Numerične metode 1 - definicije, trditve in izreki

Oskar Vavtar
po predavanjih profesorja Bora Plestenjaka

2020/21

Kazalo

1 NUMERIČNO RAČUNANJE

1.1 Uvod

Definicija 1.1 (Napaka)**.** Pri numeričnem računanju izračunamo numerični približek za točno rešitev. Razlika med približkom in točno vrednostjo je *napaka* približka. Ločimo *absolutno* in *relativno* napako.

- absolutna napaka = približek − točna vrednost
- relativna napaka = $\frac{\text{absolutna napaka}}{\text{točna vrednost}}$

Naj bo x točna vrednost, $\hat{x}$ pa približek za x.

- Če je $\hat{x} = x + d_a$, potem je $d_a = \hat{x} - x$ *absolutna napaka.*
- Če je $\hat{x} = x(1 + d_r)$, potem je $d_r = \frac{\hat{x}-x}{x}$ *relativna napaka.*

1.2 Premična pika

Definicija 1.2. Velja $\mathrm{fl}(x) = x(1 + \delta)$ za $|\delta| \leq u$, kjer je

$$u = \frac{1}{2}b^{1-t}$$

osnovna zaokrožitvena napaka:

- single: $u = 2^{-24} = 6 \times 10^{-8}$
- double: $u = 2^{-53} = 1 \times 10^{-16}$

Izrek 1. Če za število x velja, da $|x|$ leži na intervalu med najmanjšim in največjim *pozitivnim predstavljivim normaliziranim* številom, potem velja

$$\frac{|\mathrm{fl}(x) - x|}{|x|} \leq u.$$

1.3 Občutljivost problema

Definicija 1.3 (Občutljivost)**.** Če se rezultat pri majhni spremembi argumentov[1] ne spremeni veliko, je problem *neobčutljiv*, sicer pa je *občutljiv*.

1.4 Vrste napak pri numeričnem računanju

Definicija 1.4. Pri numeričnem računanje se pojavijo 3 vrste napak:

1. NEODSTRANLJIVE NAPAKE: Npr. ko podatek ni predstavljivo število. Namesto $y = f(x)$ lahko v najboljšem primeru izračunamo $\overline{y} = f(\overline{x})$, kjer je $\overline{x}$ najbližje predstavljivo število.

$$D_n = y - \overline{y} = f(x) - f(\overline{x})$$

2. NAPAKA METODE: Npr. ko na voljo nimamo željene operacije. Namesto $f(\overline{x})$ potem izračunamo $\widetilde{y} = g(\overline{x})$, kjer je $g(x)$ približek za $f(x)$, kjer znamo vrednost g izračunati s končnim številom operacij.

$$D_m = \overline{y} - \widetilde{y} = f(\overline{x}) - g(\overline{x})$$

3. ZAOKROŽITVENE NAPAKE: Pri izračunu $\widetilde{y} = g(\overline{x})$ lahko pri vsaki osnovni operaciji pride do zaokrožitvene napake, zato na koncu kot numeričen rezultat dobimo $\widehat{y}$.

$$D_z = \widetilde{y} - \widehat{y}$$

Skupna napaka:

$$D = D_n + D_m + D_z$$

V splošnem lahko ocenimo:

$$|D| \leq |D_n| + |D_m| + |D_z|$$

[1] *Motnji* oz. *perturbaciji.*

1.5 Stabilnost metode

Definicija 1.5. Če metoda za $\forall x$ vrne $\widehat{y}$, ki je *absolutno (relativno)* blizu točnemu y, je metoda *direktno stabilna.*

Če metoda za $\forall x$ vrne tak $\widehat{y}$, da $\exists \widehat{x}$ absolutno (relativno) blizu x, da je $\widehat{y} = f(\widehat{x})$ (točno), je metoda *obratno stabilna.*

V splošnem:

$$|\text{direktna napaka}| \;\leq\; |\text{občutljivost}| \cdot |\text{obratna napaka}|$$

1.6 Analiza zaokrožitvenih napak

1.6.1 Produkt $n+1$ predstavljivih števil

Algoritem. Dana so predstavljiva števila $x_0, x_1, \ldots, x_n$;
računamo $p = x_0 \cdot x_1 \cdot \ldots \cdot x_n$.

TOČNO:

```
p₀ = x₀
i = 1,...,n
    pᵢ = pᵢ₋₁·xᵢ
p = pₙ
```

NUMERIČNO:

$$
\begin{aligned}
&\widehat{p}_0 = x_0\\
&i = 1,\ldots,n\\
&\quad \widehat{p}_i = \widehat{p}_{i-1} \cdot x_i \cdot (1 + \delta_i) \quad |\delta_i| \leq u\\
&\widehat{p} = \widehat{p}_n
\end{aligned}
$$

1.6.2 Skalarni produkt vektorjev dolžine n

Algoritem. Imamo vektorje *predstavljivih* števil $a = [a_1, \ldots, a_n]^T$, $b = [b_1, \ldots, b_n]^T$. Računamo $s = \langle b^T, a\rangle = \sum_{i=1}^{n} a_i b_i$.

TOČNO:

$$
\begin{aligned}
&\mathtt{s_0 = 0}\\
&\mathtt{i = 1,...,n}\\
&\qquad \mathtt{p_i = a_i \cdot b_i}\\
&\qquad \mathtt{s_i = s_{i-1} + p_i}\\
&\mathtt{s = s_n}
\end{aligned}
$$

NUMERIČNO:

$$
\begin{aligned}
&\widehat{\mathtt{s}}_0 = 0\\
&\mathtt{i = 1,...,n}\\
&\qquad \widehat{\mathtt{p}}_i = \mathtt{a_i \cdot b_i \cdot (1 + } \alpha_i) && |\alpha_i| \le u\\
&\qquad \widehat{\mathtt{s}}_i = (\widehat{\mathtt{s}}_{i-1} + \widehat{\mathtt{p}}_i)\cdot(1 + \beta_i) && |\beta_i| \le u\\
&\widehat{\mathtt{s}} = \widehat{\mathtt{s}}_n
\end{aligned}
$$

Trditev 1. Računanje skalarnega produkta je *obratno* stabilno, ni pa *direktno* stabilno.

2 NELINEARNE ENAČBE

2.1 Uvod

Definicija 2.1. Naj bo α ničla funkcije f, ki je *zvezno odvedljiva* v okolici α:

- $f'(\alpha) \neq 0$: α je *enostavna* ničla
- $f'(\alpha) = 0$: α je *večkratna* ničla

  Če je f m-krat zvezno odveljiva in
  $$f'(\alpha) \;=\; f''(\alpha) \;=\; \ldots \;=\; f^{(m-1)}(\alpha) \;=\; 0, \quad f^{(m)} \neq 0,$$
  je α m-kratna ničla.

Trditev 2 (Občutljivost ničel)**.** Naj bo α enostavna ničla. Če v okolici $x = \alpha$ obstaja inverzna funkcija $\alpha = f^{-1}(0)$ v bistvu "računamo"vrednost inverzne funkcije. Občutljivost je enaka absolutni vrednosti odvoda inverzna funkcije:
$$|(f^{-1})'(0)| \;=\; \frac{1}{|f'(f^{-1}(0))|} \;=\; \frac{1}{|f'(\alpha)|}.$$

Večkratno ničlo lahko izračunamo le z natančnostjo $u^{\frac{1}{m}}$, kjer je m večkratnost ničle (za dvojno ničlo dobimo le polovico točnih decimalk, za trojno le tretjino...).

2.2 Bisekcija

Izrek 2. Če je f realna zvezna funkcija na $[a, b]$ in je $f(a) \cdot f(b) < 0$, potem $\exists \xi \in (a, b)$, da je $f(\xi) = 0$.

Algoritem (Bisekcija)**.** Naj velja $f(a) \cdot f(b) < 0$ in $a < b$:

```
e = b - a
while e > ε
    e = e/2, c = a + e
    if sign(f(c)) = sign(f(a))
        a = c
    else
        b = c
```

2.3 Navadna iteracija

Algoritem (Navadna iteracija)**.**

```
izberi x_0
r = 0, 1, 2,...
    x_{r+1} = g(x_r)
```

Ustavitveni kriterij:

a) $r > r_{max}$ (prekoračeno število korakov)

b) $|r_{x+1} - r_x| < \varepsilon$

Izrek 3. Naj bo $\alpha = g(\alpha)$ in naj iteracijska funkcija g na intervalu $I = [\alpha - \delta, \alpha + \delta]$ za nek $\delta > 0$ zadošča Lipschitzovem pogoju

$$|g(x) - g(y)| \;\leq\; m|x - y| \;\text{ za }\; x, y \in I,\; 0 \leq m < 1.$$

Potem za $\forall x_0 \in I$ zaporedje $x_{r+1} = g(x_r)$, $r = 0, 1, \ldots$ konvergira k α in velja

- $|x_r - \alpha| \;\leq\; m^r \cdot |x_0 - \alpha|$
- $|x_{r+1} - \alpha| \;\leq\; \dfrac{m}{1 - m} \cdot |x_{r+1} - x_r|$

Posledica. Naj bo $\alpha = g(\alpha)$, g *zvezno odvedljiva* in $|g'(\alpha)| < 1$. Potem $\exists \delta > 0$, da za $\forall x_0 \in [\alpha - \delta, \alpha + \delta]$ zaporedje $x_{r+1} = g(x_r)$ konvergira k α.

Definicija 2.2. Naj zaporedje $\{x_r\}$ konvergira proti α [2]. Pravimo, da zaporedje konvergira z redom konvergence p, če obstaja limita

$$\lim_{r\to\infty} \frac{|x_{r+1} - \alpha|}{|x_r - \alpha|^p} \; = \; C > 0.$$

Izrek 4. Naj bo iteracijska funkcija g p-krat *zvezno odvedljiva* v okolici negibne točke α. Če velja $g'(\alpha) = \ldots = g^{(p-1)}(\alpha) = 0$ in $g^{(p)}(\alpha) \neq 0$, potem zaporedje $x_{r+1} = g(x_r)$, $r = 0, 1, \ldots$, v okolici α konvergira z redom p. V primeru $p = 1$ mora za konvergenco veljati še $|g'(\alpha)| < 1$.

2.4 Tangentna metoda

Metoda.

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

Konvergenca:

$$|e_{r+1}| \; \approx \; C \cdot |e_r|^2$$

Posledica. Če je f *dvakrat zvezno odvedljiva* v okolici ničle α, potem tangentna metoda za dovolj dober začetni približek x_0 vedno konvergira k α.

Izrek 5. Naj bo funkcija f na $I = [a, \infty)$ *dvakrat zvezno odvedljiva, naraščajoča in ima ničlo na* $\alpha \in I$. Potem je α edina ničla na I in za $\forall x_0 \in I$ tangentna metoda konvergira le k α.

[2] $\lim_{r\to\infty} x_r = \alpha$

2.5 Metode brez f'

Metoda (Sekantna metoda)**.**

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

Konvergenca:

$$|e_{r+1}| \;\approx\; C \cdot |e_r| \cdot |e_{r-1}|$$

Metoda (Mullerjeva metoda)**.** Skozi točke $(x_r, f(x_r))$, $(x_{r-1}, f(x_{r-1}))$, $(x_{r-2}, f(x_{r-2}))$ potegnemo kvadratni polinom $y = p(x)$ in za x_{r+1} vzamemo tisto ničlo polinoma p, ki je bližje x_r.

Konvergenca:

$$|e_{r+1}| \;\approx\; C \cdot |e_r| \cdot |e_{r-1}| \cdot |e_{r-2}|$$

Metoda (Inverzna interpolacija)**.** Zamenjamo vlogi x in y in vzamemo kvadratni polinom $x = \mathcal{L}(y)$, ki gre skozi točke $(f(x_r), x_r)$, $(f(x_{r-1}), x_{r-1})$, $(f(x_{r-2}), x_{r-2})$. Za x_{r+1} vzamemo

$$x_{r+1} \;=\; \mathcal{L}(0).$$

Red konvergence je enak kot pri *Mullerjevi metodi.*

3 SISTEMI LINEARNIH ENAČB

3.1 Oznake in definicije

Definicija 3.1. Sistem n linearnih enačb z n neznankami pišemo v obliki

$$Ax \;=\; b, \quad A \in \mathbb{R}^{n\times n}\ (\mathbb{C}^{n\times n}),\ x, b \in \mathbb{R}^n\ (\mathbb{C}^n).$$

Definicija 3.2. Skalarni produkt vektorjev x in y je enaka

a) $x, y \in \mathbb{R}^n$:

$$y^T x \;=\; \sum_{i=1}^{n} x_i y_i \;=\; \langle x, y\rangle \;=\; \langle y, x\rangle$$

b) $x, y \in \mathbb{C}^n$:

$$y^H x \;=\; \sum_{i=1}^{n} x_i \overline{y_i} \;=\; \langle x, y, \rangle \;=\; \overline{\langle y, x\rangle}$$

Definicija 3.3. Množenje vektorja x z matriko A:

a)

$$y_i \;=\; \sum_{k=1}^{n} a_{ik} x_k \;=\; \alpha_i^T x$$

b)

$$y \;=\; \sum_{i=1}^{n} x_i a_i$$

Definicija 3.4. $A \in \mathbb{R}^{n \times n}$ je *nesingularna*, če (ekvivalentno):

a) $\det(A) \neq 0$

b) obstaja inverz A^{-1}, da je $A^{-1}A = AA^{-1} = I$

c) $rang(A) = n$

d) za $\forall x \neq 0$ je $Ax \neq 0$

e) $\ker(A) = \{x \mid Ax = 0\} = \{0\}$

Definicija 3.5. Matrika je *simetrično pozitivno definitna*, če $A = A^T$ in $x^T A x > 0$ za $x \neq 0$.

Definicija 3.6. Če za $x \neq 0$ velja $Ax = \lambda x$, je λ lastna vrednost in x lastni vektor. Vsaka matrika ima n lastnih vrednosti, ki so ničle karakterističnega polinoma

$$p(\lambda) \ := \ \det(A - \lambda I).$$

3.2 Vektorske in matrične norme

Definicija 3.7. Vektorska norma je preslikava $\|.\| : \mathbb{C}^n \to \mathbb{R}$, da velja:

1) Nenegativnost:

$$\|x\| \ \geq \ 0, \quad \|x\| = 0 \Leftrightarrow x = 0$$

2) Homogenost:

$$\|\alpha x\| \ = \ |\alpha| \cdot \|x\|$$

3) Trikotniška neenakost:

$$\|x + y\| \ \leq \ \|x\| + \|y\|$$

Definicija 3.8. Matrična norma je preslikava $\|.\| : \mathbb{C}^{n\times n} \to \mathbb{R}$, da velja:

1) Nenegativnost:
$$\|A\| \;\geq\; 0, \quad \|A\| = 0 \Leftrightarrow A = 0$$

2) Homogenost:
$$\|\alpha A\| \;=\; |\alpha| \cdot \|A\|$$

3) Trikotniška neenakost
$$\|A + B\| \;\leq\; \|A\| + \|B\|$$

4) Submultiplikativnost:
$$\|A \cdot B\| \;\leq\; \|A\| \cdot \|B\|$$

za $\forall A, B \in \mathbb{C}^{n\times n}$ in $\forall \alpha \in \mathbb{C}$.

Lema 1 (Operatorska norma)**.** Če je za poljubno vektorsko normo $\|.\|_v$ definiramo
$$\|A\| \;:=\; \max_{x\neq 0} \frac{\|Ax\|_v}{\|x\|_v},$$
je to matrična norma.

Lema 2.
$$\|A\|_1 \;=\; \max_{j=1,\ldots,n} \|a_j\|_1 \;=\; \max_{j=1,\ldots,n} \sum_{i=1}^{n} |a_{ij}|$$

Lema 3 (Spektralna norma)**.**
$$\|A\|_2 \;=\; \sigma_1(A) = \max_{i=1,\ldots,n} \sqrt{\lambda_i(A^H A)}$$

Lema 4. Za vsako matrično normo $\|.\|_m$ obstaja taka vektorska norma $\|.\|_v$, ki je z njo usklajena, kar pomeni, da za vse pare A, x velja
$$\|Ax\|_v \;\leq\; \|A\|_m \cdot \|x\|_v.$$

Lema 5. Za vsako lastno vrednost λ in poljubno matrično normo $\|A\|$ velja:

$$|\lambda| \ \leq \ \|A\|.$$

3.3 Občutljivost sistemov linearnih enačb

Lema 6. Če je $\|X\| < 1$, potem velja:

a) $I - X$ je *nesingularna*

b) $(I - X)^{-1} \ = \ \sum_{i=0}^{\infty} X^i \ = \ I + X + X^2 + \dots$

c) Če je $\|I\| = 1$, je $\|(I - X)^{-1}\| \ \leq \ \dfrac{1}{1 - \|X\|}$

Izrek 6. Naj bo A *nesingularna* in ΔA taka motnja, da je $\|\Delta A\| < \frac{1}{\|A^{-1}\|}$. Če je $Ax = b$ in $(A + \Delta A)(x + \Delta x) = b + \Delta b$, potem velja:

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

kjer je $K(A) = \|A\| \cdot \|A^{-1}\|$.

3.4 LU razcep

Izrek 7. Za matriko A je ekvivalentno:

1) Obstaja enoličen razcep $A = LU$

2) Vse vodilne podmatrike $A_k = A(1:k, 1:k)$ so nesingularne

Algoritem (LU razcep)**.**

```
j = 1,...,n-1
    i = j+1,...,n
        l_ij = a_ij / a_jj
        k = j+1,...,n
            a_ik = a_ik - l_ij a_jk
```

Zahtevnost algoritma je odvisna od števila operacij. Preštejmo osnovne računske operacije:

$$\sum_{j=1}^{n-1}\left(\sum_{i=j+1}^{n}\left(1+\sum_{k=j+1}^{n}2\right)\right) = \frac{(n-1)n}{2}+2\frac{(n-1)n(2n-1)}{6} = \frac{2}{3}n^3+\sigma(n^2)$$

Algoritem (Prema substitucija)**.**
Sistem $Ly = b$ rešujemo s premo substitucijo:

```
i = 1,...,n
    y_i = b_i - sum_{j=1}^{i-1} l_ij y_j
```

Število operacij:

$$\sum_{i=1}^{n}(1+2(i-1)) = n+2\frac{(n-1)n}{2} = n^2$$

Algoritem.

```
i = n,n-1,...,1
    x_i = 1/u_ii (y_i - sum_{j=i+1}^{n} u_ij x_j)
```

Število opracij:

$$n^2+n$$

Algoritem (LU razcep z delnim pivotiranjem).

```
j = 1,...,n-1
    poisci |a_{p_j}| = max_{j≤l≤n} |a_{l_j}|
    zamenjaj vrstici j in p
    i = j+1,...,n
        l_ij = a_ij / a_jj
        k = j+1,...,n
            a_ik = a_ik - l_ij a_jk
```

Pri pivotiranju vedno velja $|l_{ij}| \leq 1$, saj saj je $|a_{ij}| \leq |a_{jj}|$

Izrek 8. Če je $\det(A) \neq 0$, potem obstaja taka permutacijska matrika P, da obstaja LU razcep

$$PA = LU,$$

kjer je L spodnjetrikotniška matrika, U pa zgornjetrikotniška matrika.

3.5 Stabilnost reševanja sistemov in LU razcepov

Lema 7. Naj bo L *nesingularna spodnje trikotna* matrika velikosti $n \times n$. Če sistem $Ly = b$ rešimo s premo substitucijo, potem izračunani $\hat{y}$ zadošča

$$(L + \Delta L)\hat{y} = b,$$

kjer je $|\Delta L| \leq nu|L|$. Reševanje spodenje trikotnega sistema s premo substitucijo je torej obratno stabilno.

Lema 8. Naj bo U *zgornje trikotna* matrika velikosti $n \times n$, $\det U \neq 0$. Če sistem $Ux = y$ rešimo z obratno substitucijo, potem numerično izračunani $\hat{x}$ zadošča

$$(U + \Delta U)\hat{x} = y,$$

kjer je

$$|\Delta U| \leq nu|U|.$$

Reševanje z obratno substitucijo je tudi obratno stabilno.

Lema 9. Naj bo A *nesingularna* matrika velikosti $n \times n$, kjer se izvede LU razcep brez pivotiranja. Za izračunani matriki $\hat{L}$ in $\hat{U}$ velja

$$A + E \; = \; \hat{L}\hat{U},$$

kjer je

$$|E| \; \leq \; nu|\hat{L}||\hat{U}|.$$

Izrek 9. Za izračunano rešitev $\hat{x}$ sistema linearnih enačb $Ax = b$ z uporabo LU razcepa velja

$$(A + \Delta A)\hat{x} \; = \; b,$$

kjer je

$$|\Delta A| \; \leq \; 3nu|L||U| + \sigma(u^2).$$

4 SISTEMI NELINEARNIH ENAČB

4.1 Uvod

Definicija 4.1. Sistem nelinearnih enačb je sistem oblike

$$\begin{aligned} f_1(x_1, \ldots, x_n) &= 0 \\ &\vdots \\ f_n(x_1, \ldots, x_n) &= 0 \end{aligned}$$

za $f_i : \mathbb{R}^n \to \mathbb{R}$. Krajše zapišemo kot

$$F(x) \;=\; 0, \quad F = \begin{bmatrix} f_1 \\ \vdots \\ f_n \end{bmatrix},$$

kjer je $F : \mathbb{R}^n \to \mathbb{R}^n$ in $x \in \mathbb{R}^n$.

4.2 Navadna iteracija

Metoda. To je posplošitev navadne iteracije za eno spremenljivko. Poiščemo funkcijo $G : \mathbb{R}^n \to \mathbb{R}^n$, da velja

$$F(\alpha) = 0 \;\Leftrightarrow\; \alpha = G(\alpha), \quad \alpha \in \mathbb{R}^n.$$

Izberemo $x^{(0)} \in \mathbb{R}^n$ in tvorimo zaporedje

$$x^{(r+1)} \;=\; G(x^{(r)}), \quad r = 0, 1, \ldots$$

Izrek 10. Naj bo bo $G : \mathbb{R}^n \to \mathbb{R}^n$ zvezno odvedljiva na $\Omega \subset \mathbb{R}^n$. Če velja:

a) $x \in \Omega \;\Rightarrow\; G(x) \in \Omega$

b) $x \in \Omega \;\Rightarrow\; \rho(JG(x)) \leq m < 1$

potem ima G v Ω natanko eno negibno točko α in zaporedje $x^{(r+1)} = G(x^{(r)})$, $r = 0, 1, \ldots$, za $\forall x^{(0)} \in \Omega$ konvergira k α. Tu je $JG(x) = \begin{bmatrix} \frac{\partial g_1}{\partial x_1}(x) & \cdots & \frac{\partial g_1}{\partial x_n}(x) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_n}{\partial x_1}(x) & \cdots & \frac{\partial g_n}{\partial x_n}(x) \end{bmatrix}$ Jacobijeva matrika parcialnih odvodov in ρ spektralni radij[3].

4.3 Newtonova metoda

Metoda. Naj bo x približek za ničlo funkcije f. Iščemo popravek Δx, da bo $F(x + \Delta x) = 0$. Z uporabo razvoja v Taylorjevo vrsto dobimo

$$F(x + \Delta x) \ = \ F(x) + JF(x) \cdot \Delta x + \underbrace{\mathcal{O}(\|\Delta x\|^2)}_{\text{zanemarimo}}.$$

Iz tega sledi $\Delta x = -JF(x)^{-1} \cdot F(x)$, za nov približek vzamemo torej

$$x + \Delta x \ = \ x - JF(x)^{-1} \cdot F(x),$$

za iteracijsko funkcijo pa

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

Algoritem.

$x^{(0)}$
r = 0, 1,...
    $JF(x^{(r)}) \cdot \Delta x^{(r)} = -F(x^{(r)})$
    $x^{(r+1)} = x^{(r)} + \Delta x^{(r)}$

[3]Največja absolutna vrednost lastnih vrednosti.

5 LINEARNI PROBLEM NAJMANJŠIH KVADRATOV

5.1 Uvod

Metoda. To je primer predoločenega sistema, kjer imamo več enačb kot neznank. V splošnem[4] nima rešitve. Namesto tega iščemo x, ki minimizira normo ostanka $Ax - b$. Če iščemo minimum $\|Ax - b\|_2$, potem je to linearni problem najmanjših kvadratov. Ustrezen x je rešitev po *metodi najmanjših kvadratov*.

Izrek 11. Naj bo $A \in \mathbb{R}^{m\times n}$, $b \in \mathbb{R}^m$, $m \geq n$, $\text{rang}(A) = n$. Rešitev normalnega sistema

$$A^T Ax = A^T b$$

je rešitev predoločenega sistema $Ax = b$ po metodi najmanjših kvadratov.

Opomba. Pravilna izpeljava:

$$A^T(Ax - b) = 0 \quad \Rightarrow \quad A^T Ax = A^T b$$

5.2 Razcep Cholskega

Izrek 12. Dan je sistem $Bx = c$, B je simetrična pozitivno definitna[5] matrika velikosti $n \times n$. Velja:

1. Če je B simetrična pozitivno definitna $\Rightarrow$ vse njene vodilne podmatrike so simetrične pozitivno definitne.
2. Če je B simetrična pozitivno definitna $\Rightarrow$ obstaja LU razcep $B = LU$, kjer je $u_{ii} > 0$ za $i = 1, \ldots, n$.

[4] Razen v primeru kot je $b \in \text{Im}(A)$

[5] $B = B^T, \quad z^T Bz > 0$ za $z \neq 0$

3. B je simetrično pozitivno definitna $\Leftrightarrow$ obstaja nesingularna spodnje trikotna matrika U, da je $B = UU^T$, $u_{ii} > 0$ za $i = 1, \ldots, n$.

Algoritem.

$$
\begin{aligned}
&\texttt{k = 1,..., n} \\
&\quad v_{kk} = \Big(b_{kk} - \sum_{i=1}^{k-1} {v_{ki}}^2\Big)^{\frac{1}{2}} \\
&\quad \texttt{j = k + 1,..., n} \\
&\qquad v_{jk} = \frac{1}{v_{kk}}\Big(b_{jk} - \sum_{i=1}^{k-1} v_{ji} v_{ki}\Big)
\end{aligned}
$$

Število operacij:

$$\sum_{k=1}^{n} (2k + 2(n-k)k) \;=\; \frac{1}{3}n^3 + \mathcal{O}(n^2)$$

Metoda. Reševanje sistema $Bx = c$, B je simetrična pozitivno definitna matrika:

1. $B = VV^T$
2. reši $Vy = b$
3. reši $V^T x = y$

Metoda. Dana je $A \in \mathbb{R}^{m \times n}$, $m > n$, $\operatorname{rang}(A) = n$. Po metodi najmanjših kvadratov iščemo x. $A^T A$ je simetrična, uporabimo razcep Choleskegta:

1. $B = A^T A$, $c = A^T b$
2. $B = VV^T$, V je spodnje trikotna matrika
3. reši $Vy = c$
4. reši $V^T x = y$

5.3 QR razcep

Izrek 13. Naj bo $A \in \mathbb{R}^{m \times n},\ m \geq n,\ \operatorname{rang}(A) = n$. Potem obstaja enoličen razcep $A = QR$, kjer je Q matrika velikosti $m \times n$ z ortonormiranimi stolpci in R zgornje trikotna matrika velikosti $n \times n$ s pozitivnimi diagonalnimi elementi.

Algoritem.

$\texttt{k = 1,..., n}$
$\quad \texttt{q}_\texttt{k} \texttt{ = a}_\texttt{k}$
$\quad \texttt{i = 1,..., k-1}$
$\qquad \texttt{r}_\texttt{ik} \texttt{ = q}_\texttt{i}^\texttt{T} \cdot \texttt{a}_\texttt{k}$
$\qquad \texttt{q}_\texttt{k} \texttt{ = q}_\texttt{k} \texttt{ - r}_\texttt{ik} \cdot \texttt{q}_\texttt{i}$
$\quad \texttt{r}_\texttt{kk} \texttt{ = } \|\texttt{q}_\texttt{u}\|_2$
$\quad \texttt{q}_\texttt{k} \texttt{ = } \frac{1}{\texttt{r}_\texttt{ku}} \texttt{q}_\texttt{k}$

Število operacij:

$$\sum_{k=1}^{n} \left(\sum_{i=1}^{k-1} (2m + 2m) + 2m + 1 + m \right) \; = \; 2mn^2 + \mathcal{O}(mn)$$

5.4 Givensove rotacije

Algoritem.

$\tilde{\texttt{Q}} \texttt{ = I}_\texttt{m} \texttt{ (Ce potrebujemo } \tilde{\texttt{Q}}\texttt{)}$
$\texttt{i = 1,..., n}$
$\quad \texttt{k = i+1,..., m}$
$\qquad \texttt{r = (a}_\texttt{ii}^\texttt{2} \texttt{ + a}_\texttt{ki}^\texttt{2}\texttt{)}^{\frac{1}{2}}\texttt{, c = } \frac{\texttt{a}_\texttt{ii}}{\texttt{r}}\texttt{, s = } \frac{\texttt{a}_\texttt{ki}}{\texttt{r}}$
$\qquad \texttt{A([i k], i:n) = } \begin{bmatrix} \texttt{c} & \texttt{s} \\ \texttt{-s} & \texttt{c} \end{bmatrix} \texttt{ A([i k], i:n)}$
$\qquad \texttt{b([i k]) = } \begin{bmatrix} \texttt{c} & \texttt{s} \\ \texttt{-s} & \texttt{c} \end{bmatrix} \texttt{ b([i k]) (Ce resujemo Ax = b po m.n.k.)}$
$\qquad \tilde{\texttt{Q}}\texttt{([i k],:) = } \begin{bmatrix} \texttt{c} & \texttt{s} \\ \texttt{-s} & \texttt{c} \end{bmatrix} \tilde{\texttt{Q}}\texttt{([i k],:)}$
$\tilde{\texttt{Q}} \texttt{ = } \tilde{\texttt{Q}}^\texttt{T}$

Število operacij:

$$\sum_{i=1}^{n} \sum_{k=i+1}^{m} (6 + 6(n - i + 1) + 6) \;\approx\; \sum_{i=1}^{n} 6(n-i)(m-i) \;=\; 3mn^2 - n^3$$

Če potrebujemo $\tilde{Q}$, je to še dodatnih $6m^2n - 3mn^2$ operacij.

5.5 Householderjeva zrcaljenja

Algoritem.

```
Q̃ = I_m (Ce potrebujemo Q̃)
i = 1,..., n (v primeru m=n le do n-1)
    doloci w_i ∈ ℝ^(m-n+1), ki prezrcali A(i:m, i) v ±*·e_1
    A(i:m, i:n) = P_i·A(i:m, i:n)
    b(i:m) = P_i·b(i:m)
    Q̃(i:m,:) = P_i·Q̃(i:m,:)
Q̃ = Q̃^T
```

Število operacij:

$$\sum_{i=1}^{n} (2(m-i+1) + (n-i+1)4(m-i+1) + 4(m-i+1)) \;\approx$$

$$\approx\; \sum_{i=1}^{n} 4(n-i)(m-i) \;\approx\; 2mn^2 - \frac{2}{3}n^3$$

Če potrebujemo še $\tilde{Q}$ porabimo še dodatnih $4m^2n - 2mn^2$ operacij.

5.6 Singularni razcep (SVD)

Izrek 14. Za $A \in \mathbb{R}^{m\times n}$, $m \geq n$, obstaja razcep $A = U\Sigma V^T$, kjer je U ortogonalna matrika velikosti $m \times m$, V ortogonalna matrika velikosti $n \times n$ in Σ matrika velikosti $m \times n$ oblike

$$\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_n \\ & & \\ & & \end{bmatrix},$$

kjer so $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_n \geq 0$ singularne vrednosti matrike A.

Lema 10. Naj bo $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\operatorname{rang}(A) = n$. Potem je za $b \in \mathbb{R}^m$ minimum $\|Ax - b\|_2$ dosežen pri

$$x = \sum_{i=1}^{n} \frac{{u_i}^T b}{\sigma_i} v_i.$$

Trditev 3. Naj bo $A \in \mathbb{R}^{m\times n}$, $m \geq n$, $\operatorname{rang}(A) = r < n$. Če je $A = U\Sigma U^T$ singularni razcep A, potem ima izmed vseh vektorjev $x \in \mathbb{R}^n$, ki minimizirajo $\|Ax - b\|_2$, najmanjšo normo $\|x\|_2$ vektor

$$x = \sum_{i=1}^{r} \frac{{u_i}^T b}{\sigma_i} v_i.$$

Definicija 5.1. Matrike $X \in \mathbb{R}^{n\times m}$ je psevdoinverz matrike A, če zadošča naslednjim pogojem:

1. $AXA = A$
2. $XAX = X$
3. $(AX)^T = AX$
4. $(XA)^T = XA$

Če je A matrika velikosti $n \times n$, $\det(A) \neq 0$, potem je

$$A^+ \; = \; A^{-1}.$$

Če je A matrika velikosti $m \times n$, $\operatorname{rang}(A) = r$, $m \geq n$, potem je

$$A^+ \; = \; (A^T A)^{-1} A^T.$$

Izrek 15. Za $A \in \mathbb{R}^{m \times n}$, $A = U \Sigma V^T$, je

$$A^+ \; = \; V \Sigma^+ U^T,$$

kjer je

$$\Sigma^+ \; = \; \begin{bmatrix} {\sigma_1}^{-1} & & & & & \\ & \ddots & & & & \\ & & {\sigma_r}^{-1} & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix},$$

če je $\operatorname{rang}(A) = r$.

Če je $A = U \Sigma V^T$, je

$$\begin{aligned} A \; &= \; \sum_{i=1}^{r} \sigma_1 u_i {v_i}^T \\ A^+ \; &= \; \sum_{i=1}^{r} \frac{1}{\sigma_i} v_i {u_i}^T \\ A^+ b \; &= \; \sum_{i=1}^{r} \frac{{u_i}^T b}{\sigma_i} v_i \end{aligned}$$