\cdot,2}(x)$	$p_{\cdot,3}(x)$
0	1	2			
			3		
2	-1	4		$\frac{11}{4}$	
			2		$\frac{5}{2}$
4	-3	8		$\frac{5}{4}$	
			5		
6	-5	10			

Sledi torej, da je $p_{0,3}(1) = \frac{5}{2}$.

7.2. Newtonova oblika

Lagrangeeva oblika interpolacijskega polinoma je enostavna in nazorna, a računsko ni najbolj učinkovita. Primernejša in bolj vsestranska je Newtonova oblika, ki je osnovana na deljenjih diferencah. Deljena diferenca $[x_0, x_1, \ldots, x_n]f$ predstavlja vodilni koeficient polinoma stopnje največ n, ki interpolira funkcijo f v točkah $x_0, x_1, \ldots, x_n$.

Naloga 7.6. Naj bodo $x_0, x_1, \ldots, x_n$ paroma različne točke. Dokažite, da velja

$$[x_0, x_1, \ldots, x_n]f = \sum_{k=0}^{n} w_k f(x_k),$$

kjer so vrednosti w_k, $k = 0, 1, \ldots, n$, definirane tako kot v nalogi 7.4.

Rešitev. Polinom p stopnje največ n, ki interpolira vrednosti funkcije f v točkah $x_0, x_1, \ldots, x_n$, lahko zapišemo v Lagrangeeevi obliki kot

$$p = \sum_{k=0}^{n} f(x_k)\ell_{n,k}.$$

Vodilni koeficient je koeficient pri funkciji $x \mapsto x^n$ in je po definiciji Lagrangeevih baznih polinomov enak

$$\sum_{k=0}^{n} f(x_k) \frac{1}{\prod_{\substack{j=0 \\ j \neq k}}^{n} (x_i - x_k)} = \sum_{k=0}^{n} \frac{f(x_k)}{\omega'(x_k)},$$

kar po definiciji w_k potrjuje dokazovano formulo.

Naloga 7.7. Naj bosta x_0 in x_1 različni točki. Utemeljite, da za funkcijo f, ki je definirana v x_0 in x_1, velja $[x_0]f = f(x_0)$ in $[x_1]f = f(x_1)$ ter

$$[x_0, x_1]f = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Čemu ustreza vrednost $\lim_{x_1 \to x_0} [x_0, x_1]f$?

Rešitev. Uporabimo lahko rezultat naloge 7.6. Če je $\omega(x) = x - x_0$ ali $\omega(x) = x - x_1$, je odvod funkcije ω enak 1, zato je $[x_0]f = f(x_0)$ in $[x_1]f = f(x_1)$. V primeru dveh točk je funkcija ω podana s predpisom $\omega(x) = (x - x_0)(x - x_1)$, njen odvod ω' pa s predpisom $\omega'(x) = 2x - x_0 - x_1$. Velja torej

$$[x_0, x_1]f = \frac{f(x_0)}{x_0 - x_1} + \frac{f(x_1)}{x_1 - x_0} = \frac{f(x_1) - f(x_0)}{x_1 - x_0}.$$

Limita tega izraza, ko pošljemo x_1 proti x_0, ustreza odvodu funkcije f v točki x_0.

Privlačnost deljenih diferenc se skriva v rekurzivni zvezi, ki ji zadoščajo. Za dano funkcijo f in zaporedje interpolacijskih točk $x_0, x_1, \ldots, x_{k-1}, x_k$ lahko vrednost $[x_0, x_1, \ldots, x_k]f$ izrazimo kot

$$[x_0, x_1, \ldots, x_k]f = \frac{[x_1, \ldots\ldots x_k]f - [x_0, \ldots x_{k-1}]f}{x_k - x_0}.$$

Ker je deljena diferenca $[x]f$ enaka vrednosti funkcije v točki x, lahko začnemo z vrednostmi $f(x_0), f(x_1), \ldots, f(x_k)$ in vrednost $[x_0, x_1, \ldots, x_k]f$ določimo v $k-1$ korakih rekurzije. Pri tem se lahko ista interpolacijska točka v zaporedju pojavi večkrat, s čimer ponazorimo, da v določeni točki poleg vrednosti interpoliramo tudi enega ali več odvodov funkcije. Za veljavnost rekurzivne formule zadostuje že, da sta vsaj dve točki različni, saj je deljena diferenca neodvisna od vrstnega reda točk in lahko privzamemo, da so urejene naraščajoče. Če so vse točke enake, to je $x_k = x_0$, deljeno diferenco izračunamo po formuli

$$[x_0, x_1, \ldots, x_k]f = \frac{f^{(k)}(x_0)}{k!}$$

ob predpostavki, da je funkcija k-krat odvedljiva v točki x_0.

Interpolacijski polinom p funkcije f za zaporedje interpolacijskih točk $x_0, x_1, \ldots, x_n$ je v Newtonovi obliki podan kot

$$p(x) = \sum_{k=0}^{n} (x - x_0)(x - x_1) \ldots (x - x_{k-1})[x_0, x_1, \ldots x_k]f.$$

Zanj velja $p(x_i) = f(x_i)$, $i = 0, 1, \ldots, n$. Še več, če se v interpolacijskem zaporedju točka x_j pojavi m-krat, velja $p^{(r)}(x_j) = f^{(r)}(x_j)$, $r = 0, 1, \ldots, m-1$.

Naloga 7.8. Naj bo p polinom stopnje 3, ki interpolira vrednosti $f(x) = 40/(x+1)$ v točkah 0, $\frac{1}{3}$, $\frac{2}{3}$, 1. Predstavite p v Newtonovi obliki.

Rešitev. Polinom p določimo na podlagi naslednje sheme. Vrednosti deljenih diferenc so izračunane s pomočjo rekurzivne formule.

$\cdot$	$[\cdot]f$	$[\cdot,\cdot]f$	$[\cdot,\cdot,\cdot]f$	$[\cdot,\cdot,\cdot,\cdot]f$
0	40			
		-30		
$\frac{1}{3}$	30		18	
		-18		-9
$\frac{2}{3}$	24		9	
		-12		
1	20			

S pomočjo podatkov v zgornji diagonali sheme določimo interpolacijski polinom

$$p(x) = 40 - 30x + 18x\left(x - \frac{1}{3}\right) - 9x\left(x - \frac{1}{3}\right)\left(x - \frac{2}{3}\right).$$

Razvidno je, da se p v potenčni bazi izraža kot $p(x) = -9x^3 + 27x^2 - 38x + 40$. Zato se (kot je zaradi enoličnosti interpolacijskega polinoma pričakovano) ujema s polinomom v Lagrangeevi obliki, izpeljanim v nalogi 7.3.

Naloga 7.9. Določite interpolacijski polinom q stopnje 3, ki interpolira vrednosti in odvode funkcije $f(x) = 40/(x+1)$ v točkah 0 in 1.

Rešitev. Polinom q določimo na podoben način kot polinom p v nalogi 7.8, le da tokrat uporabimo zaporedje interpolacijskih točk $0, 0, 1, 1$ namesto $0, \frac{1}{3}, \frac{2}{3}, 1$. S tem, da točki 0 in 1 v zaporedju podamo dvakrat, ponazarimo, da interpoliramo tako vrednosti kot vrednosti odvoda funkcije f v točkah 0 in 1. V shemi deljenih diferenc v stolpcu, ki določa deljeno diferenco f na dveh točkah, dvakrat uporabimo odvod $f'(x) = -40/(x+1)^2$ funkcije f.

$\cdot$	$[\cdot]f$	$[\cdot,\cdot]f$	$[\cdot,\cdot,\cdot]f$	$[\cdot,\cdot,\cdot,\cdot]f$
0	40			
		-40		
0	40		20	
		-20		-10
1	20		10	
		-10		
1	20			

Koeficiente polinoma q razberemo iz zgornje diagonale sheme deljenih diferenc. Dobimo

$$q(x) = 40 - 40x + 20x^2 - 10x^2(x-1),$$

kar se razlikuje od polinoma p v nalogi 7.8.

Newtonovo obliko interpolacijskega polinoma lahko razumemo kot posplošitev razvoja funkcije f v Taylorjevo vrsto. Ob primernih prepodstavkah za funkcijo f namreč velja

$$f(x) = p(x) + (x - x_0)(x - x_1)\dots(x - x_n)[x_0, x_1, \dots, x_n, x]f,$$

kjer p označuje interpolacijski polinom funkcije f za zaporedje interpolacijskih točk $x_0, x_1, \dots, x_n$. Iz te zveze po Hermite–Genocchijevi integralski formuli sledi

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!}(x - x_0)(x - x_1)\dots(x - x_n)$$

za nek $\xi \in (\min(x, x_0), \max(x, x_n))$, kar je priročna enakost za ocenjevanje napake interpolacije na zaprtem intervalu.

Naloga 7.10. S pomočjo četrtega odvoda funkcije $f(x) = 40/(x+1)$ ocenite napaki interpolacije

$$\max_{x\in[0,1]} |f(x) - p(x)| \quad \text{in} \quad \max_{x\in[0,1]} |f(x) - q(x)|$$

funkcije f s polinomoma p in q iz nalog 7.8 in 7.9 na intervalu $[0, 1]$.

Rešitev. Za polinom p velja

$$\max_{x\in[0,1]} |f(x) - p(x)| \le \frac{1}{4!} \max_{x\in[0,1]} \left|f^{(4)}(x)\right| \max_{x\in[0,1]} \left|x(x - \tfrac{1}{3})(x - \tfrac{2}{3})(x - 1)\right|.$$

Ker je $f^{(4)}(x) = 40 \cdot 4!/(x + 1)^5$, lahko prvi del zgornje ocene navzgor omejimo s 40. Za omejitev drugega dela ocene obravnavamo maksimalni absolutni vrednosti parabol $x(x - 1)$ in $(x - \frac{1}{3})(x - \frac{2}{3})$ na intervalu $[0, 1]$. Prva je omejena z $\frac{1}{4}$, druga pa z $\frac{2}{9}$, kar pomeni, da je drugi maksimum v oceni gotovo manjši od $\frac{1}{18}$. Na podlagi tega ocenimo

$$\max_{x\in[0,1]} |f(x) - p(x)| \le \frac{40}{18} = 2.\overline{2}.$$

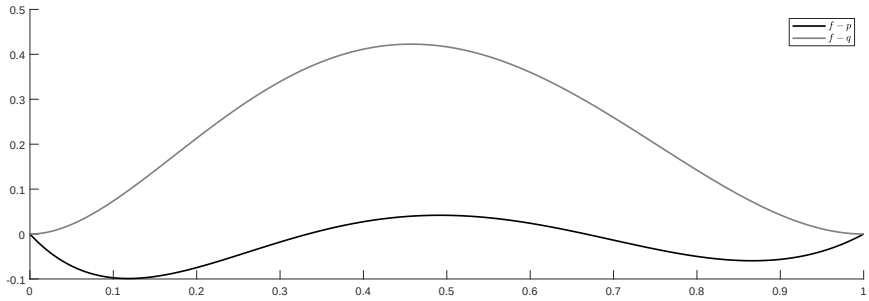
Napako interpolacije s polinom q ocenimo z

$$\max_{x\in[0,1]} |f(x) - q(x)| \le \frac{1}{4!} \max_{x\in[0,1]} \left|f^{(4)}(x)\right| \max_{x\in[0,1]} \left|x^2(x - 1)^2\right|.$$

Ker je absolutna vrednost parabole $x(x - 1)$ na intervalu $[0, 1]$ omejena z $\frac{1}{4}$, velja

$$\max_{x\in[0,1]} |f(x) - q(x)| \le \frac{40}{16} = 2.5.$$

Oceni namigujeta, da vzdolž intervala $[0, 1]$ polinom p bolje aproksimira funkcijo f kot polinom q. To potrjujeta tudi grafa razlik $f - p$ in $f - q$, prikazana na sliki 7.2.



SLIKA 7.2: *Graf razlik med funkcijo in interpolacijskima polinomoma iz naloge 7.10.*

Naloga 7.11. Dokažite, da za funkcijo $f(x) = 40/(x+1)$ in poljubno zaporedje interpolacijskih točk $x_0, x_1, \ldots, x_k$ velja

$$[x_0, x_1, \ldots, x_k]f = \frac{40(-1)^k}{(x_0+1)(x_1+1)\ldots(x_k+1)}.$$

S pomočjo te ugotovitve izboljšajte oceni napak iz naloge 7.10.

Rešitev. Dokažimo najprej, da izražava deljene diference velja, če so vse točke enake. Z indukcijo preverimo, da za k-ti odvod funkcije f velja

$$f^{(k)}(x) = \frac{40(-1)^k k!}{(x+1)^{k+1}},$$

kar pomeni, da je

$$[\underbrace{x_0, x_0, \ldots, x_0}_{k+1}]f = \frac{f^{(k)}(x_0)}{k!} = \frac{40(-1)^k}{(x_0+1)^{k+1}}.$$

Predpostavimo sedaj, da sta med točkami $x_0, x_1, \ldots, x_k$ vsaj dve različni (brez škode za splošnost sta to x_0 in x_k) in da izražava velja za deljenje diference, v katerih nastopa manj kot $k+1$ točk. Potem po rekurzivni zvezi za deljene diference velja

$$[x_0, x_1, \ldots, x_k]f = \frac{\frac{40(-1)^{k-1}}{(x_1+1)\ldots(x_{k-1}+1)(x_k+1)} - \frac{40(-1)^{k-1}}{(x_0+1)(x_1+1)\ldots(x_{k-1}+1)}}{x_k - x_0},$$

kar se poenostavi ravno v želeno izražavo.

Iz dokazanega sledi, da za poljuben $x \in [0,1]$ velja

$$\left|[0, \tfrac{1}{3}, \tfrac{2}{3}, 1, x]f\right| = \frac{9}{x+1} \leq 9, \qquad |[0,0,1,1,x]f| = \frac{10}{x+1} \leq 10.$$

S pomočjo ocen, izpeljanih v nalogi 7.10, sklepamo, da je

$$|f(x) - p(x)| \leq \frac{1}{2}, \qquad |f(x) - q(x)| \leq \frac{5}{8}.$$

Tudi ti dve oceni, ki sta precej boljši od ocen v nalogi 7.10, kažeta na to, da je polinom p boljša aproksimacija za f kot q.

Naloga 7.12. Prilagodite Hornerjev postopek za računanje vrednosti polinoma v Newtonovi obliki in izračunajte vrednosti polinomov p in q iz nalog 7.8 in 7.9 v točki $x = \frac{1}{2}$. Katera vrednost je boljši približek za $f(\frac{1}{2})$?

Rešitev. Vrednost polinoma, ki je predstavljen v obliki

$$x \mapsto a_0 + a_1(x - x_0) + \ldots + a_n(x - x_0) \cdot \ldots \cdot (x - x_{n-1}),$$

lahko v točki x izračunamo s $3n$ računskimi operacijami na podoben način kot s Hornerjevim postopkom. Edina razlika je, da moramo x v vsakem koraku postopka ustrezno zamakniti. Če podatke predstavimo s seznamoma $\boldsymbol{a} = (a_0, a_1, \ldots, a_n)$ in $\boldsymbol{X} = (x_0, x_1, \ldots, x_n)$, v Matlabu to opravimo na naslednji način.

```
n = length(a)-1;
b = a(n+1);
for i = n:-1:1
    b = a(i) + (x-X(i))*b;
end
```

Končna vrednost b ustreza vrednosti polinoma v točki x. Za izračun vrednosti polinomov p in q v točki $x = \frac{1}{2}$ so potrebni trije koraki postopka. Pri polinomu p spremenljivka b po vrsti zavzame vrednosti -9, 19.5, -26.75, 26.625, pri polinomu q pa -10, 25, -27.5, 26.25. Torej je $p(\frac{1}{2}) = 26.625$ in $q(\frac{1}{2}) = 26.25$. Ker je $f(\frac{1}{2}) = \frac{80}{3} = 26.\overline{6}$, polinom p v točki $\frac{1}{2}$ predstavlja boljši približek za funkcijo f kot polinom q.

Naloga 7.13. V Matlabu sestavite metodo, ki sprejme seznam interpolacijskih točk in seznam vrednosti funkcije v interpolacijskih točkah, vrne pa vrednosti interpolacijskega polinoma pri abscisah, podanih v dodatnem seznamu. Metoda naj temelji na konstrukciji sheme deljenih diferenc in izračunu vrednosti polinoma po prilagojenem Hornerjevem postopku. Preizkusite jo z interpolacijskimi podatki iz naloge 7.8, ki določajo interpolacijski polinom p za funkcijo f, in izračunajte

$$\max_{i=0,1,\ldots,1000} \left| f\left(\tfrac{i}{1000}\right) - p\left(\tfrac{i}{1000}\right) \right|,$$

da dobite približno vrednost napake interpolacije na intervalu $[0, 1]$.

Rešitev. Izsek kode, podan v nalogi 7.12, za izračun vrednosti polinoma, ki interpolira vrednosti funkcije f v $n+1$ točkah, dopolnimo s konstrukcijo tabele `F` velikosti $(n+1) \times (n+1)$, v katero shranjujemo vrednosti deljenih diferenc. V prvi stolpec te tabele postavimo funkcijske vrednosti iz `fX` v interpolacijskih točkah iz `X`. Spremenljivki `X` in `fX` predstavljata vrstici dolžine $n+1$, ki ju metoda prejme kot vhodna podatka. Nato v zanki dopolnimo tabelo `F` tako, da v koraku j, $j = 2, 3, \ldots, n+1$, zapolnimo prvih $n-j+2$ elementov v j-tem stolpcu z vrednostmi, ki jih izračunamo s pomočjo rekurzivne formule za deljene diference na podlagi vrednosti v prejšnjem stolpcu tabele `F` in interpolacijskih točk iz `X`.

```
n = length(X)-1;
F = [fX' NaN(n+1,n)];
for j = 2:n+1
    for i = 1:n-j+2
        F(i,j) = (F(i+1,j-1)-F(i,j-1))/(X(i+j-1)-X(i));
    end
end
```

Prva vrstica tabele `F` predstavlja koeficiente interpolacijskega polinoma v Newtonovi obliki. Za izračun vrednosti tega polinoma lahko uporabimo prilagojeni Hornerjev postopek. Pri podatkih `X = [0 1/3 2/3 1]` in `fX = [40 30 24 20]` izračunamo vrednosti polinoma v točkah `x = 0:1e-3:1`. Maksimalna absolutna razlika med vrednostmi funkcije in polinoma p v teh točkah je 0.0990.

Naloga 7.14. Metodo iz naloge 7.13 nadgradite tako, da lahko sprejme tudi seznam vrednosti odvodov funkcije ter izračuna vrednosti polinoma, ki interpolira tako vrednosti kot odvode funkcije v interpolacijskih točkah. Prilagojeno metodo preizkusite z interpolacijskimi podatki iz naloge 7.9, ki določajo polinom q za funkcijo f. Izračunajte

$$\max_{i=0,1,\ldots,1000} \left| f\left(\tfrac{i}{1000}\right) - q\left(\tfrac{i}{1000}\right) \right|$$

in rezultat primerjajte s približkom za napako iz naloge 7.13.

Rešitev. Metoda za izračun vrednosti polinoma, ki poleg vrednosti funkcije v interpolacijskih točkah interpolira tudi vrednosti njenih odvodov, mora sprejeti še eno dodatno vrstico `dfX` dolžine $n+1$, v kateri so shranjene vrednosti odvodov funkcije v interpolacijskih točkah. Shema deljenih diferenc je v tem primeru dolžine $(2n+2)\times(2n+2)$. V prvi stolpec `F` je treba na začetku vstaviti funkcijske vrednosti iz `fX` tako, da se vsaka zaporedoma pojavi dvakrat. V drugi stolpec v lihe vrstice `F` postavimo vrednosti odvodov iz `dfX`, v sode vrstice (razen zadnje) pa vrednosti, ki jih dobimo s pomočjo rekurzivne formule za deljene diference. Pred nadaljevanjem razširimo še seznam `X` tako, da vsako interpolacijsko točko podvojimo.

```
F = NaN(2*n+2);
F([1:2:2*n+1 2:2:2*n+2],1) = [fX fX]';
F([1:2:2*n+1 2:2:2*n],2) = [dfX diff(fX)./diff(X)]';
X([1:2:2*n+1 2:2:2*n+2]) = [X X]';
```

Preostanek postopka za izračun sheme deljenih diferenc je v večji meri enak prejšnjemu, treba je le ustrezno prilagoditi indekse v zanki. Začnemo s tretjim stolpcem in vztrajamo do stolpca $2n+2$.

Ocena za napako interpolacije vrednosti in odvodov funkcije f v točkah 0 in 1 s polinomom q, ki jo izračunamo v 1001 točkah na intervalu $[0,1]$, je enaka 0.4226 in nakazuje, da se polinom q slabše prilega funkciji f kot p. Z izrisom grafa f ter grafov p in q se lahko prepričamo, da se p enakomerno prilega funkciji f vzdolž celotnega intervala, medtem ko se q s funkcijo posebej dobro ujema v okolici točk 0 in 1, slabše pa v sredini intervala.

8. Odvajanje in integriranje

Osnovni način računanje približka za odvod ali integral funkcije je, da funkciji priredimo interpolacijski polinom in ga odvajamo oziroma integriramo namesto funkcije. S tem pristopom lahko izpeljemo pravila za odvode in integrale, ki temeljijo na funkcijskih vrednostih. Ta premislek lahko tudi obrnemo: pravilo za odvod ali integral zapišemo kot uteženo kombinacijo funkcijskih vrednosti in nato uteži določimo z zahtevo, da je pravilo točno za polinome čim višjih stopenj. Temu načinu izpeljave pravimo metoda nedoločenih koeficientov.

8.1. Pravila za odvajanje

Približek za vrednost odvoda funkcije f v točki x_0 je smiselno osnovati na podlagi vrednosti funkcije f v točki x_0 in njeni okolici. V odvisnosti od števila uporabljenih funkcijskih vrednosti lahko pričakujemo boljše ali slabše približke.

Naloga 8.1. Za dano funkcijo f bi radi izpeljali formulo za odvod funkcije f v točki x_0 oblike

$$F(x_0) = Af(x_0 - 2h) + Bf(x_0 - h) + Cf(x_0) + Df(x_0 + h) + Ef(x_0 + 2h),$$

kjer je $h > 0$ izbran razmik. Določite konstante A, B, C, D in E tako, da bo formula točna za vse polinome stopnje največ 4. Pomagajte si s sistemom enačb, ki ga dobite, če za funkcijo f vzamete funkcije $x \mapsto (x - x_0)^i$, $i = 0, 1, 2, 3, 4$.

Rešitev. Poljuben polinom stopnje največ 4 lahko predstavimo v predlagani potenčni bazi z zamikom x_0. Zato bo formula zanj točna, če konstante A, B, C, D in E zadoščajo enačbam

$$\begin{aligned}
0 &= A + B + C + D + E,\\
1 &= -2hA - hB + hD + 2hE,\\
0 &= 4h^2A + h^2B + h^2D + 4h^2E,\\
0 &= -8h^3A - h^3B + h^3D + 8h^3E,\\
0 &= 16h^4A + h^4B + h^4D + 16h^4E,
\end{aligned}$$

ki jih dobimo tako, da v formuli funkcijo f po vrsti nadomestimo z baznimi funkcijami $x \mapsto (x - x_0)^i$, $i = 0, 1, 2, 3, 4$. Od tod izračunamo

$$A = \frac{1}{12h}, \quad B = -\frac{2}{3h}, \quad C = 0, \quad D = \frac{2}{3h}, \quad E = -\frac{1}{12h},$$

torej je

$$F(x_0) = \frac{1}{12h}\left(f(x_0 - 2h) - 8f(x_0 - h) + 8f(x_0 + h) - f(x_0 + 2h)\right)$$

iskana formula za približek $f'(x_0)$.

Naloga 8.2. Predpostavite, da je funkcija f petkrat zvezno odvedljiva in za pravilo F iz naloge 8.1 izrazite ostanek $f'(x_0) - F(x_0)$ v obliki $Kf^{(5)}(\xi)$, kjer je K konstanta, odvisna od h, in ξ število z intervala $(x_0 - 2h, x_0 + 2h)$.

Rešitev. Pravilo $F(x_0)$ je enolično določeno z zahtevo, da je točno za polinome stopnje manjše ali enake 4. Potemtakem zaradi enoličnosti polinoma p stopnje največ 4, ki interpolira vrednosti f v točkah $x_0 - 2h$, $x_0 - h$, x_0, $x_0 + h$, $x_0 + 2h$, ustreza vrednosti $p'(x_0)$. Za vsak $x \in [x_0 - 2h, x_0 + 2h]$ obstaja število $\xi \in (x_0 - 2h, x_0 + 2h)$, odvisno od x, da velja

$$f(x) - p(x) = \frac{f^{(5)}(\xi)}{5!}(x - x_0 + 2h)(x - x_0 + h)(x - x_0)(x - x_0 - h)(x - x_0 - 2h).$$

Če zgornji izraz odvajamo in izvrednotimo v x_0, dobimo

$$f'(x_0) - F(x_0) = f'(x_0) - p'(x_0) = \frac{f^{(5)}(\xi)}{5!}4h^4 = \frac{h^4}{30}f^{(5)}(\xi),$$

kar predstavlja ostanek v iskani obliki.

Naloga 8.3. S pravilom F, izpeljanim v nalogi 8.1, bi radi poiskali čimbolj natančen približek za odvod funkcije $f(x) = 1/(x + 1)$ v točki $x_0 = \frac{1}{2}$. Napaka pri računanju odvoda je odvisna od h in je sestavljena iz dveh delov, napake metode, ki jo določa ostanek, in zaokrožitvene napake pri računanju funkcije f. Predpostavite, da je napaka pri izračunu vrednosti funkcije f manjša od ε, in določite h, pri katerem je napaka pri izračunu odvoda najmanjša.

Rešitev. Napaka metode za dano funkcijo f ustreza absolutni vrednosti ostanka. Za majhen h bo približno enaka

$$\frac{h^4}{30}\frac{5!}{(1 + 1/2)^6} = \frac{2^8}{3^6}h^4.$$

Zaokrožitveno napako lahko na podlagi formule za odvajanje ocenimo z

$$\frac{1}{h}\left|\frac{1}{12}\varepsilon_1 - \frac{2}{3}\varepsilon_2 + \frac{2}{3h}\varepsilon_3 - \frac{1}{12}\varepsilon_4\right| \leq \frac{3\varepsilon}{2h}, \qquad |\varepsilon_i| \leq \varepsilon, \ i = 1, 2, 3, 4.$$

Celotna napaka je torej oblike

$$\frac{2^8}{3^6}h^4 + \frac{3\varepsilon}{2h}$$

in bo najmanjša pri $h \approx \sqrt[5]{3^7/2^{11}\varepsilon}$.

8.2. Pravila za integriranje

Interpolacijska integracijska pravila za funkcijo f na intervalu $[a, b]$ v splošnem predstavimo v obliki

$$\int_a^b f(x)\,\mathrm{d}x = \sum_{i=0}^{n} w_i f(x_i) + R(f).$$

Izpeljemo jih z integracijo interpolacijskega polinoma za funkcijo f. Za točke $x_0 < x_1 < \ldots < x_n$ predpostavljamo, da so paroma različne; imenujemo jih vozli integracijskega pravila. Koeficienti w_i, $i = 0, 1, \ldots, n$, so uteži, ki jih lahko predstavimo kot integrale

$$w_i = \int_a^b \ell_{n,i}(x)\,\mathrm{d}x$$

Lagrangeevih baznih polinomov $\ell_{n,i}$ stopnje n, definiranih na podlagi interpolacijskih točk x_i. Izraz $R(f)$ označuje ostanek integracijskega pravila, za katerega ob predpostavki, da je f dovoljkrat zvezno odvedljiva, velja

$$R(f) = \int_a^b \frac{f^{(n+1)}(\xi_x)}{(n+1)!}\omega(x)\,\mathrm{d}x,$$

kjer je $\omega(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$ in $\xi_x \in (a, b)$ število, odvisno od parametra x.

Naloga 8.4. Izpeljite interpolacijsko integracijsko pravilo za integral funkcije f na intervalu $[a, b]$ z vozloma $x_0 = a$ in $x_1 = b$. Ob predpostavki, da je funkcija f dvakrat zvezno odvedljiva, dokažite, da obstaja $\xi \in (a, b)$, da za ostanek velja

$$R(f) = -\frac{(b-a)^3}{12} f''(\xi),$$

ter utemeljite, da je pravilo točno za vse linearne funkcije. Kaj geometrijsko predstavlja približek za integral funkcije f?

Rešitev. Uteži integracijskega pravila

$$\int_a^b f(x)\,\mathrm{d}x = w_0 f(a) + w_1 f(a) + R(f)$$

sta podani z

$$w_0 = \int_a^b \frac{x-b}{a-b}\,\mathrm{d}x = \frac{b-a}{2}, \quad w_1 = \int_a^b \frac{x-a}{b-a}\,\mathrm{d}x = \frac{b-a}{2},$$

ostanek pa z

$$R(f) = \frac{1}{2}\int_a^b f''(\xi_x)(x-a)(x-b)\,\mathrm{d}x.$$

Slednjega lahko poenostavimo. Ker je funkcija f'' zvezna na intervalu $[a,b]$, je omejena, zato obstajata realni konstanti k in K, da velja $k \leq f''(\xi_x) \leq K$ za vsak $x \in [a,b]$. Nadalje velja

$$K(x-a)(x-b) \leq f''(\xi_x)(x-a)(x-b) \leq k(x-a)(x-b),$$

saj je parabola $x \mapsto (x-a)(x-b)$ na intervalu $[a,b]$ negativnega predznaka. Po integraciji zgornjih neenačb dobimo

$$K\int_a^b (x-a)(x-b)\,\mathrm{d}x \leq \int_a^b f''(\xi_x)(x-a)(x-b)\,\mathrm{d}x \leq k\int_a^b (x-a)(x-b)$$

in iz

$$\int_a^b (x-a)(x-b)\,\mathrm{d}x = -\frac{1}{6}(b-a)^3$$

sledi

$$k \leq -\frac{6}{(b-a)^3}\int_a^b f''(\xi_x)(x-a)(x-b)\,\mathrm{d}x \leq K.$$

Ponovno upoštevamo, da je f'' zvezna na intervalu $[a,b]$, kar pomeni, da zavzame vse vrednosti med k in K. Torej obstaja tako število $\xi \in (a,b)$, da je

$$f''(\xi) = -\frac{6}{(b-a)^3}\int_a^b f''(\xi_x)(x-a)(x-b).$$

Integral funkcije f lahko torej predstavimo kot

$$\int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\left(f(a)+f(b)\right) - \frac{(b-a)^3}{12}f''(\xi).$$

Če je funkcija f linearna, je odvod f'' enak 0 in $R(f)=0$. To pomeni, da je pravilo točno. Geometrijsko vrednost $\frac{b-a}{2}\left(f(a)+f(b)\right)$ predstavlja ploščino trapeza z oglišči $(a,0)$, $(b,0)$, $(b,f(b))$, $(a,f(a))$, zato se je pravila prijelo ime trapezno pravilo.

Poseben primer interpolacijskih pravil so Newton–Cotesova pravila, pri katerih so vozli x_i ekvidistantni ($x_i = x_{i-1}+h$ za $h=(b-a)/n$). Ločimo pravila zaprtega tipa, pri katerih prvi in zadnji vozel ustrezata robovoma intervala ($x_0=a$ in $x_n=b$) ter pravila odprtega tipa, pri katerih točki x_0 in x_n izpustimo (prvi in zadnji vozel sta določena z $x_1=a+h$ in $x_{n-1}=b-h$). Slednja navadno uporabljamo, kadar funkcija na robovih intervala ni omejena. Pri pravilih zaprtega tipa za $n=1$ dobimo trapezno pravilo:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{h}{2}\left(f(x_0)+f(x_1)\right) - \frac{1}{12}h^3 f^{(2)}(\xi),$$

pri $n=2$ pa Simpsonovo pravilo:

$$\int_a^b f(x)\,\mathrm{d}x = \frac{1}{3}h\left(f(x_0)+4f(x_1)+f(x_2)\right) - \frac{1}{90}h^5 f^{(4)}(\xi).$$

Pri tem je $\xi \in (a,b)$. Pravilu za $n=3$ pravimo 3/8-pravilo.

Naloga 8.5. Izpeljite Newton–Cotesovo integracijsko pravilo zaprtega tipa s štirimi vozli ($n = 3$), ki je oblike

$$\int_a^b f(x)\,\mathrm{d}x = Af(x_0) + Bf(x_1) + Cf(x_2) + Df(x_3) + Ef^{(r)}(\xi), \qquad \xi \in (a, b).$$

Konstante A, B, C in D določite tako, da je pravilo točno za polinome čim višjih stopenj. Pomagajte si s funkcijami $x \mapsto (x - a)^i$, $i = 0, 1, 2, 3$, ki sestavljajo bazo za prostor polinomov stopnje 3. Nato na podoben način določite še konstanti E in r ob predpostavki, da je f dovoljkrat zvezno odvedljiva.

Rešitev. Ko integracijsko pravilo uporabimo za funkcije $x \mapsto (x - a)^i$, $i = 0, 1, 2, 3$, dobimo

$$\begin{aligned} 3h &= A + B + C + D, \\ \tfrac{1}{2}(3h)^2 &= hB + 2hC + 3hD, \\ \tfrac{1}{3}(3h)^3 &= h^2B + 4h^2C + 9h^2D, \\ \tfrac{1}{4}(3h)^4 &= h^3B + 8h^3C + 27h^3D. \end{aligned}$$

Konstante, ki zadoščajo zgornjemu sistemu enačb, so

$$A = \frac{3}{8}h, \quad B = \frac{9}{8}h, \quad C = \frac{9}{8}h, \quad D = \frac{3}{8}h.$$

Oglejmo si še ostanek integracijskega pravila za funkcijo $x \mapsto (x - x_0)^4$. Ker je

$$\frac{(3h)^5}{5} - \frac{3h}{8}\left(0 + 3h^4 + 3(2h)^4 + (3h)^4\right) = -\frac{9}{10}h^5,$$

je $r = 4$ in $E = -3/80h^5$. S tem je 3/8-pravilo

$$\int_a^b f(x)\,\mathrm{d}x = \frac{3}{8}\left(f(x_0) + 3f(x_1) + 3f(x_2) + f(x_3)\right) - \frac{3}{80}h^5 f^{(4)}(\xi).$$

v celoti določeno.

Med Newton–Cotesovi pravili odprtega tipa je najenostavnejše sredinsko pravilo z enim vozlom ($n = 2$):

$$\int_a^b f(x)\,\mathrm{d}x = 2hf(x_1) + \frac{1}{3}h^3 f^{(2)}(\xi).$$

Pogosto se uporablja Milnovo pravilo s tremi vozli, ki ustreza izbiri $n = 4$.

Naloga 8.6. Izpeljite Newton–Cotesovo integracijsko pravilo odprtega tipa s tremi vozli ($n = 4$), ki je oblike

$$\int_a^b f(x)\,\mathrm{d}x = Af(x_1) + Bf(x_2) + Cf(x_3) + Df^{(r)}(\xi), \quad \xi \in (a, b).$$

Konstante A, B in C določite tako, da je pravilo točno za polinome čim višjih stopenj. Z določitvijo konstant D in r izrazite tudi ostanek pravila ob predpostavki, da je funkcija f dovoljkrat zvezno odvedljiva.

Rešitev. Z ustrezno izbiro treh prostih konstant lahko zagotovimo, da bo pravilo točno za vse parabole. To je res natanko tedaj, ko je pravilo točno za funkcije $x \mapsto (x-a)^i$, $i = 0, 1, 2$, oziroma ko konstante A, B in C zadoščajo zvezam

$$4h = A + B + C, \quad 8h^2 = hA + 2hB + 3hC, \quad 64h^3/3 = h^2A + 4h^2B + 9h^2C.$$

Rešitev dobljenega sistema enačb je $A = 8h/3$, $B = -4h/3$ in $C = 8h/3$. Nadalje opazimo, da je

$$\int_a^b (x-a)^3 \,\mathrm{d}x = 64h^4 = \frac{4h}{3}\left(2h^3 - 8h^3 + 54h^3\right)$$

in

$$\int_a^b (x-a)^4 \,\mathrm{d}x = \frac{1024h^5}{5} \neq \frac{4h}{3}\left(2h^4 - 16h^4 + 162h^4\right),$$

od kjer sledi, da je pravilo točno za polinome stopnje manjše ali enake 3, ne pa tudi za polinome stopnje 4. To pomeni, da je $r = 4$. Iz druge enačbe potem sledi še, da je $4!D = 1024h^5/5 - 592h^5/3 = 112h^5/15$, torej je $D = 14h^5/45$. Na ta način smo izpeljali Milnovo pravilo

$$\int_a^b f(x)\,\mathrm{d}x = \frac{4}{3}h\left(2f(x_1) - f(x_2) + 2f(x_3)\right) + \frac{14}{45}h^5 f^{(4)}(\xi).$$

s štirimi vozli.

Z dodajanjem vozlov (to je z večanjem števila n) ne dobimo nujno pravil, ki bi nam zagotovila boljše približke za integral funkcije. Primernejša nadgradnja osnovnih Newton–Cotesovih pravil so tako imenovana sestavljena pravila, pri katerih interval $[a, b]$ razdelimo na m enako dolgih podintervalov, nato pa na vsakem izmed njih uporabimo osnovno pravilo, določeno z majhnim številom n.

Naloga 8.7. Naj $S_1 f$ označuje Simpsonovo pravilo na intervalu $[a, b]$, $S_2 f$ pa integracijsko pravilo, sestavljeno iz dveh Simpsonovih pravil na intervalu $[a, b]$. Dokažite, da lahko Milnovo pravilo, izpeljano v nalogi 8.6, izrazimo kot kombinacijo pravil $S_1 f$ in $S_2 f$.

Rešitev. Integracijski pravili $S_1 f$ in $S_2 f$ sta podani z

$$S_1 f = \frac{2h}{3}\left(f(x_0) + 4f(x_2) + f(x_4)\right),$$
$$S_2 f = \frac{h}{3}\left(f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + f(x_4)\right),$$

kjer je $h = (b-a)/4$ in $x_i = a + ih$, $i = 0, 1, \ldots, 4$. Ker v Milnovem pravilu ne nastopata funkcijski vrednosti v točkah x_0 in x_4, vzamemo

$$2S_2 f - S_1 f = \frac{4h}{3}\left(2f(x_1) - f(x_2) + 2f(x_3)\right),$$

kar ravno ustreza formuli, izpeljani v nalogi 8.6.

Naloga 8.8. Naj bo

$$x_j = a + jh, \quad h = \frac{b-a}{m}, \quad j = 0, 1, \ldots, m,$$

kjer je m izbrano sodo naravno število. Predpostavite, da za sestavljeno trapezno pravilo

$$T_h f = \frac{h}{2}\left(f(x_0) + 2f(x_1) + \ldots + 2f(x_{m-1}) + f(x_m)\right)$$

velja

$$\int_a^b f(x)\,\mathrm{d}x = T_h f + c_2 h^2 + c_4 h^4 + \ldots, \quad c_2, c_4, \ldots \in \mathbb{R}.$$

Utemeljite, zakaj je $T_h^1 f = (4T_h f - T_{2h} f)/3$ boljši približek za integral kot $T_h f$, in ugotovite, kateremu sestavljenemu pravilu ustreza $T_h^1 f$.

Rešitev. Na podlagi izražave ostanka za sestavljeno trapezno pravilo sklepamo, da je

$$\int_a^b f(x)\,\mathrm{d}x - T_h^1 f = \frac{1}{3}\left(4c_2 h^2 + 4c_4 h^4 + \ldots - c_2(2h)^2 - c_4(2h)^4 - \ldots\right) = \mathcal{O}(h^4),$$

kar pomeni, da je $T_h^1 f$ približek za integral, ki je za dva reda boljši od $T_h f$. Z upoštevanjem izražav $T_h f$ in T_{2hf} lahko pravilo $T_h^1 f$ zapišemo kot

$$T_h^1 f = \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{\frac{m}{2}} f(x_{2i-1}) + 2\sum_{i=1}^{\frac{m}{2}-1} f(x_{2i}) + f(x_m)\right),$$

kar ustreza ravno sestavljenemu Simpsonovemu pravilu za korak h.

Naloga 8.9. Naj bo

$$x_j = a + jh, \quad h = \frac{b-a}{2m}, \quad j = 0, 1, \ldots, 2m,$$

kjer je m izbrano naravno število. V Matlabu implementirajte sestavljeno Simpsonovo pravilo

$$S_m f = \frac{h}{3}\left(f(x_0) + 4\sum_{i=1}^{m} f(x_{2i-1}) + 2\sum_{i=1}^{m-1} f(x_{2i}) + f(x_{2m})\right).$$

Preizkusite ga z računanjem integrala funkcije $f(x) = e^{-x^2}$ na intervalu $[0, 3]$ pri izbirah $m = 3k + 1$, $k = 0, 1, 2, 3, 4, 5$. Primerjajte dobljene vrednosti s točno vrednostjo integrala. Kakšen je najmanjši m, pri katerem se $S_m f$ od točne rešitve $\sqrt{\pi}\,\mathrm{erf}(3)/2$ absolutno razlikuje za manj kot 10^{-10}?

Rešitev. Funkcijo, ki izračuna približek po sestavljenem Simpsonovem pravilu, lahko zasnujemo tako, da sprejme seznam `fX`, ki vsebuje vrednosti funkcije f v točkah iz seznama `X = a:h:b`, kjer je `h = (b-a)/(2*m)`. Približek za integral `Sf` funkcije f na intervalu $[a, b]$ lahko potem izračunamo z naslednjim ukazom.

```
Sf = h*(fX(1) + 4*sum(fX(2:2:end-1)) + ...
                2*sum(fX(3:2:end-2)) + fX(end))/3;
```

Tabela 8.1 vsebuje napake $e_m = |\sqrt{\pi}\,\mathrm{erf}(3)/2 - S_m f|$, dobljene pri numeričnem integriranju funkcije $f(x) = e^{-x^2}$ na intervalu $[0, 3]$ pri različnih izbirah m. Točna vrednost je izračunana s pomočjo vgrajene funkcije `erf`. Nadalje izračunamo, da je $m = 50$ najmanjši parameter, pri katerem je napaka sestavljenega pravila manjša od 10^{-10}.

m	1	4	7	10	13	16
e_m	$1.8 \cdot 10^{-1}$	$1.7 \cdot 10^{-6}$	$2.3 \cdot 10^{-7}$	$5.9 \cdot 10^{-8}$	$2.1 \cdot 10^{-8}$	$9.3 \cdot 10^{-9}$

TABELA 8.1: *Napake pri integraciji s sestavljenim Simpsonovim pravilom v nalogi 8.9.*

Naloga 8.10. V Matlabu implementirajte adaptivno Simpsonovo pravilo. Preizkusite ga z integracijo funkcije $g(x) = 1/\sqrt{x + 10^{-6}}$ na intervalu $[0, 1]$, ki iz začetne vrednosti 1000 na levem robu intervala zelo hitro pade proti 0. Približno določite toleranco, pri kateri se rezultat po absolutni vrednosti razlikuje za manj kot 10^{-15} od točne rešitve $(\sqrt{10^6 + 1} - 1)/500$, in primerjajte čas izvajanja metode s časom izvajanja sestavljenega Simpsonovega pravila pri $m = 10^8$, ki da primerljivo dober približek.

Rešitev. Funkcijo za računanje integrala po adaptivnem Simpsonovem pravilu zasnujemo rekurzivno. Sprejme funkcijo f ter interval $[a, b]$, na katerem jo integriramo, poleg tega pa še toleranco (`tol`) za absolutno napako približka. V funkciji najprej izračunamo približek `Sf1` po osnovnem Simpsonovem pravilu na intervalu $[a, b]$ in približek `Sf2` po sestavljenem Simpsonovem pravilu na dveh podintervalih intervala $[a, b]$. Na podlagi teh dveh vrednosti lahko napako integracije ocenimo z `abs(Sf2-Sf1)/15`, kar sledi iz izpeljave Richardsonove ekstrapolacije. Če je ocenjena napaka manjša od tolerance, zaključimo in približek Richardsonove integracije prištejemo približku za integral, sicer pa rekurzivno nadaljujemo na levem in desnem podintervalu intervala $[a, b]$ z razpolovljeno toleranco ter rezultata rekurzivnih klicev seštejemo.

```
function Sf = adsimpson(f,a,b,tol)

c = (a+b)/2;
d = (a+c)/2;
e = (c+b)/2;

fa = f(a);
```

```
fc = f(c);
fb = f(b);

Sf1 = (b-a)*(fa + 4*fc + fb)/6;
Sf2 = (b-a)*(fa + 4*f(d) + 2*fc + 4*f(e) + fb)/12;

r = (Sf2-Sf1)/15;
if abs(r) <= tol
    Sf = Sf2 + r;
else
    Sfa = adsimpson(f,a,c,tol/2);
    Sfb = adsimpson(f,c,b,tol/2);
    Sf = Sfa + Sfb;
end

end
```

Izkaže se, da je tak postopek izredno učinkovit pri integraciji funkcij, ki na enem delu intervala strmo rastejo ali padajo, na drugem pa so zelo položne oziroma se blago spreminjajo. V našem primeru lahko z adaptivnim Simpsonovim pravilom približek za integral, ki se od točne rešitve absolutno razlikuje za manj kot 10^{-15}, izračunamo približno 20-krat hitreje kot s sestavljenim Simpsonovim pravilom.

Poseben tip interpolacijskih pravil so Gaussova pravila, pri katerih vozle določimo z zahtevo, da je pravilo točno za polinome najvišje možne stopnje. Posledično so taka pravila navadno uporabna le v primerih, ko poznamo funkcijske vrednosti na celotnem intervalu integracije, je pa tak pristop izredno učinkovit, saj se izkaže, da lahko z $n+1$ vozli dobimo pravilo, ki je točno za polinome stopnje $2n+1$.

Naloga 8.11. Naj bo funkcija f definirana na intervalu $[-1,1]$. Določite uteži α_0, α_1 in vozla x_0, x_1 v Gaussovem pravilu

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \alpha_0 f(x_0) + \alpha_1 f(x_1) + K f^{(4)}(\xi), \quad \xi \in (-1,1).$$

Določite tudi konstanto K, ki nastopa v ostanku pravila.

Rešitev. V zastavljenem integracijskem pravilu proste parametre predstavljata uteži α_0 in α_1 ter vozla x_0 in x_1. Zato domnevamo, da jih lahko določimo tako, da bo pravilo točno za polinome stopnje manjše ali enake 3 (vsak tak polinom ima namreč 4 proste parametre). Uporabimo integracijsko pravilo za funkcije $x \mapsto x^i$, $i = 0,1,2,3$,

in zahtevajmo, da je za te funkcije pravilo točno. Dobimo sistem nelinearnih enačb

$$\begin{aligned} 2 &= \alpha_0 + \alpha_1, \\ 0 &= \alpha_0 x_0 + \alpha_1 x_1, \\ 2/3 &= \alpha_0 x_0^2 + \alpha_1 x_1^2, \\ 0 &= \alpha_0 x_0^3 + \alpha_1 x_1^3. \end{aligned}$$

Iz teh enačb lahko hitro sklepamo, da so uteži in vozla neničelni. Če drugo enačbo množimo z x_0^2 in od nje odštejemo četrto, dobimo

$$\alpha_1 x_1 (x_0^2 - x_1^2) = 0,$$

torej mora biti $x_0^2 = x_1^2$. Po prvi in tretji enačbi je potem

$$2x_0^2 = (\alpha_0 + \alpha_1) x_0^2 = \frac{2}{3}.$$

Brez škode za splošnost privzamemo, da je $x_0 < x_1$, torej je $x_0 = -\sqrt{3}/3$ in $x_1 = \sqrt{3}/3$. To tudi pomeni, da sta uteži α_0 in α_1 enaki 1. Ostanek izpeljanega integracijskega pravila določimo s pomočjo funkcije $x \mapsto x^4$. Po predpostavki, da je za splošno funkcijo f oblike $Kf^{(4)}(\xi)$, iz

$$\int_{-1}^{1} x^4 \,\mathrm{d}x - \left(-\sqrt{3}/3\right)^4 - \left(\sqrt{3}/3\right)^4 = \frac{2}{5} - \frac{2}{9} = \frac{8}{45},$$

sledi $K = 1/135$. Integral lahko torej zapišemo kot

$$\int_{-1}^{1} f(x) \,\mathrm{d}x = f(-\sqrt{3}/3) + f(\sqrt{3}/3) + \frac{1}{135} f^{(4)}(\xi).$$

Izpeljava Gaussovih interpolacijskih pravil z metodo nedoločenih koeficientov je v splošnem težavna, saj je za določitev vozlov in uteži treba rešiti sistem nelinearnih enačb. K sreči se izkaže, da lahko vozle določimo neodvisno od uteži kot ničle polinoma stopnje $n+1$, ki je ortogonalen na vse polinome stopnje največ n. Pri tem je skalarni produkt $\langle g, h \rangle$ s kvadratom integrabilnih funkcij g in h definiran s predpisom

$$\langle g, h \rangle = \int_{-1}^{1} g(x) h(x) \,\mathrm{d}x.$$

Ko določimo vozle, lahko uteži pravila izpeljemo po standardnem postopku z metodo nedoločenih koeficientov.

Naloga 8.12. Izpeljite Gaussovo integracijsko pravilo

$$\int_{-1}^{1} f(x) \,\mathrm{d}x = \alpha_0 f(x_0) + \alpha_1 f(x_1) + \alpha_2 f(x_2) + K f^{(6)}(\xi), \quad \xi \in (-1, 1),$$

ki je točno za polinome stopnje največ 5. Vozle x_0, x_1, x_2 določite s pomočjo kubičnega polinoma, ki je ortogonalen na vse polinome stopnje največ 2, uteži α_0, α_1, α_2 pa z metodo nedoločenih koeficientov. Določite tudi konstanto K v ostanku pravila, izražava katerega temelji na predpostavki, da je funkcija f šestkrat zvezno odvedljiva.

Rešitev. Kubični polinom, ortogonalen na polinome stopnje manjše ali enake 2, lahko določimo z Gram–Schmidtovim postopkom za ortogonalizacijo baznih funkcij $x \mapsto x^i$, $i = 0, 1, 2, 3$. Skalarni produkt funkcij g in h je podan s predpisom

$$\langle g, h \rangle = \int_{-1}^{1} g(x)h(x)\,\mathrm{d}x.$$

Po standardnem postopku dobimo bazne polinome

$$x \mapsto \frac{1}{\sqrt{2}}, \quad x \mapsto \frac{\sqrt{3}}{\sqrt{2}}x, \quad x \mapsto \frac{3\sqrt{5}}{2\sqrt{2}}\left(x^2 - \frac{1}{3}\right), \quad x \mapsto \left(x + \frac{\sqrt{3}}{\sqrt{5}}\right)x\left(x - \frac{\sqrt{3}}{\sqrt{5}}\right).$$

Pri tem zadnji polinom ni normiran, saj nas zanimajo le njegove ničle, ki ustrezajo vozlom $x_0 = -\sqrt{3}/\sqrt{5}$, $x_1 = 0$, $x_2 = \sqrt{3}/\sqrt{5}$. Sedaj lahko uteži določimo z reševanjem sistema linearnih enačb

$$\alpha_0 + \alpha_1 + \alpha_2 = 2, \quad -\frac{\sqrt{3}}{\sqrt{5}}\alpha_0 + \frac{\sqrt{3}}{\sqrt{5}}\alpha_2 = 0, \quad \frac{3}{5}\alpha_0 + \frac{3}{5}\alpha_2 = \frac{2}{3},$$

ki jih dobimo, če v nastavek za integracijsko formulo po vrsti vstavimo funkcije $x \mapsto x^i$, $i = 0, 1, 2$, za katere zahtevamo, da je pravilo točno. Iz tega sledi $\alpha_0 = 5/9$, $\alpha_1 = 8/9$ in $\alpha_2 = 5/9$. Izpeljano pravilo

$$\int_{-1}^{1} f(x)\,\mathrm{d}x = \frac{1}{9}\left(5f\left(-\frac{\sqrt{3}}{\sqrt{5}}\right) + 8f(0) + 5f\left(\frac{\sqrt{3}}{\sqrt{5}}\right)\right) + Kf^{(6)}(\xi)$$

je po konstrukciji točno za polinome stopnje manjše ali enake 5. Konstanto K v ostanku pravila določimo tako, da vstavimo funkcijo $x \mapsto x^6$. Iz

$$\frac{2}{7} = \frac{1}{9}\left(5\frac{27}{125} + 5\frac{27}{125}\right) + 6!K$$

sledi $K = 1/15750$.

9. Diferencialne enačbe

Eden izmed klasičnih problemov matematične analize je reševanje navadne diferencialne enačbe

$$y'(x) = f(x, y(x)), \quad x > x_0,$$

kjer je f dana funkcija dveh spremenljivk, y pa neznana funkcija, ki jo iščemo. Ob ustreznih predpostavkah o funkciji f funkcija y v okolici x_0 obstaja in je enolična, če predpišemo začetni pogoj $y(x_0) = y_0$. Temu pravimo začetni problem ali Cauchyjeva naloga.

9.1. Runge–Kutta metode

Pomemben razred numeričnih metod za reševanje začetnih problemov so Runge–Kutta metode. Gre za diskretne metode, pri katerih približek za funkcijo y iščemo samo v določenih točkah desno od x_0. Približek y_n za točno vrednost $y(x_n)$ v točki $x_n = x_0 + nh$, $h > 0$, določimo na podlagi formule

$$y_n = y_{n-1} + \sum_{i=1}^{s} \gamma_i k_i$$

za izbrano stopnjo s in izbrane koeficiente γ_i, $i = 1, 2, \ldots, s$. Za slednje navadno velja, da se seštejejo v 1, kar pomeni, da je vsota $\sum_{i=1}^{s} \gamma_i k_i$ povprečje premikov k_i. Ti so v splošnem definirani kot

$$k_i = hf\left(x_{n-1} + \alpha_i h, y_{n-1} + \sum_{j=1}^{s} \beta_{i,j} k_j\right)$$

in odvisni od izbire parametrov α_i in $\beta_{i,j}$. Parametre Runge–Kutta metode ekonomično podamo z Butcherjevo shemo

$$\begin{array}{c|cccc} \alpha_1 & \beta_{1,1} & \beta_{1,2} & \cdots & \beta_{1,s} \\ \alpha_2 & \beta_{2,1} & \beta_{2,2} & \cdots & \beta_{2,s} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha_s & \beta_{s,1} & \beta_{s,2} & \cdots & \beta_{s,s} \\ \hline & \gamma_1 & \gamma_2 & \cdots & \gamma_s \end{array}.$$

Ker vsaka taka metoda za izračun približka y_n uporabi le približek y_{n-1}, jih uvrščamo med enočlenske metode.

Naloga 9.1. Utemeljite, da sta eksplicitna Eulerjeva metoda

$$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1}),$$

in implicitna Eulerjeva metoda

$$y_n = y_{n-1} + hf(x_n, y_n).$$

enostopenjski Runge–Kutta metodi in zapišite njuni Butcherjevi shemi.

Rešitev. Iz formule za eksplicitno Eulerjevo metodo je razvidno, da ustreza enostopenjski Runge–Kutta metodi s parametri $\alpha_1 = 0$, $\beta_{1,1} = 0$ in $\gamma_1 = 1$. Implicitno Eulerjevo metodo pa dobimo, če vzamemo $\alpha_1 = 1$, $\beta_{1,1} = 1$ in $\gamma_1 = 1$, saj je

$$y_n = y_{n-1} + hf(x_n, y_n) = y_{n-1} + hf\left(x_n, y_{n-1} + hf(x_n, y_n)\right).$$

Pripadajoči Butcherjevi shemi sta

$$\begin{array}{c|c} 0 & 0 \\ \hline & 1 \end{array}, \qquad \begin{array}{c|c} 1 & 1 \\ \hline & 1 \end{array}.$$

Naloga 9.2. Izboljšana Eulerjeva metoda je podana z Butcherjevo shemo

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1/2 & 1/2 & 0 \\ \hline & 0 & 1 \end{array}.$$

Z razvojem funkcije y v Taylorjevo vrsto dokažite, da je lokalna napaka te metode reda 3, to je $y(x_n) - y_n = \mathcal{O}(h^3)$ ob predpostavki $y_{n-1} = y(x_{n-1})$.

Rešitev. Oglejmo si Taylorjev razvoj

$$y(x_n) = y(x_{n-1}) + hy'(x_{n-1}) + \tfrac{1}{2}h^2 y''(x_{n-1}) + \mathcal{O}(h^3)$$

funkcije y okoli točke x_{n-1}. Ker je y rešitev diferencialne enačbe, velja

$$\begin{aligned} y'(x_{n-1}) &= f\left(x_{n-1}, y(x_{n-1})\right), \\ y''(x_{n-1}) &= f_x\left(x_{n-1}, y(x_{n-1})\right) + f_y\left(x_{n-1}, y(x_{n-1})\right) y'(x_{n-1}), \end{aligned}$$

zato lahko razvoj ob predpostavki $y_{n-1} = y(x_{n-1})$ zapišemo kot

$$y(x_n) = y_{n-1} + hf_{n-1} + \tfrac{1}{2}h^2\left(f_x(x_{n-1}, y_{n-1}) + f_y(x_{n-1}, y_{n-1})f_{n-1}\right) + \mathcal{O}(h^3),$$

pri čemer smo z f_{n-1} označili vrednost $f(x_{n-1}, y_{n-1})$. Primerjajmo to s približkom y_n, ki ga dobimo z izboljšano Eulerjevo metodo. Po Butcherjevi shemi je

$$y_n = y_{n-1} + hf\left(x_{n-1} + \tfrac{1}{2}h, y_{n-1} + \tfrac{1}{2}hf_{n-1}\right)$$

in iz Taylorjevega razvoja

$$f(x_{n-1} + \tfrac{1}{2}h, y_{n-1} + \tfrac{1}{2}hf_{n-1}) \approx f_{n-1} + \tfrac{1}{2}hf_x(x_{n-1}, y_{n-1}) + \tfrac{1}{2}hf_{n-1}f_y(x_{n-1}, y_{n-1})$$

funkcije f z napako $\mathcal{O}(h^2)$ sledi, da je $y(x_n) - y_n = \mathcal{O}(h^3)$. Kot je posredno razvidno iz izpeljav, je to za red bolje kot pri eksplicitni in implicitni Eulerjevi metodi.

Naloga 9.3. V Matlabu implementirajte funkcije, ki izvedejo eksplicitno, implicitno in izboljšano Eulerjevo metodo za reševanje začetnega problema. Preizkusite jih z reševanjem enačbe

$$y'(x) = 2xy(x) - \frac{y(x)}{x+1}, \quad x > 0,$$

pri začetnem pogoju $y(0) = 1$. Poišite približke za vrednosti y v točkah $x_n = \frac{n}{20}$, $n = 1, 2, \ldots, 30$, ter analizirajte grafe napak (točna rešitev začetnega problema je $y(x) = e^{x^2}/(x+1)$).

Rešitev. Funkcije za izvedbe metod zasnujemo tako, da sprejmejo funkcijo `f` dveh spremenljivk, ki določa diferencialno enačbo, seznam točk `x`, ki vsebuje točke, v katerih iščemo približek za rešitev, in vrednost `y0`, ki predstavlja začetno vrednost (v prvi točki iz seznama `x`). Funkcija za izvedbo eksplicitne Eulerjeve metode je preprosta in v vsakem koraku zanke izračuna približek v naslednji točki.

```
m = length(x) - 1;
h = diff(x);
y = [y0 zeros(1,m)];
for n = 2:m+1
    y(n) = y(n-1) + h(n-1)*f(x(n-1),y(n-1));
end
```

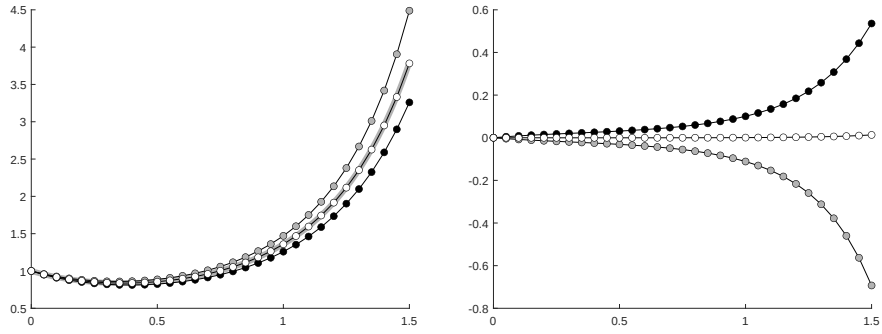
Funkcija, ki izvede implicitno Eulerjevo metodo, v vsakem koraku metode približek za rešitev najprej določi po eksplicitni Eulerjevi metodi, nato pa iterativno računa približek po implicitni Eulerjevi metodi, dokler se približka v iteraciji absolutno ne razlikujeta za manj od `tol`, ki je dodaten parameter funkcije.

```
m = length(x) - 1;
h = diff(x);
y = [y0 zeros(1,m)];
for n = 2:m+1
    y(n) = y(n-1) + h(n-1)*f(x(n-1),y(n-1));
    while true
        yns = y(n);
        y(n) = y(n-1) + h(n-1)*f(x(n),y(n));
        if abs(yns-y(n)) < tol
            break;
        end
    end
end
```

Pri implementaciji izboljšane Eulerjeve metode v vsakem koraku najprej izračunamo premika `k1` in `k2`, nato pa še približek metode.

```
m = length(x) - 1;
h = diff(x);
y = [y0 zeros(1,m)];
for n = 2:m+1
    k1 = f(x(n-1),y(n-1));
    k2 = f(x(n-1)+h(n-1)/2, y(n-1)+h(n-1)*k1/2);
    y(n) = y(n-1) + h(n-1)*k2;
end
```

Opisane funkcije preizkusimo z vhodnimi podatki `f = @(x,y) 2*x.*y - y./(x+1)`, `x = linspace(0,1.5,31)`, `y0 = 1` in `tol = 1e-10`. Rezultate podaja slika 9.1. Slika 9.1a prikazuje izračunane približke, slika 9.1b pa napake (razlike med točno vrednostjo in približki). Črne točke označuje eksplicitno Eulerjevo metodo, sive točke implicitno Eulerjevo metodo, bele točke pa izboljšano Eulerjevo metodo. Siv pas na sliki 9.1a je graf točne rešitve začetnega problema Maksimalna absolutna razlika med točnimi vrednostmi in približki v točkah iz `x` je pri eksplicitni Eulerjevi metodi 0.5358, pri implicitni Eulerjevi metodi 0.6929, pri izboljšani Eulerjevi metodi pa 0.0132. To jasno kaže, da je izboljšana Eulerjeva metoda natančnejša od drugih dveh, kot nakazuje rezultat v nalogi 9.2.



(A) (B)

Slika 9.1: *Primerjava Eulerjevih metod na primeru iz naloge 9.3.*

Naloga 9.4. Heunova metoda je podana z Butcherjevo shemo

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1 & 0 \\ \hline & 1/2 & 1/2 \end{array}.$$

Zapišite približek y_n za vrednost rešitve diferencialne enačbe $y'(x) = -2y(x)$ v odvisnosti od začetnega pogoja $y(0) = y_0$ in ugotovite, kako velik je lahko razmik

h pri izvedbi metode, da se numerična rešitev y_n, ko gre n proti neskončno, obnaša kot točna rešitev, ko gre x proti neskončno.

Rešitev. Po Butcherjevi shemi je približek y_n pri Heunovi metodi podan z

$$y_n = y_{n-1} + \tfrac{1}{2}(k_1 + k_2), \qquad k_1 = hf(x_{n-1}, y_{n-1}), \quad k_2 = hf(x_n, y_{n-1} + hk_1).$$

Če vzamemo $f(x, y) = -2y$, je $k_1 = -2hy_{n-1}$ in $k_2 = -2y_{n-1}h(1-2h)$ ter posledično

$$y_n = y_{n-1} + \tfrac{1}{2}h\left(-2hy_{n-1} - 2y_{n-1}h(1-2h)\right) = y_{n-1}\left(1 - 2h + 2h^2\right).$$

Zato lahko y_n v odvisnosti od y_0 izrazimo kot

$$y_n = y_0\left(1 - 2h + 2h^2\right)^n.$$

Točna rešitev enačbe $y'(x) = -2y(x)$ pri začetnem pogoju $y(0) = y_0$ je $y(x) = y_0e^{-2x}$ in gre proti 0, ko pošljemo x proti neskončno. Če želimo, da se tako obnaša tudi y_n, ko pošljemo n proti neskončno, mora biti $\left|1 - 2h + 2h^2\right| < 1$. Ker je $1-2h+2h^2 > 0$ za vsak $h > 0$, je to ekvivalentno pogoju $h(h-1) < 0$ oziroma $h < 1$.

Naloga 9.5. Izpeljite trapezno metodo za reševanje začetnega problema tako, da enačbo $y'(x) = f(x, y(x))$ integrirate na intervalu $[x_{n-1}, x_n]$, integral $f(x, y(x))$ nadomestite s trapeznim pravilom in izrazite $y(x_n)$. Z zapisom Butcherjeve sheme dokažite, da gre za Runge–Kutta metodo stopnje 2. Podobno kot v nalogi 9.4 obravavajte približek y_n za rešitev enačbe $y'(x) = -2y(x)$ v odvisnosti od začetnega pogoja $y(0) = y_0$.

Rešitev. Z integracijo leve strani diferencialne enačbe dobimo

$$\int_{x_{n-1}}^{x_n} y'(x)\,\mathrm{d}x = y(x_n) - y(x_{n-1}),$$

z integracijo desne strani pa ob uporabi trapeznega pravila

$$\int_{x_{n-1}}^{x_n} f(x, y(x))\,\mathrm{d}x = \frac{1}{2}h\left(f(x_{n-1}, y(x_{n-1})) + f(x_n, y(x_n))\right) + \mathcal{O}(h^3).$$

Če nadomestimo $y(x_{n-1})$ z y_{n-1} in $y(x_n)$ z y_n ter pozabimo na člen $\mathcal{O}(h^3)$, pridemo do formule

$$y_n = y_{n-1} + \frac{1}{2}h\left(f(x_{n-1}, y_{n-1}) + f(x_n, y_n)\right),$$

ki določa trapezno metodo. Iz izpeljave je razvidno, da je podobno kot pri izboljšani Eulerjevi metodi (obravnavani v nalogi 9.2) tudi pri tej metodi lokalna napaka reda 3. Pripadajoča Butcherjeva shema je

$$\begin{array}{c|cc} 0 & 0 & 0 \\ 1 & 1/2 & 1/2 \\ \hline & 1/2 & 1/2 \end{array}.$$

Oglejmo si še, kako se y_n izraža v odvisnosti od y_0 v primeru, ko je $f(x,y) = -2y$. Iz

$$y_n = y_{n-1} + \frac{1}{2}h(-2y_{n-1} - 2y_n) = (1-h)y_{n-1} - hy_n$$

sledi

$$y_n = \frac{1-h}{1+h}y_{n-1} = \left(\frac{1-h}{1+h}\right)^n y_0,$$

kar pomeni, da je $\lim_{n\to\infty} y_n = 0$ neodvisno od $h > 0$. Zaključimo lahko, da je trapezna metoda bolj stabilna od Heunove metode (naloga 9.4), saj pri reševanju tega začetnega problema nimamo omejitve na dolžino koraka. Izkaže se, da je to tudi v splošnem prednost implicitnih metod pred eksplicitnimi.

9.2. Veččlenske metode

Pri numeričnem reševanju začetnega problema bi lahko v rekurzivni formuli, ki določa približek y_n za vrednost rešitve y v točki x_n, poleg približka y_{n-1} uporabili tudi prejšnje približke $(y_{n-2}, y_{n-3}, \ldots)$. To je osnovna ideja veččlenskih metod.

Naloga 9.6. Izpeljite dvočlensko Adams–Bashforthovo metodo za reševanje začetnega problema tako, da enačbo $y'(x) = f(x, y(x))$ integrirate na intervalu $[x_{n-1}, x_n]$, funkcijo $x \mapsto f(x, y(x))$ pa nadomestite z linearno funkcijo, ki jo interpolira v točkah x_{n-2} in x_{n-1} (s tem dosežemo, da je metoda eksplicitna).

Rešitev. Linearno funkcijo, ki interpolira funkcijo $x \mapsto f(x, y(x))$ v točkah x_{n-2} in x_{n-1}, lahko zapišemo kot

$$x \mapsto f\left(x_{n-1}, y(x_{n-1})\right) + \frac{f\left(x_{n-1}, y(x_{n-1})\right) - f\left(x_{n-2}, y(x_{n-2})\right)}{x_{n-1} - x_{n-2}}(x - x_{n-1}).$$

Če slednjo integriramo na intervalu $[x_{n-1}, x_n]$, dobimo

$$hf\left(x_{n-1}, y(x_{n-1})\right) + \frac{1}{2}h^2\frac{f\left(x_{n-1}, y(x_{n-1})\right) - f\left(x_{n-2}, y(x_{n-2})\right)}{x_{n-1} - x_{n-2}}.$$

To pomeni, da je

$$y(x_n) = y(x_{n-1}) + \frac{1}{2}h\left(3f\left(x_{n-1}, y(x_{n-1})\right) - f\left(x_{n-2}, y(x_{n-2})\right)\right) + \mathcal{O}(h^3)$$

in približek za $y(x_n)$ v tej metodi izračunamo kot

$$y_n = y_{n-1} + \frac{1}{2}h(3f(x_{n-1}, y_{n-1}) - f(x_{n-2}, y_{n-2})).$$

Literatura

[1] B. Plestenjak, *Razširjen uvod v numerične metode*, DMFA – založništvo, Ljubljana 2015.

[2] J. Grošelj, *Pozdravljen, Matlab: osnove Matlaba za študente numerične matematike*, Elektronski vir: `https://www.fmf.uni-lj.si/~groseljj/matlab/pozdravljen_matlab.pdf`, Ljubljana 2018.