# Uvod v numerične metode

Luka Horjak (lh0919@student.uni-lj.si)

2. junij 2023

# Kazalo

# Uvod

Ta dokument je povzetek zapiskov pri predmetu Uvod v numerične metode iz leta 2020/21. Povzetek sem napisal v letu 2022/23, predavatelj (in avtor originalnih zapiskov) je v tem letu bil prof. dr. Bor Plestenjak.

Zapiski niso popolni. Manjka večina zgledov, ki pomagajo pri razumevanju definicij in izrekov. Poleg tega nisem dokazoval čisto vsakega izreka, pogosto sem kakšnega označil kot očitnega ali pa le nakazal pomembnejše korake v dokazu.

Zelo verjetno se mi je pri pregledu zapiskov izmuznila kakšna napaka – popravki so vselej dobrodošli.

# 1 Računalniški zapis realnih števil

## 1.1 Predstavljiva števila

**Definicija 1.1.1.** *Numerična metoda* je postopek, ki s končnim številom elementarnih operacij izračuna približek iskane numerične vrednosti.

**Definicija 1.1.2.** *Množica predstavljivih števil* je množica

$$P(b,t,L,U) = \left\{ \pm \sum_{i=1}^{t} c_i b^{e-i} \;\middle|\; \forall i \colon 0 \leq c_i < b \wedge L \leq e \leq U \right\}.$$

Pravimo, da je število *normalizirano*, če v tem zapisu velja $c_1 \neq 0$, sicer pravimo, da je *subnormalizirano*.

**Definicija 1.1.3.** Ob zgornjih oznakah številu $u = \frac{1}{2}b^{1-t}$ pravimo *osnovna zaokrožitvena napaka*.

**Izrek 1.1.4.** Če za realno število $x$ velja

$$\min P(b,t,L,U) \leq |x| \leq \max P(b,t,L,U),$$

velja

$$\min_{y \in P(b,t,L,U)} \frac{|y-x|}{|x|} \leq u.$$

**Definicija 1.1.5.** Število $y \in P(b,t,L,U)$, za katerega je zgornja vrednost najmanjša, označimo s $\mathrm{fl}(x)$.

**Opomba 1.1.5.1.** Za računske operacije po standardu IEEE za poljubno operacijo $* \in \{+,-,\cdot,\div\}$ velja[1]

$$\mathrm{fl}(x * y) = (x * y) \cdot (1+\delta)$$

za nek $\delta \in [-u,u]$. Podobno velja za korenjenje.

[1] Izjema so overflowi in underflowi.

## 1.2 Numerične napake

**Definicija 1.2.1.** Naj bo $f\colon \mathbb{R} \to \mathbb{R}$ funkcija in $x \in \mathbb{R}$ število.

i) *Neodstranljiva napaka* pri računanju vrednosti $f(x)$ je razlika $D_n = f(x) - f(\mathrm{fl}(x))$.

ii) *Napaka metode* je razlika $D_m = f(\mathrm{fl}(x)) - \widetilde{f}(\mathrm{fl}(x))$, kjer je $\widetilde{f}$ numerična metoda za funkcijo $f$.

iii) *Zaokrožitvena napaka* je razlika $D_z = \widetilde{f}(\mathrm{fl}(x)) - g(\mathrm{fl}\, x)$, kjer je $g$ funkcija, ki jo dobimo tako, da vsako operacijo v metodi $\widetilde{f}$ komponiramo s fl.

Skupno napako izračunamo kot $D = D_n + D_m + D_z$.

**Definicija 1.2.2.** Naj bo $f\colon \mathbb{R}^n \to \mathbb{R}^m$ preslikava. *Stopnja občutljivosti* preslikave $f$ v točki $a$ je limita[2]

$$\lim_{x \to 0} \frac{\|f(a+x) - f(a)\|}{\|x\|}.$$

**Opomba 1.2.2.1.** Za odvedljive realne funkcije realne spremenljivke je stopnja občutljivosti kar $|f'(a)|$.

**Definicija 1.2.3.** Naj bo $f\colon R^n \to \mathbb{R}^m$ preslikava in $\widetilde{f}\colon \mathbb{R}^n \to \mathbb{R}^m$ numerična metoda za $f$.

i) Metoda $\widetilde{f}$ je *direktno stabilna*, če je napaka $\frac{\|\widetilde{f}(x) - f(x)\|}{f(x)}$ »majhna«.[3]

ii) Metoda $\widetilde{f}$ je *obratno stabilna*, če je napaka

$$\min_{f(y) = \widetilde{f}(x)} \frac{\|x - y\|}{x}$$

»majhna«.

**Opomba 1.2.3.1.** Direktna napaka je načeloma manjša od obratne napake pomnožene z občutljivostjo metode.

**Trditev 1.2.4.** Pri računanju produkta $p = \prod_{i=0}^{n} x_i$ za relativno zaokrožitveno napako $\gamma$ velja $|\gamma| \leq nu + o(u)$.

*Dokaz.* Produkt izračunamo z naslednjim algoritmom:

1: $p = x_0$
2: **for** $i = 1, \ldots, n$ **do**
3: $\quad p \leftarrow p \cdot x_i$
4: **end for**

Za rezultat $\widetilde{p}$, ki ga na koncu dobimo, tako zaradi zaokroževanja dobimo

$$\widetilde{p} = p \cdot \prod_{i=1}^{n} (1 + \delta_i),$$

kjer je $|\delta_i| \leq u$ za vsak $i$. Tako z Bernoullijevo neenakostjo ocenimo

$$1 - nu \leq (1-u)^n \leq 1 + \gamma \leq (1+u)^n = 1 + nu + o(u). \qquad \square$$

[2] Ob predpostavki, da obstaja.
[3] Omejena z dovolj majhnim številom.

**Opomba 1.2.4.1.** Računanje produkta je direktno stabilno. Ni težko preveriti, da je tudi obratno stabilno.

**Trditev 1.2.5.** Računanje skalarnega produkta je obratno, ne pa direktno stabilno.

*Dokaz.* Skalarni produkt vektorjev $x$ in $y$ izračunamo z naslednjim algoritmom:

1: $s = 0$
2: **for** $i = 1, \ldots, n$ **do**
3: $\quad s \leftarrow s + x_i \cdot y_i$
4: **end for**

Dejanski numerični rezultat je zaradi zaokroževanja enak

$$\widetilde{s} = \sum_{i=1}^{n} x_i y_i (1 + \alpha_i) \cdot \prod_{j=i}^{n} (1 + \beta_j),$$

pri čemer je $\beta_1 = 0$. Označimo

$$1 + \gamma_i = (1 + \alpha_i) \cdot \prod_{j=i}^{n} (1 + \beta_j).$$

Tedaj je $\widetilde{s} = \langle x, \widetilde{y} \rangle$, kjer vzamemo $\widetilde{y}_i = y_i \cdot (1 + \gamma_i)$. Metoda je torej obratno stabilna, saj je $|\gamma_i| \leq nu + o(u)$. Velja pa

$$\frac{|\widetilde{s} - s|}{|s|} \leq \frac{\langle |x|, |y| \rangle}{|\langle x, y \rangle|} \cdot (nu + o(u)),$$

kar ni omejeno navzgor. Če sta vektorja $x$ in $y$ skoraj pravokotna, je lahko napaka zelo velika. Metoda ni direktno stabilna. □

# 2 Nelinearne enačbe

## 2.1 Definicije in bisekcija

**Definicija 2.1.1.** *Nelinearna enačba* je enačba oblike $f(x) = 0$, kjer je $f\colon \mathbb{R} \to \mathbb{R}$ (gladka) funkcija.

**Definicija 2.1.2.** Naj bo $f\colon \mathbb{R} \to \mathbb{R}$ funkcija z ničlo $\alpha$. Predpostavimo, da je $f$ gladka[4] v okolici $\alpha$. Pravimo, da je $\alpha$ $m$-kratna ničla, če je

$$m = \min\left\{n \in \mathbb{N} \;\middle|\; f^{(n)}(\alpha) \neq 0\right\}.$$

**Opomba 2.1.2.1.** Če je $\alpha$ $m$-kratna ničla funkcije $f$, za $x$ v okolici $\alpha$ velja

$$|f(x)| = \frac{\left|f^{(m)}(\xi)\right|}{m!} \cdot |x - \alpha|^m,$$

zato velja

$$|x - \alpha| = \sqrt[m]{\frac{m!\,|f(x)|}{|f^{(m)}(\xi)|}} = O\left(|f(x)|^{\frac{1}{m}}\right).$$

Natančnost računanja ničel torej pada z redom ničle.

**Izrek 2.1.3.** Če je $f\colon [a, b] \to \mathbb{R}$ zvezna funkcija in je $f(a)f(b) < 0$, obstaja tak $c \in (a, b)$, da je $f(c) = 0$.

*Dokaz.* Glej dokaz izreka 3.5.3 v skripti predmeta Analiza 1 v 1. letniku. □

**Opomba 2.1.3.1.** Ničlo poiščemo z metodo *bisekcije*. Algoritem za bisekcijo je naslednji:

```
e = b − a
while e > ε do
    e ← e/2
    c ← a + e
    if sgn(f(c)) = sgn(f(a)) then
        a ← c
    else
        b ← c
    end if
end while
```

Do napake $\varepsilon$ pridemo v

$$k = \left\lceil \log_2\left(\frac{\varepsilon}{b - a}\right)\right\rceil$$

korakih.

[4] Dovoljkrat zvezno odvedljiva.

## 2.2 Navadna iteracija

**Izrek 2.2.1.** Naj bo $g\colon I = [\alpha - \delta, \alpha + \delta] \to \mathbb{R}$ skrčitev s faktorjem $m < 1$ in fiksno točko $\alpha$. Tedaj za vsak $x_0 \in I$ zaporedje, podano z rekurzivno zvezo $x_n = g(x_{n-1})$, konvergira k $\alpha$. Pri tem velja

$$|x_r - \alpha| \leq m^r \, |x_0 - \alpha| \quad \text{in} \quad |x_{r+1} - \alpha| \leq \frac{m}{1-m} \cdot |x_{r+1} - x_r| \, .$$

*Dokaz.* Za dokaz konvergence in prve neenakosti glej dokaz izreka 7.4.2 v skripti predmeta Analiza 1 v 1. letniku. Velja

$$\begin{aligned} |x_{r+1} - \alpha| &= \left| \sum_{n=r+1}^{\infty} x_n - x_{n+1} \right| \\ &\leq \sum_{n=r+1}^{\infty} |x_n - x_{n+1}| \\ &\leq \sum_{n=1}^{\infty} m^n \, |x_r - x_{r-1}| \\ &= \frac{m}{1-m} \cdot |x_r - x_{r-1}| \, . \end{aligned}$$

□

**Posledica 2.2.1.1.** Naj bo $g\colon I \to \mathbb{R}$ zvezno odvedljiva funkcija s fiksno točko $\alpha \in \operatorname{Int} I$. Če je $|g'(\alpha)| < 1$, obstaja tak $\delta > 0$, da za vsak $x_0 \in [\alpha - \delta, \alpha + \delta]$ zaporedje, podano z rekurzivno zvezo $x_n = g(x_{n-1})$, konvergira k $\alpha$.

*Dokaz.* The proof is obvious and need not be mentioned. □

**Definicija 2.2.2** (Navadna iteracija)**.** *Navadna iteracija* je numerična metoda, ki rešitev enačbe $g(x) = x$ aproksimira z uporabo začetnega približka $x_0$ po predpisu $\widetilde{x} = g^n(x_0)$ za dovolj velik $n$.

**Definicija 2.2.3.** Naj bo $g\colon I \to \mathbb{R}$ funkcija s fiksno točko $\alpha \in \operatorname{Int} I$, v kateri je zvezno odvedljiva. Pravimo, da je točka $\alpha$ *privlačna*, če velja $|g'(\alpha)| < 1$, in *odbojna*, če velja $|g'(\alpha)| > 1$.

**Definicija 2.2.4.** Naj zaporedje $(x_n)_{n=1}^{\infty}$ konvergira proti $\alpha$. Pravimo, da je $p$ *red konvergence*, če obstaja taka konstanta $C > 0$, da je

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

**Opomba 2.2.4.1.** Pravimo, da je konvergenca linearna, če je $p = 1$, kvadratična, če je $p = 2$, superlinearna, če je $1 < p < 2$ in podobno za večje $p$. Če je red koncergence enak $p$, število točnih decimalnih mest z iteracijo narašča asimptotsko kot $A^p$.

**Izrek 2.2.5.** Naj bo $g\colon I \to \mathbb{R}$ gladka na $I$ in $\alpha \in \operatorname{Int} I$ njena fiksna točka. Naj bo

$$p = \min \left\{ n \in \mathbb{N} \;\middle|\; g^{(n)}(\alpha) \neq 0 \right\}.$$

Tedaj zaporedje, podano z rekurzivnim predpisom $x_n = g(x_{n-1})$ in začetnim členom v okolici $\alpha$, konvergira z redom $p$.[5]

[5] Pri $p = 1$ zahtevamo še $|g'(\alpha)| < 1$.

*Dokaz.* Po Taylorjevem izreku velja

$$x_{n+1} - \alpha = \frac{g^{(p)}(\xi)}{p!}(x_n - \alpha)^p$$

za nek $\xi$ med $\alpha$ in $x_n$. Sledi, da je

$$\lim_{n\to\infty} \frac{|x_{n+1} - \alpha|}{|x_n - \alpha|^p} = \lim_{n\to\infty} \frac{\left|\frac{g^{(p)}(\xi)}{p!}(x_n - \alpha)^p\right|}{|x_n - \alpha|^p} = \frac{\left|g^{(p)}(\alpha)\right|}{p!}.$$ □

**Definicija 2.2.6.** *Tangentna metoda* je navadna iteracija za funkcijo

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

**Opomba 2.2.6.1.** Če je $\alpha$ enostavna ničla za $f$, sledi

$$g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{f'(\alpha)^2} = 0,$$

zato je red konvergence vsaj kvadratičen.[6] Če je $\alpha$ ničla reda $m \geq 2$, pa se izkaže, da velja

$$g'(\alpha) = 1 - \frac{1}{m},$$

zato je konvergenca linearna.

**Izrek 2.2.7.** Naj bo $f\colon [a, \infty) \to \mathbb{R}$ dvakrat zvezno odvedljiva, naraščajoča in konveksna funkcija z ničlo $\alpha \geq a$. Tedaj je $\alpha$ edina ničla $f$ in za vsak $x_0 \geq a$ tangentna metoda konvergira k $\alpha$.

*Dokaz.* Naj bo $e_n = x_n - \alpha$ za vsak $n \in \mathbb{N}$. Ker velja

$$0 = f(\alpha) = f(x_n) + f'(x_n)(\alpha - x_n) + \frac{f''(\xi_n)}{2}(\alpha - x_n)^2,$$

sledi

$$0 = \frac{f(x_n)}{f'(x_n)} + \alpha - x_n + \frac{f''(\xi_n)}{2f'(x_n}(\alpha - x_n)^2,$$

od koder lahko izrazimo

$$e_{n+1} = \alpha - x_n + \frac{f(x_n)}{f'(x_n)} = \frac{f''(\xi_n)}{2f'(x_n)} \cdot e_n^2.$$

Tako sledi

$$x_{n+1} - \alpha = \frac{f''(\xi_n)}{2f'(x_n)} \cdot e_n^2 \geq 0,$$

zato je $x_k \geq \alpha$ za vse $k \in \mathbb{N}$. Za $n \geq 1$ je tako

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \leq x_n,$$

zato je zaporedje padajoče in posledično konvergentno, saj je omejeno. □

[6] Če velja še $f''(\alpha) = 0$, je red celo vsaj kubičen.

## 2.3 Sekantne metode

**Definicija 2.3.1.** *Sekantna metoda* je numerična metoda, ki približek ničle funkcije $f$ aproksimira z uporabo začetnih približkov $x_0$ in $x_1$ ter rekurzivne zveze

$$x_{n+1} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

**Opomba 2.3.1.1.** Izkaže se, da velja $e_{n+1} \approx ce_ne_{n-1}$, od koder sledi, da je red konvergence enak $p = \frac{1+\sqrt{5}}{2}$, zato je konvergenca superlinearna.

**Definicija 2.3.2.** *Mullerjeva metoda* je numerična metoda, pri kateri naslednji člen zaporedja izračunamo kot ničlo kvadratnega polinoma, ki interpolira vrednosti v prejšnjih treh točkah.

**Opomba 2.3.2.1.** Za red konvergence v tem primeru velja $p^3 = p^2 + p + 1$, oziroma $p \approx 1{,}84$, zato je konvergenca superlinearna.

**Opomba 2.3.2.2.** Za razliko od prejšnjih metod lahko s to metodo za realne funkcije najdemo kompleksne ničle tudi pri realnih začetnih približkih.

**Definicija 2.3.3.** *Inverzna interpolacija* je numerična metoda, pri kateri poiščemo kvadratni polinom $p$, ki interpolira točke $(f(x_n), x_n)$, $(f(x_{n-1}), x_{n-1})$ in $(f(x_{n-2}), x_{n-2})$, in naslednji člen izračunamo po predpisu $x_{n+1} = p(0)$.

**Opomba 2.3.3.1.** Kot pri Mullerjevi metodi je tudi tu red konvergence enak $p \approx 1{,}84$.

**Opomba 2.3.3.2.** Zgoraj naštete metode lahko med seboj seveda tudi kombiniramo.

# 3 Sistemi linearnih enačb

## 3.1 Vektorske in matrične norme

**Definicija 3.1.1.** *Vektorska norma* je preslikava $\|\cdot\| : \mathbb{C}^n \to \mathbb{R}$, za katero za vse $x, y \in \mathbb{C}^n$ in $\alpha \in \mathbb{C}$ velja

i) $\|x\| \geq 0$ z enakostjo natanko tedaj, ko je $x = 0$,

ii) $\|\alpha x\| = |\alpha| \cdot \|x\|$,

iii) $\|x + y\| \leq \|x\| + \|y\|$.

**Trditev 3.1.2.** Za vektor $x \in \mathbb{C}^n$ veljajo ocene

$$\begin{aligned} \|x\|_2 \leq \|x\|_1 &\leq \sqrt{n}\, \|x\|_2 , \\ \|x\|_\infty \leq \|x\|_2 &\leq \sqrt{n}\, \|x\|_\infty , \\ \|x\|_\infty \leq \|x\|_1 &\leq n\, \|x\|_\infty . \end{aligned}$$

**Definicija 3.1.3.** *Matrična norma* je preslikava $\|\cdot\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, za katero za vse $A, B \in \mathbb{C}^{n\times n}$ in $\alpha \in \mathbb{C}$ velja

i) $\|A\| \geq 0$ z enakostjo natanko tedaj, ko je $A = 0$,

ii) $\|\alpha A\| = |\alpha| \cdot \|A\|$,

iii) $\|A + B\| \leq \|A\| + \|B\|$,

iv) $\|A \cdot B\| \leq \|A\| \cdot \|B\|$.

**Definicija 3.1.4.** *Frobeniusova norma* je preslikava

$$\|A\|_F = \left( \sum_{i,j=1}^{n} |a_{i,j}|^2 \right)^{\frac{1}{2}} .$$

**Trditev 3.1.5.** Frobeniusova norma je matrična norma.

*Dokaz.* Prve tri lastnosti matrične norme sledijo iz tega, da je Frobeniusova norma vektorska norma na prostoru $\mathbb{C}^{n^2}$. Velja pa

$$\begin{aligned} \|A \cdot B\|_F^2 &= \left( \sum_{i,j=1}^{n} \left| \sum_{k=1}^{n} a_{i,k} b_{k,j} \right|^2 \right) \\ &\leq \sum_{i,j=1}^{n} \left( \left( \sum_{k=1}^{n} |a_{i,k}|^2 \right) \cdot \left( \sum_{k=1}^{n} |b_{k,j}|^2 \right) \right) \\ &= \sum_{i=1}^{n} \left( \left( \sum_{k=1}^{n} |a_{i,k}|^2 \right) \cdot \left( \sum_{j,k=1}^{n} |b_{k,j}|^2 \right) \right) \\ &= \left( \sum_{i,k=1}^{n} |a_{i,k}|^2 \right) \cdot \left( \sum_{j,k=1}^{n} |b_{k,j}|^2 \right) \\ &= \|A\|_F^2 \cdot \|B\|_F^2 . \end{aligned}$$

□

**Trditev 3.1.6.** Za poljubno vektorsko normo $\|\cdot\|_v$ je tudi

$$\|A\| = \sup_{x \neq 0} \frac{\|Ax\|_v}{\|x\|_v}$$

matrična norma.[7]

*Dokaz.* The proof is obvious and need not be mentioned. □

**Definicija 3.1.7.** Za $p \geq 1$ definiramo

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}.$$

**Trditev 3.1.8.** Velja

$$\|A\|_1 = \max_j \sum_{i=1}^{n} |a_{i,j}| .$$

*Dokaz.* Velja

$$\|Ax\|_1 = \sum_{i=1}^{n} \left| \sum_{j=1}^{n} a_{i,j} x_j \right| \leq \sum_{j=1}^{n} \left( |x_j| \cdot \sum_{i=1}^{n} |a_{i,j}| \right) \leq \|x\|_1 \cdot \max_j \sum_{i=1}^{n} |a_{i,j}| .$$

Ni težko videti, da je enakost dosežena za $x = e_k$, kjer je $k$ indeks, pri katerem je zgornja vsota največja. □

**Trditev 3.1.9.** Velja

$$\|A\|_\infty = \left\|A^\top\right\|_1 .$$

*Dokaz.* Velja

$$\|Ax\|_\infty = \max_i \left| \sum_{j=1}^{n} a_{i,j} x_j \right| \leq \max_i \sum_{j=1}^{n} |a_{i,j} x_j| \leq \max_i \sum_{j=1}^{n} |a_{i,j}| \cdot \|x\|_\infty = \left\|A^\top\right\|_1 \cdot \|x\|_\infty .$$

Enakost je očitno dosežena za[8]

$$x_j = \frac{|a_{k,j}|}{a_{k,j}},$$

kjer je $k$ indeks, pri katerem je zgornja vsota največja. □

**Definicija 3.1.10.** Število $\sigma \geq 0$ je *singularna vrednost* matrike $A$, če je $\sigma^2$ lastna vrednost matrike $A^{\mathsf{H}}A$.

**Opomba 3.1.10.1.** Ker je $\left\langle x, A^{\mathsf{H}}Ax \right\rangle = \langle Ax, Ax \rangle \geq 0$, so vse lastne vrednosti matrike $A^{\mathsf{H}}A$ nenegativne.

**Trditev 3.1.11.** Naj bo $\sigma$ največja singularna vrednost matrike $A$. Tedaj velja

$$\|A\|_2 = \sigma.$$

[7] To je tudi splošna definicija operatorske norme.

[8] Če je $a_{k,j} = 0$, vzamemo $x_j = 1$.

*Dokaz.* Naj bodo $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n$ singularne vrednosti matrike $A$, štete z večkratnostmi. Tedaj obstaja ortonormirana baza lastnih vektorjev $u_i$. Sedaj lahko zapišemo

$$\|Ax\|_2^2 = \left\langle x, A^{\mathsf{H}}x \right\rangle = \left\langle \sum_{i=1}^{n} \alpha_i u_i, \sum_{i=1}^{n} \alpha_i \sigma_i^2 u_i \right\rangle = \sum_{i=1}^{n} \alpha_i^2 \sigma_i^2 \leq \sigma_1^2 \|x\|_2 .$$

Enakost je očitno dosežena v primeru $x = u_1$. □

**Trditev 3.1.12.** Za vsako matriko $A \in \mathbb{C}^{n\times n}$ veljajo ocene

$$\begin{aligned}
\frac{1}{\sqrt{n}} \|A\|_F &\leq \|A\|_2 \leq \|A\|_F , \\
\frac{1}{\sqrt{n}} \|A\|_1 &\leq \|A\|_2 \leq \sqrt{n} \|A\|_1 , \\
\frac{1}{\sqrt{n}} \|A\|_\infty &\leq \|A\|_2 \leq \sqrt{n} \|A\|_\infty .
\end{aligned}$$

**Lema 3.1.13.** Za vsako matrično normo $\|\cdot\|_m$ obstaja taka vektorska norma $\|\cdot\|_v$, da za vsak $x \in \mathbb{C}^n$ in $A \in \mathbb{C}^{n\times n}$ velja $\|Av\|_v \leq \|A\|_m \cdot \|v\|_v$.

*Dokaz.* Vzamemo $\|x\|_v = \|(x, 0, \ldots, 0)\|_m$. □

**Lema 3.1.14.** Za vsako lastno vrednost $\lambda$ matrike $A$ in normo $\|\cdot\|$ velja $|\lambda| \leq \|A\|$.

*Dokaz.* Za inducirano vektorsko normo velja

$$|\lambda| \cdot \|x\| = \|Ax\| \leq \|A\| \cdot \|x\| .$$ □

**Lema 3.1.15.** Frobeniusova in spektralna norma sta invariantni za množenje z unitarno matriko.

*Dokaz.* Velja

$$\|A\|_F^2 = \sum_{i=1}^{n} \|a_i\|^2 = \sum_{i=1}^{n} \|UA_i\|^2 = \|UA\|^2$$

in

$$\|UA\|_2 = \sup_{x\neq 0} \frac{\|UAx\|}{\|x\|} = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \|A\|_2 .$$

Velja tudi $\|AU\|_F = \left\|U^{\mathsf{H}}A^{\mathsf{H}}\right\|_F = \left\|A^{\mathsf{H}}\right\|_F = \|A\|_F$ in podobno za $\|\cdot\|_2$. □

**Lema 3.1.16.** Če je $\|X\| < 1$, je $I - X$ nesingularna in velja

$$(I - X)^{-1} = \sum_{n=0}^{\infty} X^n.$$

Če je $\|I\| = 1$, velja še

$$\left\|(I - X)^{-1}\right\| \leq \frac{1}{1 - \|X\|}.$$

*Dokaz.* Če je $\|X\| < 1$, so absolutne vrednosti lastnih vrednosti matrike $X$ omejene z 1. Sledi, da je $I - X$ obrnljiva. Ker je

$$(I - X)\sum_{i=1}^{n} X^i = I - X^{n+1},$$

sledi

$$\lim_{n\to\infty} \left\| (I - X)\sum_{i=1}^{n} X^i - I \right\| = \lim_{n\to\infty} \left\| X^{n+1} \right\| \leq \lim_{n\to\infty} \|X\|^{n+1} = 0.$$

Za zadnjo neenakost opazimo, da velja

$$\left\| \sum_{i=0}^{\infty} X^i \right\| \leq \sum_{i=0}^{i} \|X\|^i = \frac{1}{1 - \|X\|}. \qquad \square$$

## 3.2 Občutljivost sistema linearnih enačb

**Definicija 3.2.1.** Naj bo $\mathbb{C}^{n\times n}$ nesingularna matrika in $\|\cdot\|$ matrična norma z $\|I\| = 1$. *Občutljivost* matrike $A$ je število

$$K(A) = \|A\| \cdot \left\|A^{-1}\right\|.$$

**Opomba 3.2.1.1.** Velja $K(A) \geq 1$.

**Izrek 3.2.2.** Naj bo $A$ nesingularna matrika in $\Delta A$ taka matrika, da je $\|\Delta A\| \cdot \|A^{-1}\| < 1$. Če je $Ax = b$ in $(A + \Delta A)(x + \Delta x) = b + \Delta b$, velja

$$\frac{\|\Delta x\|}{\|x\|} \leq \frac{K(A)}{1 - K(A)\frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta A\|}{\|A\|} + \frac{\|\Delta b\|}{\|b\|}\right).$$

*Dokaz.* Iz enačb lahko izrazimo

$$b + \Delta A \cdot x + A \cdot \Delta x + \Delta A \cdot \Delta x = b + \Delta b,$$

oziroma

$$\Delta A \cdot x + A \cdot \Delta x + \Delta A \cdot \Delta x = \Delta b.$$

Tako lahko izrazimo

$$A(I + A^{-1}\Delta A)\Delta x = \Delta b - \Delta A \cdot x.$$

Matrika $I + A^{-1}\Delta A$ je po predpostavkah obrnljiva, zato lahko izrazimo

$$\Delta x = (I + A^{-1}\Delta A)^{-1}A^{-1}(\Delta b - \Delta A \cdot x),$$

od koder dobimo

$$\begin{aligned}
\|\Delta x\| &\leq \frac{1}{1 - \|A^{-1}\Delta A\|} \cdot \left\|A^{-1}\right\| \cdot \|b - \Delta A \cdot x\| \\
&\leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\Delta A\|} \cdot (\|\Delta b\| + \|\Delta A\| \cdot \|x\|) \\
&= \frac{\|A\| \cdot \|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|A\| \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta b\|}{\|A\|} + \frac{\|\Delta A\|}{\|A\|} \cdot \|x\|\right) \\
&\leq \frac{K(A)}{1 - K(A) \cdot \frac{\|\Delta A\|}{\|A\|}} \cdot \left(\frac{\|\Delta b\|}{\|b\|} + \frac{\|\Delta A\|}{\|A\|}\right) \cdot \|x\|.
\end{aligned}$$

□

## 3.3 LU razcep

**Definicija 3.3.1.** *LU razcep* matrike $A$ je razcep $A = L \cdot U$, kjer je $L$ spodnjetrikotna matrika, ki ima na diagonali same 1, $U$ pa nesingularna zgornjetrikotna matrika.

**Izrek 3.3.2.** LU razcep matrike $A$ obstaja natanko tedaj, ko so vsi vodilni minorji matrike $A$ neničelni.

*Dokaz.* Naj $P_k$ označuje vodilno $k \times k$ podmatriko matrike $P$. Denimo, da ima matrika $A$ LU razcep. Ni težko opaziti, da je $A_k = L_k \cdot U_k$. Ker sta tako $L_k$ kot $U_k$ nesingularni, je taka tudi $A$.

Naj bodo sedaj vsi vodilni minorji matrike $A$ nesingularni. Izrek dokažemo z indukcijo po dimenziji. Za $n = 1$ je trditev očitna. Sedaj naj bo

$$A = \begin{bmatrix} A' & b \\ c^\top & d \end{bmatrix},$$

kjer je $A' \in \mathbb{C}^{n \times n}$, $b, c \in \mathbb{C}^n$ in $d \in \mathbb{C}$. Naj bo še $A' = LU$. Opazimo, da velja

$$\begin{bmatrix} A' & b \\ c^\top & d \end{bmatrix} = \begin{bmatrix} L & 0 \\ c^\top U^{-1} & 1 \end{bmatrix} \cdot \begin{bmatrix} U & L^{-1}b \\ 0 & d - c^\top U^{-1} L^{-1} b \end{bmatrix}.$$

Označimo $u = d - c^\top U^{-1} L^{-1} b$. Ker velja

$$0 \neq \det A = \det U \cdot u,$$

je $u \neq 0$, zato sta dobljeni matriki nesingularni. Ni težko videti, da je ta razcep enoličen, saj morata zgornja leva bloka biti enaka $L$ in $U$. □

**Opomba 3.3.2.1.** LU razcep matrike dobimo z naslednjim algoritmom:

```
for j = 1 to n − 1 do
    for i = j + 1 to n do
        ℓ_{i,j} = a_{i,j} / a_{j,j}
        for k = j + 1 to n do
            a_{i,k} ← a_{i,k} − ℓ_{i,j} · a_{j,k}
        end for
    end for
end for
```

Z algoritmom v vsakem koraku izračunamo matriko $L_j$ in $U_j$, za kateri velja

$$\prod_{i=0}^{j-1} L_{j-i} \cdot A = U_j.$$

Prvih $j$ stolpcev $U_j$ je pri tem nastavljenih na pravo obliko. Na koncu dobimo

$$A = \prod_{i=1}^{n} L_i^{-1} \cdot U.$$

Časovna zahtevnost algoritma je $\frac{2}{3}n^3 + O(n^2)$.

**Opomba 3.3.2.2.** Sistem linearnih enačb $Ax = b$ z LU razcepom rešimo tako, da izračunamo razcep $A = LU$, nato pa rešimo enačbi $Ly = b$ in $Ux = y$. Rešitvi dobimo eksplicitno kot

$$y_k = b_k - \sum_{i=1}^{k-1} \ell_{k,i} y_i$$

in

$$x_k = \frac{1}{u_{k,k}} \left( y_k - \sum_{i=k+1}^{n} u_{k,i} x_i \right).$$

Časovna zahtevnost reševanja je $\frac{2}{3}n^3 + O(n^2)$.

**Izrek 3.3.3.** Če je $A$ nesingularna matrika, obstaja taka permutacijska matrika $P$, da obstaja LU razcep matrike $PA$.

*Dokaz.* Na vsakem koraku algoritma za LU razcep matriko pomnožimo s permutacijsko matriko $P_k$, ki na mesto $a_{k,k}$ postavi element z največjo absolutno vrednostjo v tem stolpcu.[9] Na koncu dobimo enačbo

$$\prod_{i=0}^{j-1} L_{j-i} P_{j-i} \cdot A = U.$$

Ker $P_i$ komutira z vsemi $L_j$ za $j < i$, lahko enačbo prepišemo v

$$\prod_{i=0}^{j-1} P_{j-i} \cdot A = LU.$$ □

**Opomba 3.3.3.1.** Časovno zahtevnost smo povečali za $O(n^2)$. Pri kompletnem pivotiranjem se ta poveča za $O(n^3)$.

**Opomba 3.3.3.2.** Pri LU razcepu s pivotiranjem velja $\ell_{i,j} \leq 1$.

**Lema 3.3.4.** Naj bo $L$ nesingularna spodnjetrikotna matrika dimenzije $n \times n$. Če sistem $Ly = b$ rešimo s premo susbstitucijo, izračunani vektor $\tilde{y}$ zadošča $(L + \Delta L)\tilde{y} = b$, kjer je $|\Delta L| \leq nu\,|L| + o(u)$.[10]

*Dokaz.* Velja

$$\tilde{y}_i = \frac{b_i - \sum_{k=1}^{i-1} \ell_{i,k} \tilde{y}_k (1 + \delta_{i,k})}{\ell_{i,i}(1 + \alpha_i)(1 + \beta_i)}$$

Pri tem je $|\delta_{i,k}| \leq (i-1)u + o(u)$ in $|\alpha_i|, |\beta_i| \leq u + o(u)$. Če definiramo še $(1 + \delta_{i,i}) = (1 + \alpha_i)(1 + \beta_i)$, velja še $|\delta_{i,i}| \leq nu$.[11] Enačbo lahko sedaj preoblikujemo v

$$b_i = \sum_{k=1}^{i} \ell_{i,k} \tilde{y}_k (1 + \delta_{i,k}),$$

kar je ekvivalentno

$$\Delta L = [\ell_{i,k} \delta_{i,k}]_{i,k}.$$ □

[9] Temu algoritmu pravimo *LU razcep z delnim pivotiranjem.* Če na mesto $a_{k,k}$ postavimo največji element v matriki, dobimo *LU razcep s kompletnim pivotiranjem* in obliko $PAQ = LU$.

[10] Tu absolutne vrednosti in neenakosti gledamo po komponentah.

[11] Velja namreč $\alpha_1 = 0$, zato neenakost drži tudi za $n = 1$.

**Opomba 3.3.4.1.** Podobno velja za zgornjetrikotne matrike.

**Lema 3.3.5.** Naj bo $A$ nesingularna matrika velikosti $n \times n$, pri kateri se izvede LU razcep brez pivotiranja. Za izračunani matriki $\widetilde{L}$ in $\widetilde{U}$ velja

$$\left|A - \widetilde{L}\widetilde{U}\right| \leq nu \left|\widetilde{L}\right| \left|\widetilde{U}\right| + o(u).$$

**Izrek 3.3.6.** Za izračunano vrednost $\tilde{x}$ sistema $Ax = b$ z LU razcepom velja enakost $(A + \Delta A)\tilde{x} = b$, pri čemer je

$$|\Delta A| \leq 3nu\,|L| \cdot |U| + o(u).$$

*Dokaz.* Izrazimo lahko

$$b = (A - E + \Delta L \cdot U + L \cdot \Delta U + \Delta L \cdot \Delta U)\tilde{x},$$

kjer je $E = A - LU$. Z uporabo lem tako dobimo

$$|\Delta A| \leq 3nu\,|L|\,|U| + o(u). \qquad \square$$

**Posledica 3.3.6.1.** Velja

$$\|\Delta A\|_\infty \leq 3nu\,\|L\|_\infty \|U\|_\infty .$$

**Opomba 3.3.6.2.** Če LU razcep računamo brez pivotiranja, je $\|L\|_\infty$ neomejen, zato metoda ni obratno stabilna.

**Definicija 3.3.7.** *Pivotna rast* LU razcepa je količina

$$g = \frac{\max |u_{i,j}|}{\max |a_{i,j}|}.$$

**Lema 3.3.8.** Za LU razcep s pivotiranjem velja

$$\|\Delta A\|_\infty \leq 3gn^3u\,\|A\|_\infty .$$

*Dokaz.* Velja $\|U\|_\infty \leq ng\,\|A\|_\infty$ in $\|L\|_\infty \leq n$. $\square$

**Lema 3.3.9.** Pri delnem pivotiranju je pivotna rast omejena z $2^{n-1}$.

*Dokaz.* Elemente matrike $U$ določimo z ukazom $a_{i,k} \leftarrow a_{i,k} - \ell_{i,j} \cdot a_{j,k}$. Ker je $\ell_{i,j} \leq 1$ in vsak element spremenimo kvečjemu $n-1$-krat, sledi, da je $g \leq 2^{n-1}$. $\square$

**Opomba 3.3.9.1.** LU razcep z delnim pivotiranjem v splošnem ni obratno stabilen, v veliki večini primerov pa je.

**Lema 3.3.10.** Pri kompletnem pivotiranju velja

$$g \leq \left(n \cdot \prod_{k=1}^{n-1} \sqrt[k]{k+1}\right)^{\frac{1}{2}} \approx n^{\frac{1}{2}+\frac{\ln n}{4}}.$$

LU razcep s kompletnim pivotiranjem je obratno stabilen.

## 3.4 Razcep Choleskega

**Izrek 3.4.1.** Naj bo $A$ simetrična pozitivno definitna matrika. Tedaj velja

i) vse njene vodilne podmatrike so pozitivno definitne,

ii) obstaja LU razcep brez pivotiranja za matriko $A$, pri čemer ima matrika $U$ pozitivno diagonalo,

iii) obstaja nesingularna spodnjetrikotna matrika $V$ s pozitivno diagonalo, za katero je $A = VV^\top$.

*Dokaz.*

i) Za $x = (x_1, \ldots, x_k, 0, \ldots, 0)$ velja

$$\langle x, A_k x\rangle = \langle x, Ax\rangle > 0.$$

ii) Vse matrike $A_k$ so pozitivno definitne in zato nesingularne. Velja pa

$$u_{k,k} = \frac{\det U_k}{\det U_{k-1}} = \frac{\det A_k}{\det A_{k-1}} > 0.$$

iii) Naj bo $A = LU$. Po prejšnji točki vemo, da ima $U$ pozitivno diagonalo – naj bo $U = DM$, kjer ima $M$ na diagonali same 1. Sledi, da je

$$L \cdot DM = A = A^\top = M^\top \cdot D^\top L^\top.$$

Iz enoličnosti LU razcepa sledi, da je $L = M^\top$. Sedaj preprosto vzamemo

$$V = L \cdot \sqrt{D}.$$ □

**Opomba 3.4.1.1.** Če velja $A = VV^\top$, je očitno $A$ simetrična pozitivno definitna matrika.

**Definicija 3.4.2.** *Razcep Choleskega* matrike $A$ je razcep $A = V \cdot V^\top$, pri čemer ima $V$ pozitivne diagonalne elemente.

**Opomba 3.4.2.1.** Razcep Choleskega je enolično določen. Po enoličnosti LU razcepa je namreč $V$ enolično določena do skalarnega faktorja natančno, od koder hitro sledi enoličnost.

**Opomba 3.4.2.2.** Razcep Choleskega izračunamo z naslednjim algoritmom:

1: **for** $k = 1$ to $n$ **do**
2: $\quad v_{k,k} = \left(a_{k,k} - \sum_{i=1}^{k-1} v_{k,i}^2\right)^{\frac{1}{2}}$
3: $\quad$ **for** $j = k+1$ to $n$ **do**
4: $\quad\quad v_{j,k} = \frac{1}{v_{k,k}} \left(a_{j,k} - \sum_{i=1}^{k-1} v_{k,i} v_{i,j}\right)$
5: $\quad$ **end for**
6: **end for**

Časovna zahtevnost algoritma je $\frac{1}{3}n^3 + O(n^2)$.

**Opomba 3.4.2.3.** Reševanje sistema $Ax = b$ z razcepom Choleskega je obratno stabilno.

**Opomba 3.4.2.4.** Za simetrične matrike $A$ lahko sistem $Ax = b$ rešujemo z razcepom

$$PAP^\top = LDL^\top,$$

kjer je $P$ permutacijska, $L$ spodnjetrikotna in $D$ diagonalna matrika. Časovna zahtevnost reševanja sistema je pri tem $\frac{1}{3}n^3 + O(n^2)$.

# 4 Nelinearni sistemi

## 4.1 Navadna iteracija

**Izrek 4.1.1.** Naj bo $G\colon \mathbb{R}^n \to \mathbb{R}^n$ zvezno odvedljiva na $\Omega \subseteq \mathbb{R}^n$. Če je $G(\Omega) \subseteq \Omega$ in obstaja tak $m < 1$, da za vse $x \in \Omega$ velja[12]

$$\rho(JG(x)) \leq m,$$

ima $G$ v $\Omega$ natanko eno fiksno točko $\alpha$ in zaporedje $x_n = G^n(x_0)$ konvergira k $\alpha$ za vsak $x_0 \in \Omega$.[13]

*Dokaz.* Naj bo $\varphi(t) = \|F(a + t(b-a)) - F(a)\|$ za preslikavo $F\colon \Omega \to \mathbb{R}^n$. Tedaj je

$$\begin{aligned}\varphi'(t) &= \frac{1}{\varphi(t)} \cdot \sum_{i=1}^{n} (F_i(s) - F_i(a)) \sum_{j=1}^{n} (b_j - a_j) \frac{\partial F_i}{\partial x_j}(s) \\ &= \frac{1}{\varphi(t)} \cdot \langle F(s) - F(a), JF(s) \cdot (b-a) \rangle ,\end{aligned}$$

kjer je $s = a + t(b-a)$. Po Lagrangeevem izreku tako obstaja tak $c \in (0,1)$, da je

$$\varphi(1) - \varphi(0) = \varphi'(c).$$

Za $p = a + c(b-a)$ tako dobimo

$$\|F(b) - F(a)\| = \frac{1}{\|F(p) - F(a)\|} \cdot \langle F(p) - F(a), JF(p) \cdot (b-a) \rangle \leq \|JF(p) \cdot (b-a)\| .$$

Posebej, za $F = G^n$ dobimo

$$\|G^n(b) - G^n(a)\| \leq \|JG(p_n)^n \cdot (b-a)\| \leq \|JG(p_n)^n\| \cdot \|b-a\| .$$

Naj bo $m < m_1 < m_2 < 1$. Za vsak $x$ obstaja tak $c(x)$, da za vse $n \in \mathbb{N}$ velja[14]

$$\|JG(x)^n\| < c(x) \cdot m_1^n.$$

Sedaj naj bo

$$A = \{a + t(b-a) \mid t \in [0,1]\}$$

in

$$U_c = \{x \in A \mid \exists p < m_2 \colon \forall n \in \mathbb{N} \colon \ \|JG(x)^n\| < cp^n\} .$$

To je očitno pokritje množice $A$. Naj bo $x \in U_c$ in $x_1$ v taki okolici $x$, da za matriko $M = JG(x_1) - JG(x)$ velja $\|M\| < m_2 - p$. Tedaj dobimo

$$\|JG(x_1)^n\| \leq \sum_{i=0}^{n} \binom{n}{i} \left\|JG(x)^{n-i}\right\| \left\|M^i\right\| < c \cdot (\|M\| + m_2)^n .$$

Sledi, da je $x_1 \in U_c$, zato je to pokritje odprto. Zaradi kompaktnosti množice $A$ obstaja končno podpokritje. Ekvivalentno obstaja tak $c$, da je

$$\|JG(x)^n\| < cm_2^n$$

[12] Tu $\rho(A)$ označuje *spektralni radij*, za matrike je to absolutna vrednost največje lastne vrednosti.
[13] Tega dokaza ni v profesorjevih zapiskih, za teoretični izpit ga ni potrebno znati.
[14] Glej dokaz leme 6.1.17 iz skripte predmeta Uvod v funkcionalno analizo magistrskega študija.

za vsak $x \in A$ in $n \in \mathbb{N}$. Za vse $p > q$ tako dobimo

$$\|G^p(a) - G^q(a)\| \leq \sum_{k=q}^{p-1} \left\|G^{k+1}(a) - G^k(a)\right\| < c \cdot \|G(a) - a\| \cdot \frac{m_2^q \cdot (1 - m_2^{p-q})}{1 - m_2}.$$

Od tod ni težko videti, da je zaporedje $G^n(a)$ Cauchyjevo in zato konvergentno. Limita je zaradi zveze

$$\|G^n(b) - G^n(a)\| < c m_2^n \|b - a\|$$

neodvisna od izbire začetne točke, za limito $\alpha$ zaporedja pa velja še

$$G(\alpha) = G\left(\lim_{n\to\infty} G^n(a)\right) = \lim_{n\to\infty} G^{n+1}(a) = \alpha. \qquad \square$$

**Opomba 4.1.1.1.** Če je $\alpha = G(\alpha)$ in $\rho(JG(\alpha)) < 1$, je $\alpha$ privlačna fiksna točka. Zadošča že, da velja $\|JG(\alpha)\| < 1$ za neko matrično normo.

**Definicija 4.1.2.** *Seidelova iteracija* je numerična metoda, ki rešitev enačbe $G(x) = x$ aproksimira z začetnim približkom $x^{(0)}$ in rekurzivnim predpisom

$$x_i^{(r+1)} = G_i(x_1^{(r+1)}, \ldots, x_{i-1}^{(r+1)}, x_i^{(r)}, \ldots).$$

**Definicija 4.1.3.** *Newtonova metoda* je navadna iteracija za preslikavo

$$G(x) = x - JF(x)^{-1} \cdot F(x).$$

**Opomba 4.1.3.1.** Metoda zagotovo konvergira, če je začetni približek dovolj dober. Za korak iteracije moramo izračunati $\Delta x$ kot rešitev enačbe $JF(x) \cdot \Delta x = -F(x)$, za kar potrebujemo $O(n^3)$ operacij.

**Lema 4.1.4.** Matrika $A$, za katero je $Ax = y \neq 0$ in ima najmanjšo normo, je oblike

$$A = \frac{yx^\top}{\|x\|^2}.$$

*Dokaz.* Očitno velja $Ax = y$ in $\|A\| \geq \frac{\|y\|}{\|x\|}$. Ker je

$$\left\|yx^\top z\right\| = \|y\| \cdot \left|x^\top z\right| \leq \|y\| \cdot \|x\| \cdot \|z\|$$

z enakostjo, ko je $z = x$, sledi

$$\left\|\frac{yx^\top}{\|x\|^2}\right\| = \frac{\|y\|}{\|x\|}. \qquad \square$$

**Definicija 4.1.5.** *Broydenova metoda* je metoda, pri kateri $\Delta x$ izračunamo po predpisu

$$B_r \Delta x = -F(x).$$

Pri tem za $B_r$ vzamemo matriko, za katero je

$$B_{r+1} \cdot \left(x^{(r+1)} - x^{(r)}\right) = F\left(x^{(r+1)}\right) - F\left(x^{(r)}\right)$$

in je najbližje matriki $B_r$.

**Opomba 4.1.5.1.** Velja

$$B_{r+1} = B_r + \frac{F\left(x^{(r+1)}\right)\left(\Delta x^{(r)}\right)^\top}{\left\|\Delta x^{(r)}\right\|^2}.$$

**Opomba 4.1.5.2.** Za reševanja sistema z $B_{r+1}$ si pomagamo s Sherman-Morrisonovo formulo

$$\left(A + uv^\top\right)^{-1} = A^{-1} + \frac{A^{-1}uv^\top A^{-1}}{1 + \langle v, A^{-1}u\rangle}.$$

# 5 Linearni problemi najmanjših kvadratov

## 5.1 QR razcep

**Definicija 5.1.1.** Naj bo $A \in \mathbb{R}^{m\times n}$ in $b \in \mathbb{R}^m$, kjer je $m \geq n$. *Rešitev po metodi najmanjših kvadratov* sistema $Ax = b$ je vektor $x \in \mathbb{R}^n$, za katerega je $\|Ax - b\|$ minimalna.

**Opomba 5.1.1.1.** Predpostavimo, da je $\operatorname{rang} A = n$, sicer ne dobimo enolične rešitve.

**Izrek 5.1.2.** Naj bo $A \in \mathbb{R}^{m\times n}$ in $b \in \mathbb{R}^m$, kjer je $m \geq n$ in $\operatorname{rang} A = n$. Rešitev normalnega sistema $A^\top Ax = A^\top b$ je rešitev sistema $Ax = b$ po metodi najmanjših kvadratov.

*Dokaz.* Vrednost $\|Ax - b\|$ je minimalna, ko je $Ax - b \perp \operatorname{im} A$. To je ekvivalentno

$$\langle Ax - b, Ay\rangle = 0$$

za vsak $y \in \mathbb{R}^n$, oziroma

$$\left\langle A^\top Ax - A^\top b, y\right\rangle = 0. \qquad \square$$

**Opomba 5.1.2.1.** Sistemu $A^\top Ax = A^\top b$ pravimo *normalni sistem*.

**Opomba 5.1.2.2.** Normalnen sistem lahko rešimo z razcepom Choleskega. Časovna zahtevnost celotnega postopka je $O(n^2m)$.

**Definicija 5.1.3.** *QR razcep* matrike $A \in \mathbb{R}^{m\times n}$ je razcep $A = Q \cdot R$, kjer je $Q \in \mathbb{R}^{m\times n}$ matrika z ortonormiranimi stolpci in $R \in \mathbb{R}^{n\times n}$ zgornjetrikotna matrika s pozitivno diagonalo.

**Izrek 5.1.4.** Naj bo $A \in \mathbb{R}^{m\times n}$, kjer je $m \geq n$ in $\operatorname{rang} A = n$. Tedaj ima matrika $A$ enoličen QR razcep.

*Dokaz.* Če je $A = QR$, je $A^\top A = R^\top Q^\top QR = R^\top R$, zato je $R$ enolično določena z razcepom Choleskega. Od tod lahko izračunamo še $Q = AR^{-1}$. Ker je

$$Q^\top Q = \left(R^\top\right)^{-1} A^\top AR^{-1} = \left(R^\top\right)^{-1} R^\top RR^{-1} = I,$$

tako dobljeni matriki ustrezata pogojem QR razcepa. $\square$

**Opomba 5.1.4.1.** Podobno velja za kompleksne matrike, pri čemer zahtevamo, da je $Q$ unitarna.

**Opomba 5.1.4.2.** QR razcep matrike izračunamo z Gram-Schmidtovo ortogonalizacijo. Algoritem je naslednji:

1: **for** $k = 1$ to $n$ **do**
2: $\quad q_k = a_k$
3: $\quad$ **for** $i = 1$ to $k - 1$ **do**
4: $\quad\quad r_{i,k} = \langle q_i, a_k\rangle$
5: $\quad\quad q_k \leftarrow q_k - r_{i,k}q_i$
6: $\quad$ **end for**
7: $\quad r_{k,k} = \|q_k\|$
8: $\quad q_k \leftarrow \frac{1}{r_{k,k}} q_k$
9: **end for**

**Opomba 5.1.4.3.** V praksi raje uporabljamo *modificirano Gram-Schmidt metodo*. Razlika je v vrstici **??**, kjer ukaz nadomestimo z $r_{i,k} = \langle q_i, q_k \rangle$. Metoda je stabilnejša.

**Trditev 5.1.5.** Naj bo

$$\begin{bmatrix} A & b \end{bmatrix} = \begin{bmatrix} Q & q_{n+1} \end{bmatrix} \cdot \begin{bmatrix} R & z \\ 0 & \rho \end{bmatrix}$$

QR razcep. Rešitev po metodi najmanjših kvadratov sistema $Ax = b$ je vektor $x$, za katerega je $Rx = z$.

*Dokaz.* Izračunamo lahko

$$Ax - b = \begin{bmatrix} A & b \end{bmatrix} \cdot \begin{bmatrix} x \\ -1 \end{bmatrix} = Q(Rx - z) - \rho q_{n+1}.$$

Ker je $q_{n+1} \perp \operatorname{im} Q$, je minimum dosežen, ko je $Q(Rx - z) = 0$. □

**Opomba 5.1.5.1.** Enostavno je videti, da z razcepom $A = QR$ za rešitev po metodi najmanjših kvadratov velja tudi $Rx = Q^\top b$. Izkaže se, da je zgornji postopek stabilnejši od reševanja tega sistema.

**Opomba 5.1.5.2.** Poznamo tudi razširjeni QR razcep – pri tem dobimo $A = \widetilde{Q} \cdot \widetilde{R}$, kjer je $\widetilde{Q}$ kvadratna ortogonalna matrika in $\widetilde{R}$ zgornjetrapezna matrika. Tedaj za

$$A = \widetilde{Q} \cdot \widetilde{R} = \begin{bmatrix} Q & Q_1 \end{bmatrix} \cdot \begin{bmatrix} R \\ 0 \end{bmatrix}$$

velja

$$Ax - b = \begin{bmatrix} Rx - Q^\top b \\ -Q_1^\top b \end{bmatrix}.$$

Minimum je tako dosežen pri $Rx = Q^\top b$ in je enak $\left\| Q_1^\top b \right\|$.

**Definicija 5.1.6.** *Givensova rotacija* je matrika

$$R_{i,j}(c,s)^\top = \begin{bmatrix} I & & & & \\ & c & & s & \\ & & I & & \\ & -s & & c & \\ & & & & I \end{bmatrix},$$

kjer je $(c, s) = (\cos\varphi, \sin\varphi)$.

**Opomba 5.1.6.1.** Če izberemo

$$(c,s) = \frac{1}{\|(a_{i,j}, a_{j,j}\|} \cdot (a_{j,j}, a_{i,j}),$$

ima matrika

$$R_{i,j}(c,s)^\top \cdot A$$

na mestu $(i,j)$ element 0. Z zaporednimi uporabami takih transformacij lahko dobimo modificiran QR razcep matrike $A$.

**Opomba 5.1.6.2.** Algoritem za modificiran QR razcep z Givensovimi rotacijami je naslednji:

1: $Q = I$
2: **for** $i = 1$ to $n$ **do**
3: **for** $k = i + 1$ to $m$ **do**
4: $r = \|(a_{i,i}, a_{k,i})\|$
5: $c = \frac{a_{i,i}}{r}$
6: $s = \frac{a_{k,i}}{r}$
7: $A \leftarrow R_{i,k}(c,s)^\top \cdot A$
8: $b \leftarrow R_{i,k}(c,s)^\top \cdot b$
9: $Q \leftarrow R_{i,k}(c,s)^\top \cdot Q$
10: **end for**
11: **end for**
12: $Q \leftarrow Q^\top$

Pri tem upoštevamo, v vrsticah **??**, **??** in **??** zadošča posodobiti le zadnje stolpce dveh vrstic, kar bistveno zmanjša število potrebnih operacij. Časovna zahtevnost algoritma je $O(6m^2n - n^3)$. Če izpustimo vrstico **??**, se zahtevnost zmanjša na $O(3mn^2 - n^3)$.[15]

**Definicija 5.1.7.** Naj bo $w \in \mathbb{R}^n$ neničeln vektor. Preslikavi

$$P = I - \frac{2}{\|w\|^2} ww^\top$$

pravimo *Householderjevo zrcaljenje*.

**Trditev 5.1.8.** Velja $P = P^\top$ in $P^2 = I$.

*Dokaz.* The proof is obvious and need not be mentioned. □

**Trditev 5.1.9.** Naj bo $x = \alpha w + u$, kjer je $u \perp w$. Tedaj je $Px = -\alpha w + u$.

*Dokaz.* Dovolj je opaziti, da je $w^\top w = \|w\|^2$ in $w^\top u = 0$. □

**Posledica 5.1.9.1.** Če je $\|x\| = \|y\| \neq 0$, za $w = x - y$ velja $Px = y$.

*Dokaz.* Zapišemo lahko

$$x = \frac{x+y}{2} + \frac{w}{2} \quad \text{in} \quad y = \frac{x+y}{2} - \frac{w}{2}.$$

Sedaj je dovolj opaziti, da je $\langle x + y, w \rangle = 0$. □

[15] To naredimo, kadar nas matrika $\widetilde{Q}$ ne zanima.

**Opomba 5.1.9.2.** Householderjeva zrcaljenja lahko uporabimo za izračun QR razcepa. Postopoma v želeno obliko nastavljamo stolpce. Če imamo v $i$-tem stolpcu elemente $a_1, \ldots, a_n$, vzamemo[16]

$$w = (a_i + \operatorname{sgn}(a_i) \|(a_i, \ldots, a_n)\|, a_{i+1}, \ldots, a_n)$$

in matriko $A$ z leve zmnožimo z

$$\begin{bmatrix} I & 0 \\ 0 & P \end{bmatrix}.$$

**Opomba 5.1.9.3.** QR razcep s Householderjevimi zrcaljenji dobimo z naslednjim algoritmom:

1: $Q = I$
2: **for** $i = 1$ to $\min(n, m-1)$ **do**
3: $\quad w = (a_{i,i} + \operatorname{sgn}(a_{i,i}) \|(a_{i,i}, \ldots, a_{n,i})\|, a_{i+1,i}, \ldots, a_{n,i})$
4: $\quad P_i = I - \frac{2}{\|w\|^2} w w^\top$
5: $\quad A \leftarrow P_i \cdot A$
6: $\quad b \leftarrow P_i \cdot b$
7: $\quad Q \leftarrow P_i \cdot Q$
8: **end for**

V vrstici **??** $P_i$ dopolnimo zgoraj levo z identiteto do dimenzije $m \times m$. V vrsticah **??**, **??** in **??** zadošča posodobiti zadnjih $i$ vrstic, kar bistveno zmanjša število operacij. Časovna zahtevnost algoritma je $O\left(4m^2n - \frac{2}{3}n^3\right)$. Če izpustimo vrstico **??**, se zahtevnost zmanjša na $O\left(2mn^2 - \frac{2}{3}n^3\right)$.

**Opomba 5.1.9.4.** Občutljivost rešitve predoločenega sistema po metodi najmanjših kvadratov je odvisna od $K(A) + \|Ax - b\| K(A)^2$.[17] Pri normalnem sistemu dobimo faktor $K(A^\top A) = K(A)^2$, pri reševanju s QR razcepom pa samo $K(R) = K(A)$.

[16] Tu je izjemoma $\operatorname{sgn}(0) = 1$. S popravkom predznaka se izognemo numerični nestabilnosti v primeru, ko je $|a_i| \approx \|(a_i, \ldots, a_n)\|$.

[17] Tu za matrično normo seveda vzamemo kar $\|\cdot\|_2$.

# 6 Problem lastnih vrednosti

## 6.1 Potenčna metoda

**Definicija 6.1.1.** Neničelen vektor $x$ je *levi lastni vektor* matrike $A$, če velja

$$x^{\mathsf{H}} A = \lambda x^{\mathsf{H}}.$$

**Lema 6.1.2.** Če ima matrika $A$ lastno vrednost $\lambda$ z desnim lastnim vektorjem $x$ in lastno vrednost $\mu \neq \lambda$ z levim lastnim vektorjem $y$, je $x \perp y$.

*Dokaz.* Velja

$$\lambda \langle x, y \rangle = y^{\mathsf{H}} A x = \mu \langle x, y \rangle .$$ □

**Izrek 6.1.3** (Schur)**.** Za vsako matriko $A$ obstajata unitarna matrika $U$ in zgornjetrikotna matrika $T$, za kateri je $A = UTU^{\mathsf{H}}$.

*Dokaz.* Naj bo $A = XJX^{-1}$, kjer je $J$ Jordanova forma matrike $A$.[18] Preverimo lahko, da lahko zapišemo $X = QR$, kjer je $Q$ unitarna in $R$ zgornjetrikotna. Tako dobimo

$$A = Q \cdot RJR^{-1} \cdot Q^{\mathsf{H}}.$$ □

**Opomba 6.1.3.1.** Matriki $T$ pravimo *Schurova forma.*

**Opomba 6.1.3.2.** Matrika $S$ je *realna Schurova forma*, če za ortogonalno matriko $Q$ velja $A = QSQ^{\top}$ in je $S$ kvazizgornjetrikotna.[19] Realna Schurova forma obstaja za vse realne matrike $A$.

**Opomba 6.1.3.3.** Lastne vrednosti bi v teoriji lahko računali z Jordanovo formo, a ta metoda ni stabilna.

**Definicija 6.1.4.** *Potenčna metoda* je navadna iteracija za preslikavo $F(x) = \frac{Ax}{\|Ax\|}$.

**Opomba 6.1.4.1.** Algoritem za potenčno metodo je naslednji:

1: $z = z_0$
2: **for** $k = 1$ to $N$ **do**
3: $\quad y = Az$
4: $\quad z \leftarrow \frac{y}{\|y\|}$
5: **end for**

S tem se izognemo dvojnemu računanju $Az$.

**Izrek 6.1.5.** Naj za lastne vrednosti matrike $A$ velja $|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$.[20] Tedaj potenčna metoda konvergira k lastnemu vektorju za $\lambda_1$ za skoraj vse $z_0$.

*Dokaz.* Naj bo $A = XJX^{-1}$, pri čemer je $J$ Jordanova forma in $X = \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}$. Dodatno lahko predpostavimo, da za vsak $i$ velja $Ax_i = \lambda_i x_i$. Sedaj lahko zapišemo

$$A^N x = A^N \sum_{i=1}^{n} \alpha_i x_i = \sum_{i=1}^{n} \left( \sum_{j=i}^{i+k_i-1} \alpha_j \lambda_i^{N+i-j} \frac{N!}{(N+i-j)!} \right) x_i.$$

[18] Na predavanjih je bil izrek dokazan samo za diagonalne $J$.
[19] Na diagonali ima lahko bloke $1 \times 1$ ali $2 \times 2$.
[20] Pravimo, da je $\lambda_1$ *dominantna* lastna vrednost.

Če velja $\alpha_1 \neq 0$, opazimo, da velja

$$\lim_{N\to\infty} \frac{A^N x}{\lambda_1^N} = \alpha_1 x_1,$$

saj je

$$\lim_{N\to\infty} \left|\frac{\lambda_i}{\lambda_1}\right|^N \cdot \frac{N!}{(N-k)!} \leq \lim_{N\to\infty} \left|\frac{\lambda_i}{\lambda_1}\right|^N \cdot N^k = 0$$

za vsak $i \neq 1$, pri $i = 1$ pa dobimo ravno $\alpha_1 \lambda^N x_1$. Sledi, da je

$$\lim_{N\to\infty} \frac{A^n x}{\|A^N x\|} = \frac{\alpha_1 x_1}{\|\alpha_1 x_1\|} = c \cdot x_1. \qquad \square$$

**Opomba 6.1.5.1.** Konvergenca je linearna in je odvisna od $\left|\frac{\lambda_2}{\lambda_1}\right|$.

**Trditev 6.1.6.** Vrednost $\lambda$, ki minimizira $\|Az - \lambda z\|$, reši enačbo

$$\|z\|^2 \lambda = \langle Az, z\rangle .$$

*Dokaz.* Problem je ekvivalenten metodi najmanjših kvadratov za enačbo $z \cdot \lambda = Az$. Sledi, da je

$$z^{\mathsf{H}} z\lambda = z^{\mathsf{H}} Az. \qquad \square$$

**Definicija 6.1.7.** *Rayleighov kvocient* za matriko $A$ in vektor $z \neq 0$ je vrednost

$$\rho(z, A) = \frac{\langle Az, z\rangle}{\|z\|^2}.$$

**Trditev 6.1.8.** Za Rayleighov kvocient veljajo naslednje lastnosti:

i) Za vsak $\alpha \neq 0$ je $\rho(z, A) = \rho(\alpha z, A)$.

ii) Za lastni par $(x, \lambda)$ je $\rho(x, A) = \lambda$.

iii) Minimum izraza $\|Az - \lambda z\|$ je dosežen pri $\lambda = \rho(z, A)$.

*Dokaz.* The proof is obvious and need not be mentioned. $\square$

**Opomba 6.1.8.1.** Rayleighov kvocient uporabimo kot ustavitveni pogoj za potenčno metodo. Dobimo naslednji algoritem:

1: $z = z_0$
2: **while** $\|Az - \rho(z, A)z\| \geq \varepsilon$ **do**
3: $\quad y = Az$
4: $\quad z \leftarrow \frac{y}{\|y\|}$
5: **end while**

**Opomba 6.1.8.2.** Ko najdemo dominantno lastno vrednost, lahko induktivno poiščemo še ostale. Obstaja Householderjevo zrcaljenje $U$, za katero je $Ue_1 = x_1$. Dobimo novo matriko

$$B = U^{\mathsf{H}} AU = \begin{bmatrix} \lambda_1 & b^\top \\ 0 & C \end{bmatrix}.$$

Preostale lastne vrednosti matrike $A$ so ravno lastne vrednosti matrike $C$.

## 6.2 Inverzna in ortogonalna iteracija

**Definicija 6.2.1.** Naj bo $\sigma$ približek lastne vrednosti matrike $A$. *Inverzna iteracija* je potenčna metoda za matriko $(A - \sigma I)^{-1}$.

**Opomba 6.2.1.1.** Lastne vrednosti matrike $(A-\sigma I)^{-1}$ so $\frac{1}{\sigma-\lambda_i}$. Če je $\sigma$ dovolj dober približek za $\lambda$, bo $\frac{1}{\sigma-\lambda}$ dominantna lastna vrednost, zato bo metoda konvergirala k približku lastnega vektorja.

**Lema 6.2.2.** Naj bo $S = \begin{bmatrix} S_1 & S_2 \end{bmatrix}$ nesingularna matrika. Če je

$$B = S^{-1}AS = \begin{bmatrix} B_{1,1} & B_{1,2} \\ B_{2,1} & B_{2,2} \end{bmatrix},$$

stolpci $S_1$ razpenjajo invariantni podprostor za $A$ natanko tedaj, ko je $B_{2,1} = 0$.

*Dokaz.* Velja

$$AS_1 = S_1B_{1,1} + S_2B_{2,1}.$$

Ekvivalenca je zdaj očitna. □

**Definicija 6.2.3.** Invarianten podprostor je *dominanten*, če ga razpenjajo lastni vektorji $x_1, x_2, \ldots, x_p$, katerih lastne vrednosti zadoščajo $|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|$.

**Definicija 6.2.4.** *Ortogonalna iteracija* je metoda, s katero aproksimiramo bazo dominantnega invariatnega podprostora dimenzije $p$ z uporabo začetnega približka $Z_0 \in \mathbb{C}^{n\times p}$ in rekurzivno zvezo $Z_{k+1} = Q_k$, kjer je $AZ_k = Q_kR_k$ QR razcep.

**Izrek 6.2.5.** Denimo, da za lastne vrednosti $A$ velja $|\lambda_1| \geq \cdots \geq |\lambda_p| > |\lambda_{p+1}| \geq \cdots \geq |\lambda_n|$. Tedaj za skoraj vse $Z_0$ ortogonalna iteracija konvergira proti ortonormirani bazi za dominantni invariantni podprostor dimenzije $p$.

*Dokaz.* Naj bo $|\lambda_p| > \lambda > |\lambda_{p+1}|$. Naj bo[21] $A = XJX^{-1}$ in

$$Z_0 = X \cdot \begin{bmatrix} W_1 \\ W_2 \end{bmatrix},$$

pri čemer je $\det W_1 \neq 0$. Sedaj opazimo, da velja

$$\operatorname{im} Z_{k+1} = \operatorname{im} Q_k = \operatorname{im}(AZ_k),$$

zato je

$$\operatorname{im} Z_k = \operatorname{im}(A^kZ_0).$$

Ker je

$$A^kZ_0 = XJ^kW = X \cdot \begin{bmatrix} J_1^kW_1 \\ J_2^kW_2 \end{bmatrix} = X \cdot \begin{bmatrix} I \\ J_2^kW_2W_1^{-1}J_1^{-k} \end{bmatrix} \cdot J_1^kW_1.$$

[21] Na predavanjih je bil izrek dokazan samo za diagonalne $J$.

Matrika $J_1$ je namreč nesingularna, saj je $|\lambda_i| > |\lambda_n| \geq 0$. Ni težko opaziti, da velja

$$\lim_{N\to\infty} \lambda^N J_1^{-N} = \lim_{N\to\infty} \lambda^{-N} J_2^N = 0,$$

zato je

$$\lim_{k\to\infty} J_2^k W_2 W_1^{-1} J_1^{-k} = 0.$$

Sledi, da $A^k Z_0$ konvergira k

$$\begin{bmatrix} X_1 \\ 0 \end{bmatrix}. \qquad \square$$

**Opomba 6.2.5.1.** Podobno opazimo, da če za nek $r < p$ velja $|\lambda_r| > |\lambda_{r+1}|$, tudi prvih $r$ stolpcev $Z_k$ konvergira k ortonormirani bazi dominantnega invariantnega podprostora dimenzije $r$.

**Izrek 6.2.6.** Naj za lastne vrednosti matrike $A$ velja $|\lambda_1| > \cdots > |\lambda_n|$. Tedaj za skoraj vse $Z_0 \in \mathbb{C}^{n\times n}$ zaporedje $Z_k^\top A Z_k$ za $Z_k$ iz ortogonalne iteracije konvergira k Schurovi formi.

*Dokaz.* Za poljuben $p$ naj bo $Z_k = \begin{bmatrix} X_k & Y_k \end{bmatrix}$, kjer je $X_k \in \mathbb{C}^{n\times p}$. Tedaj je

$$Z_k^\top A Z_k = \begin{bmatrix} X_k^\top A X_k & X_k^\top A Y_k \\ Y_k^\top A X_k & Y_k^\top A Y_k \end{bmatrix}.$$

Zaporedje $Y_k^\top A X_k$ konvergira k 0, saj $\operatorname{im} X_k$ konvergira k invariantnemu podprostoru za $A$ in so stolpci matrik $Y_k$ ter $X_k$ pravokotni. Ker to velja za vsak $p$, v limiti dobimo zgornjetrikotno matriko. $\square$

**Opomba 6.2.6.1.** Če je $A$ realna matrika s kompleksnimi lastnimi vrednostmi, ortogonalna iteracija konvergira k realni Schurovi formi.

**Opomba 6.2.6.2.** Časovna zahtevnost enega koraka ortogonalne iteracije je $O(n^3)$, konvergenca poddiagonalnih elementov pa je linearna – odvisna od razmerja $\left|\frac{\lambda_j}{\lambda_i}\right|$.

## 6.3 QR iteracija

**Definicija 6.3.1.** Naj bo $A$ kvadratna matrika. *QR iteracija* je numerična metoda, s katero aproksimiramo Schurovo formo matrike $A$ z zaporedjem, podanim z $A_0 = A$ in rekurzivno zvezo

$$A_{k+1} = R_k Q_k,$$

pri čemer je $A_k = Q_k R_k$ QR razcep.

**Izrek 6.3.2.** Naj bo $A$ nesingularna matrika. Za matriko $A_k$ iz QR iteracije velja $A_k = Z_k^\top A Z_k$, kjer je $Z_k$ matrika, ki jo dobimo pri ortogonalni iteraciji z začetnim približkom $Z_0 = I$.

*Dokaz.* Trditev dokažemo z indukcijo po $k$. Baza $k = 0$ je trivialna. Naj bo $A_k = QR$ in $AZ_k = Z_{k+1}R'$. Tako dobimo

$$QR = A_k = Z_k^\top A Z_k = Z_k^\top Z_{k+1} R'.$$

Ker je $Z_k^\top Z_{k+1}$ ortogonalna, smo dobili nov QR razcep. Iz enoličnosti razcepa sledi, da je $Q = Z_k^\top Z_{k+1}$ in $R = R'$. Tako sledi

$$A_{k+1} = RQ = R' Z_k^\top Z_{k+1} = \left(Z_{k+1}^\top A Z_k\right) Z_k^\top Z_{k+1} = Z_{k+1}^\top A Z_{k+1}. \qquad \square$$

**Posledica 6.3.2.1.** Če za lastne vrednosti matrike $A$ velja $|\lambda_1| > \cdots > |\lambda_n|$, matrike $A_k$ konvergirajo proti Schurovi formi.

*Dokaz.* The proof is obvious and need not be mentioned. $\square$

**Opomba 6.3.2.2.** Časovna zahtevnost enega koraka ortogonalne iteracije je $O(n^3)$, konvergenca poddiagonalnih elementov pa je linearna – odvisna od razmerja $\left|\frac{\lambda_j}{\lambda_i}\right|$.

**Definicija 6.3.3.** Matrika $H$ je *zgornja Hessenbergova*, če je $h_{i,j} = 0$ za vsak $i > j + 1$.

**Trditev 6.3.4.** Če je $A$ zgornja Hessenbergova matrika, je vsak vmesni rezulta QR iteracije zgornja Hessenbergova matrika.

*Dokaz.* Opazimo, da je v QR razcepu $A = QR$ tudi $Q$ zgornja Hessenbergova, zato je taka tudi $RQ$. $\square$

**Opomba 6.3.4.1.** Če pred QR iteracijo $A$ s primerno ortogonalno matriko pretvorimo v Hessenbergovo matriko $H = Q^\top A Q$, je časovna zahtevnost enega koraka QR iteracije samo $O(n^2)$, saj potrebujemo le $n - 1$ Givensovih rotacij. Pretvorbo dosežemo s Householderjevimi zrcaljenji, ki imajo skupaj časovno zahtevnost $O(n^3)$.

**Definicija 6.3.5.** Zgornja Hessenbergova matrika $H$ je *nerazcepna*, če za vsak $i$ velja $h_{i+1,i} \neq 0$.

**Opomba 6.3.5.1.** Če $H$ ni nerazcepna, jo lahko ločimo na dva dela in lastne vrednosti iščemo v vsakem posebej. Numerično pogoj preverimo z

$$|a_{i+1,i}| \leq \varepsilon \cdot \left(|a_{i,i}| + |a_{i+1,i+1}|\right).$$

**Definicija 6.3.6.** Naj bo $A$ kvadratna matrika. *QR iteracija s premiki* je modifikacija QR iteracije, pri kateri za rekurzivno zvezo vzamemo

$$A_{k+1} = R_k Q_k + \sigma_k I,$$

pri čemer je $A_k - \sigma_k I = Q_k R_k$ QR razcep.

**Opomba 6.3.6.1.** Matriki $A_{k+1}$ in $A_k$ sta si še vedno unitarno podobni, saj velja

$$A_{k+1} = Q_k^{\mathsf{H}} A_k Q_k.$$

**Lema 6.3.7.** Naj bo $\sigma$ lastna vrednost nerazcepne zgornje Hessenbergove matrike $A$. Če je $A - \sigma I = QR$ in $B = RQ + \sigma I$, je $b_{n,n-1} = 0$ in $b_{n,n} = \sigma$.

*Dokaz.* Ker je prvih $n-1$ stolpcev matrike $A - \sigma I$ linearno neodvisnih, sledi $r_{i,i} \neq 0$ za vse $i < n$. Sledi, da je $r_{n,n} = 0$, saj je $R$ singularna. Zadnja vrstica matrike $RQ$ je tako ničelna. □

**Opomba 6.3.7.1.** Za izbiro premikov imamo več opcij:

i) Enojni premik – vzamemo $\sigma_k = a_{n,n}^{(k)}$. Ta ne deluje, če niso vse lastne vrednosti matrike $A$ realne.

ii) Dvojni premik[22] – naredimo dva enojna premika, in sicer z lastnima vrednostma matrike

$$\begin{bmatrix} a_{n-1,n-1}^{(k)} & a_{n-1,n}^{(k)} \\ a_{n,n-1}^{(k)} & a_{n,n}^{(k)} \end{bmatrix}.$$

**Opomba 6.3.7.2.** Francisov premih ohranja realnost elementov matrike $A$. Če je $A - \sigma_1 I = QR$, $B = RQ + \sigma_1 I$, $B - \sigma_2 I = Q'R'$ in $C = R'Q' + \sigma_2 I$, velja

$$\begin{aligned} QQ'R'R &= Q(B - \sigma_2 I)R \\ &= Q(Q^{\mathsf{H}} AQ - \sigma_2 I)R \\ &= (AQ - \sigma_2 I)QR \\ &= (AQ - \sigma_2 I)(A - \sigma_1 I) \in \mathbb{R}^{n\times n}. \end{aligned}$$

Dobili smo QR razcep matrike, ki je enolično določen. Ker je matrika realna, je tak tudi razcep, zato je $QQ' \in \mathbb{R}^{n\times n}$. Tako dobimo

$$C = (Q')^{\mathsf{H}} BQ' = (Q')^{\mathsf{H}} Q^{\mathsf{H}} AQQ' \in \mathbb{R}^{n\times n}.$$

**Opomba 6.3.7.3.** Izkaže se, da lahko Francisov premik naredimo v $O(n^2)$ operacijah.

[22] Tudi Francisov

# 7 Polinomska interpolacija

## 7.1 Lagrangeeva interpolacija

**Izrek 7.1.1.** Za paroma različne točke $x_0, \ldots, x_n$ in vrednosti $y_0, \ldots, y_n$ obstaja natanko en polinom $p$, za katerega je $\deg p \leq n$ in $p(x_i) = y_i$ za vse $i$.

*Dokaz.* Če sta $p$ in $q$ dva taka polinoma, ima njuna razlika stopnjo največ $n$, ima pa $n+1$ ničel. Sledi, da imamo kvečjemu en tak polinom. Ni pa težko videti, da

$$p(x) = \sum_{i=0}^{n} y_i \cdot \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}$$

zadošča danim pogojem. □

**Izrek 7.1.2.** Naj bo $f \in \mathcal{C}^{n+1}([a,b])$ funkcija in $x_0, \ldots, x_n \in [a,b]$ paroma različne točke. Naj bo $p$ interpolacijski polinom za $f$ in točke $x_i$. Tedaj za vsak $x \in [a,b]$ obstaja tak $\xi$, da je

$$\min\left(\{x\} \cup \{x_i \mid 0 \leq i \leq n\}\right) < \xi < \max\left(\{x\} \cup \{x_i \mid 0 \leq i \leq n\}\right)$$

in je

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n} (x - x_i).$$

*Dokaz.* Brez škode za splošnost naj bo $x \neq x_i$ za vsak $i$. Definirajmo

$$w(t) = \prod_{i=0}^{n} (t - x_i)$$

in

$$g(t) = f(t) - p(t) - \frac{f(x) - p(x)}{w(x)} w(t).$$

Tako velja $g(x) = 0$ in $g \in \mathcal{C}^{n+1}([a,b])$. Po Rollovem izreku ima $g^{(n+1)}$ vsaj eno ničlo $\xi$ na želenem intervalu. Tako dobimo

$$0 = f^{(n+1)}(\xi) - \frac{f(x) - p(x)}{w(x)} (n+1)!.$$ □

## 7.2 Deljene diference

**Definicija 7.2.1.** Za paroma različne točke $x_0, \ldots, x_k$ in funkcijo $f$ je *deljena diferenca* $[x_0, \ldots, x_k]f$ koeficient pred $x^k$ pripadajočega interpolacijskega polinoma. Če se katera izmed točk ponovi, tam interpoliramo tudi vrednosti odvodov.

**Izrek 7.2.2.** Za paroma različne točke $x_0, \ldots, x_n$ velja

$$p(x) = \sum_{i=0}^{n} \left( [x_0, \ldots, x_i]f \cdot \prod_{j=0}^{i-1} (x - x_j) \right).$$

*Dokaz.* Trditev dokažemo z indukcijo, baza je trivialna. Velja

$$p_{n+1}(x) = p_n(x) + c \prod_{i=0}^{n} (x - x_i),$$

kjer je

$$c = \frac{f^{(k)}(x_{n+1}) - p_n^{(k)}(x_{n+1})}{\left( \prod_{i=0}^{n} (x_{n+1} - x_i) \right)^{(k)}}.$$

Očitno je $c$ vodilni koeficient $p_{n+1}$. □

**Izrek 7.2.3.** Veljajo naslednje trditve:

i) Deljena diferenca $[x_0, \ldots, x_n]f$ je simetrična funkcija.

ii) Deljena diferenca je linearen funkcional.

iii) Če je $x_0 \neq x_n$, velja

$$[x_0, \ldots, x_n]f = \frac{[x_1, \ldots, x_n]f - [x_0, \ldots, x_{n-1}]f}{x_n - x_0}$$

*Dokaz.* Simetričnost in linearnost sta očitni. Če je $q$ interpolacijski polinom za točke $x_0, \ldots, x_{n-1}$ in $r$ interpolacijski polinom za $x_1, \ldots, x_n$, velja

$$p(x) = \frac{x - x_n}{x_0 - x_n} q(x) + \frac{x - x_0}{x_n - x_0} r(x).$$

S primerjavo koeficientov dobimo rekurzivno formulo. □

**Opomba 7.2.3.1.** Če je $x_i = x_0$ za vsak $i$, deljeno diferenco namesto z rekurzivno formulo izračunamo kot

$$[x, \ldots, x]f = \frac{f^{(n)}(x)}{n!}.$$

**Izrek 7.2.4.** Za vsako funkcijo $f \in \mathcal{C}^k([a, b])$ velja

$$[x_0, \ldots, x_k]f = \int_0^1 \int_0^{t_1} \cdots \int_0^{t_{k-1}} f^{(k)} \left( x_0 + \sum_{i=1}^{k} t_i (x_i - x_{i-1}) \right) dt_k \, dt_{k-1} \ldots dt_1$$

*Dokaz.* Izrek očitno velja za $k = 0$. Najprej predpostavimo, da je $x_k \neq x_{k-1}$. Tedaj sledi

$$\int_0^{t_{k-1}} f^{(k)}\left(X + t_k(x_k - x_{k-1})\right)\, dt_k = \frac{f^{(k-1)}\left(X + t_{k-1}(x_k - x_{k-1})\right) - f^{(k-1)}\left(X\right)}{x_k - x_{k-1}},$$

kjer je

$$X = x_0 + \sum_{i=1}^{k-1} t_i(x_i - x_{i-1}).$$

Po indukcijski predpostavki je ta izraz enak

$$\frac{[x_0, \ldots, x_{k-2}, x_k]f - [x_0, \ldots, x_{k-1}]f}{x_k - x_{k-1}} = [x_0, \ldots, x_k]f.$$

Če je $x_i = x_0$ za vsak $i$, pa dobimo

$$\int_0^1 \int_0^{t_1} \cdots \int_0^{t_{k-1}} f^{(k)}\left(x_0\right)\, dt_k\, dt_{k-1} \ldots\, dt_1 = \frac{f^{(k)}(x_0)}{k!}.$$ □

**Posledica 7.2.4.1.** Za vsako funkcijo $f \in \mathcal{C}^k([a, b])$ in točke $x_0, \ldots, x_n$ velja

$$[x_0, \ldots, x_n]f = \frac{f^{(k)}(\xi)}{k!}$$

za $\xi \in (\min\{x_i \mid 0 \leq i \leq n\}, \max\{x_i \mid 0 \leq i \leq n\})$.

*Dokaz.* Volumen množice, po kateri integriramo, je enak $\frac{1}{k!}$. Izrek tako sledi iz izreka o vmesni vrednosti in zveznosti funkcije $f^{(k)}$. □

**Izrek 7.2.5.** Za funkcijo $f$ in interpolacijski polinom $p$ na točkah $x_0, \ldots, x_n$ velja

$$f(x) = p(x) + [x_0, \ldots, x_n, x]f \cdot \prod_{i=0}^{n}(x - x_i).$$

*Dokaz.* Desna stran enačbe je interpolacijski polinom za $f$ na točkah $x_0, \ldots, x_n, x$. □

**Posledica 7.2.5.1.** Za $f \in \mathcal{C}^{n+1}([a, b])$ naj bo $p$ njen interpolacijski polinom za točke $x_0, \ldots, x_n$.[23] Tedaj za vsak $x \in [a, b]$ obstaja tak $\xi$, da velja

$$\min\left(\{x\} \cup \{x_i \mid 0 \leq i \leq n\}\right) \leq \xi \leq \max\left(\{x\} \cup \{x_i \mid 0 \leq i \leq n\}\right)$$

in

$$f(x) - p(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{i=0}^{n}(x - x_i).$$

*Dokaz.* The proof is obvious and need not be mentioned. □

**Opomba 7.2.5.2.** To je pravzaprav posplošitev Taylorjevega izreka.

[23] Ne nujno različne.

**Opomba 7.2.5.3.** Boljšo aproksimacijo funkcije lahko poskusimo dobiti tako, da povečamo število interpolacijskih točk. Včasih je pri tem namesto ekvidistantnih smiselno vzeti Čebiševe točke

$$x_i = \lambda \cos\left(\frac{j\pi}{n}\right) + \mu.$$

**Opomba 7.2.5.4.** Kljub večjemu številu točk polinomi pogosto ne konvergirajo proti $f$. V takem primeru je bolj smiselno $f$ aproksimirati odsekoma. Na odsekih lahko vzamemo linearne funkcije, lahko pa vzamemo polinome tretje stopnje in dosežemo, da je dobljena funkcija zvezno odvedljiva.

# 8 Numerično integriranje

## 8.1 Kvadraturne formule

**Definicija 8.1.1.** *Kvadraturna formula* je enačba

$$\int_a^b f(x)\,dx = \sum_{i=0}^{n} \alpha_i f(x_i) + R(f),$$

Pri čemer velja

$$\alpha_i = \int_a^b \prod_{j\neq i} \frac{x - x_j}{x_i - x_j}\,dx.$$

Tu so $\alpha_i$ uteži, $x_i$ vozli in $R(f)$ napaka.

**Opomba 8.1.1.1.** Kvadratna formula je vedno točna za polinome stopnje največ $n$. Če izberemo primerne vozle, lahko dosežemo, da je formula točna za polinome stopnje največ $2n + 1$.[24]

**Definicija 8.1.2.** *Newton-Cotesove formule* so kvadratne formule, pri katerih za vozle vzamemo ekvidistantne točke

$$x_i = a + ih,$$

kjer je $h = \frac{b-a}{n}$. Formula je *zaprta*, če sta krajišči intervala vozla, sicer je *odprta*.

**Opomba 8.1.2.1.** Naj bo $f \in \mathcal{C}^2([a, b])$. Za $n = 1$ dobimo

$$f(x) = f(a) \cdot \frac{x - b}{a - b} + f(b) \cdot \frac{x - a}{b - a} + \frac{f''(\xi_x)}{2}(x - a)(x - b),$$

od koder z izrekom o vmesni vrednosti izpeljemo

$$\int_a^b f(x)\,dx = \frac{h}{2} \cdot (f(a) + f(b)) - \frac{h^3}{12} f''(\xi).$$

**Opomba 8.1.2.2.** Če je $n$ sod, ne Newton-Cotesova formula točna za polinome stopnje največ $n + 1$.

**Opomba 8.1.2.3.** Iz Peanovega izreka lahko izpeljemo, da za $f \in \mathcal{C}^r([a, b])$ velja

$$R(f) = ch^m f^{(r)}(\xi),$$

kjer je $r$ najnižja stopnja polinoma, za katerega je $R \neq 0$.

**Opomba 8.1.2.4.** Naj bo $f \in \mathcal{C}^3([a, b])$. Za $n = 2$ dobimo[25]

$$\int_a^b f(x)\,dx = \frac{h}{3} \cdot (f(x_0) + 4f(x_1) + f(x_2)) + \int_a^b \frac{f'''(\xi_x)}{6} \cdot (x - x_0)(x - x_1)(x - x_2)\,dx.$$

Z uporabo prejšnje opombe lahko izpeljemo še

$$R(f) = -\frac{h^4}{90} \cdot f^{(4)}(\xi).$$

[24] Gaussova kvadraturna formula, glej poglavje **??**.
[25] Tudi Simpsonovo pravilo.

**Opomba 8.1.2.5.** Napaka metode kvadraturnih formul je enaka $R(f)$. Če je $\widetilde{f(x_i)}$ naš približek za $f(x_i)$ in velja $\left|\widetilde{f(x_i)} - f(x_i)\right| < \varepsilon$, lahko neodstranljivo napako ocenimo kot

$$|D_n| = \left|\sum_{i=0}^{n} \alpha_i \cdot \left(\widetilde{f(x_i)} - f(x_i)\right)\right| < \varepsilon \cdot \sum_{i=0}^{n} |\alpha_i| .$$

Če imajo vse $\alpha_i$ enak predznak, sledi $|D_n| < \varepsilon \cdot (b-a)$.

**Definicija 8.1.3.** *Sestavljeno pravilo* za integriranje je enačba, ki jo dobimo tako, da integral odsekoma aproksimiramo s kvadraturno formulo.

**Opomba 8.1.3.1.** Za odsekoma linearno aproksimiacijo dobimo sestavljeno trapezno pravilo

$$\int_a^b f(x)\,dx = \frac{h}{2} \cdot \left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right) - \sum_{i=0}^{n-1} \frac{h^3}{12} f''(\xi_i),$$

kar lahko z izrekom o povprečni vrednosti poenostavimo do

$$\frac{h}{2} \cdot \left(f(x_0) + 2\sum_{i=1}^{n-1} f(x_i) + f(x_n)\right) - (b-a) \cdot \frac{h^2}{12} f''(\xi).$$

**Opomba 8.1.3.2.** Za odsekoma kvadratne polinome pri sodih $n$ dobimo sestavljeno Simpsonovo pravilo

$$\int_a^b f(x)\,dx = \frac{h}{3} \cdot \left(f(x_0) + 4\sum_{\substack{0<i<n \\ 2\nmid i}} f(x_i) + 2\sum_{\substack{0<i<n \\ 2\mid i}} f(x_i) + f(x_n)\right) - (b-a) \cdot \frac{h^4}{180} f^{(4)}(\xi).$$

## 8.2 Gaussove kvadraturne formule

**Trditev 8.2.1.** Naj bo $\rho$ nenegativna funkcija. Tedaj je

$$\langle f, g\rangle = \int_a^b f(x)\overline{g(x)}\rho(x)\,dx$$

skalarni produkt.

*Dokaz.* The proof is obvious and need not be mentioned. □

**Definicija 8.2.2.** Naj bodo $p_n(x) = x^n$ polinomi. Polinomom, ki jih dobimo z Gram-Schmidtovo ortogonalizacijo na $p_i$ z zgornjim skalarnim produktom pravimo *ortogonalni polinomi* za interval $[a, b]$ in utež $\rho$.

**Lema 8.2.3.** Vse ničle ortogonalnih polinomov so enostavne, realne in ležijo znotraj intervala $(a, b)$.

*Dokaz.* Naj bo $p$ ortogonalen polinom in

$$q(x) = \prod_{i=1}^{k}(x - x_i)^{\alpha_i},$$

kjer so $x_i$ vse ničle polinoma $p$, ki ležijo na $(a, b)$, in

$$\alpha_i = \begin{cases} 1, & 2 \nmid \operatorname{ord}_{x_i}(p), \\ 0, & 2 \mid \operatorname{ord}_{x_i}(p). \end{cases}$$

Očitno je $\langle p, q\rangle \neq 0$, zato je $\deg p = \deg q$. Sledi, da vse ničle polinoma $p$ ležijo na intervalu $(a, b)$ in so enostavne. □

**Definicija 8.2.4.** *Gaussova kvadraturna formula* je kvadraturna formula, pri kateri za vozle vzamemo ničle ortogonalnega polinoma $p_{n+1}$ (z utežjo 1).

**Trditev 8.2.5.** Naj bo $f$ polinom stopnje največ $2n + 1$. Tedaj je Gaussova kvadraturna formula točna za $f$.

*Dokaz.* Naj bo $q$ ortogonalen polinom stopnje $n + 1$. Vozli so njegove ničle. Za polinoma $g$ in $h$ stopnje največ $n$ lahko zapišemo $f = gq + h$. Tako sledi

$$\int_a^b f(x)\rho(x)\,dx = \int_a^b g(x)q(x)\rho(x)\,dx + \int_a^b h(x)\rho(x)\,dx = 0 + \sum_{i=0}^{n}\alpha_i h(x_i) = \sum_{i=0}^{n}\alpha_i f(x_i). \quad \square$$

**Lema 8.2.6.** Uteži Gaussovih kvadraturnih formul so pozitivne.

*Dokaz.* Naj bo $q$ ortogonalen polinom. Polinom $p_k(x) = \frac{q(x)^2}{(x-x_k)^2}$ je stopnje $2n$, zato je

$$0 < \int_a^b p_k(x)\rho(x)\,dx = \sum_{i=0}^{n}\alpha_i p_k(x_i) = \alpha_k p_k(x_k). \quad \square$$

# 9 Navadne diferencialne enačbe

## 9.1 Eulerjeva metoda

**Definicija 9.1.1.** Naj bo $D = [a, b] \times \Omega \subseteq \mathbb{R}^2$. Funkcija $f \colon D \to \mathbb{R}$ je *Lipschitzova* glede na $y$ s konstanto $L$, če za vsaka $(x, y_1)$ in $(x, y_2)$ v $D$ velja

$$|f(x, y_1) - f(x, y_2)| \leq L \cdot |y_1 - y_2| .$$

**Opomba 9.1.1.1.** Če je $f$ Lipschitzova glede na $y$, ima navadna diferencialna enačba $y' = f(x, y)$ z začetnim pogojem $y(x_0) = y_0$ enolično lokalno rešitev ne glede na $x_0$ in $y_0$.

**Opomba 9.1.1.2.** Občutljivost problema določimo z zvezo

$$|\tilde{y}(x) - y(x)| \leq e^{L(x-x_0)} |\tilde{y}_0 - y_0| + \frac{e^{L(x-x_0)}}{L} \cdot \left\| \tilde{f} - f \right\|_\infty .$$

**Definicija 9.1.2.** *Eksplicitna Eulerjeva metoda* je numerična metoda, s katero izračunamo približek vrednosti $y$ v točki $x_n = x_0 + nh$. Približke vrednosti v točkah $x_k$ dobimo tako, da se premaknemo v smeri tangente.

**Opomba 9.1.2.1.** Algoritem za Eulerjevo metodo je naslednji:

1: $y = y_0$
2: **for** $i = 0$ to $n - 1$ **do**
3: $\quad y \leftarrow y + hf(x + ih, y)$
4: **end for**

**Opomba 9.1.2.2.** Poznamo tudi *implicitno Eulerjevo metodo*, pri kateri $y_{n+1}$ izračunamo iz zveze

$$y_{n+1} = y_n + hf(x + ih, y_{n+1}).$$

**Definicija 9.1.3.** *Taylorjeva metoda* je numerična metoda, podobna Eulerjevi. Pri tem $y(x + h)$ izračunamo kot

$$y(x + h) = \sum_{n=0}^{N} \frac{h^n}{n!} \cdot y^{(n)}(x).$$

**Definicija 9.1.4.** Pravimo, da ima metoda lokalno napako reda $k$, če se pri točni vrednosti $y(x_n)$ izračunani približek $y_{n+1}$ ujema z razvojem $y(x_n + h)$ v Taylorjevo vrsto okoli $x_n$ do vključno člena s $h^k$.

**Opomba 9.1.4.1.** Ker velja $y'' = f_x + f \cdot f_y$, lahko odvode $y$ izrazimo z odvodi funkcije $f$.

## 9.2 Metode tipa Runge-Kutta

**Definicija 9.2.1.** *Metoda Runge-Kutta* je numerična metoda, s katero izračunamo približek vrednosti $y$ v točki $x_n = x_0 + nh$. Približke vrednosti $y$ v točkah $x_k$ dobimo tako, da izračunamo koeficiente

$$k_i = hf\left(x_n + \alpha_i h, y_n + \sum_{j=1}^{m} \beta_{i,j} k_j\right)$$

in rekurzivno izračunamo

$$y_{n+1} = y_n + \sum_{i=1}^{m} \gamma_i k_i.$$

**Opomba 9.2.1.1.** Metoda je eksplicitna, če velja $\beta_{i,j} = 0$ za vse $j \geq i$.

**Opomba 9.2.1.2.** Da ima metoda lokalno napako reda 1, mora veljati

$$\alpha_i = \sum_{j=1}^{i} \beta_{i,j} \quad \text{in} \quad \sum_{i=1}^{m} \gamma_i = 1.$$